Submitted to the Annals of Statistics

OPTIMIZATION HIERARCHY FOR FAIR STATISTICAL DECISION PROBLEMS

BY ANIL ASWANI[1,a] AND MATT OLFAT[1,b]

[1] *Industrial Engineering and Operations Research, University of California, Berkeley,*
[a] *aaswani@berkeley.edu;* [b] *molfat@berkeley.edu*

Data-driven decision-making has drawn scrutiny from policy makers due to fears of potential discrimination, and a growing literature has begun to develop fair statistical techniques. However, these techniques are often specialized to one model context and based on ad-hoc arguments, which makes it difficult to perform theoretical analysis. This paper develops an optimization hierarchy, which is a sequence of optimization problems with an increasing number of constraints, for fair statistical decision problems. Because our hierarchy is based on the framework of statistical decision problems, this means it provides a systematic approach for developing and studying fair versions of hypothesis testing, decision-making, estimation, regression, and classification. We use the insight that qualitative definitions of fairness are equivalent to statistical independence between the output of a statistical technique and a random variable that measures attributes for which fairness is desired. We use this insight to construct an optimization hierarchy that lends itself to numerical computation, and we use tools from variational analysis and random set theory to prove that higher levels of this hierarchy lead to consistency in the sense that it asymptotically imposes this independence as a constraint in corresponding statistical decision problems. We demonstrate numerical effectiveness of our hierarchy using several data sets, and we use our hierarchy to fairly perform automated dosing of morphine.

1. Introduction. There is growing concern that improperly designed data-driven approaches to decision-making may display biased or discriminatory behavior. Such concerns are justified by examples of unfair algorithms that have been deployed in the real world [4, 10, 38, 74]. In response, researchers have started to develop approaches to encourage fairness in various statistical or learning problems [8, 21, 26, 27, 28, 37, 40, 45, 50, 61, 65, 75, 76, 81, 108, 114]. Fair statistics and learning approaches seek to estimate a model that predicts a response variable using a vector of predictor variables, while ensuring that model predictions are fair (we discuss quantitative measures of fairness in the next subsection) with respect to a finite set of variables that characterize designated protected attributes (e.g., gender or race).

1.1. *Quantitative Measures of Fairness.* The existing literature proposes several different fairness measures. We believe the underlying (and unifying) idea behind these measures is they approximately quantify independence between the output of a model and the variables of protected attributes. This way of thinking about fairness was first noticed by [53].

Consider fairness for classification: Let $(X, Y, Z) \in \mathbb{R}^p \times \{\pm 1\} \times \{\pm 1\}$ be a jointly distributed random variable consisting of a vector of predictors, a binary class label, and a binary protected attribute. Let $\delta(x)$ be a score for a classifier that operates on X, and suppose the classifier makes binary predictions $\widehat{Y}(x,t) = \operatorname{sign}(t - \delta(x))$ for a given threshold t of the score. Since binary classifiers output a ± 1 that can be mapped to desirable/undesirable decisions, one measure of fairness is $KS = \max_{t \in \mathbb{R}} |\mathbb{P}[\widehat{Y}(X,t) = +1 | Z = +1] - \mathbb{P}[\widehat{Y}(X,t) =$

MSC2020 subject classifications: Primary 62C12, 62F12; secondary 49J53, 60D05.
Keywords and phrases: Fairness, Independence, Optimization, Statistical Learning.

$+1|Z=-1]|$. This measures how similar the probability of making a prediction of a given binary class is between the two groups specified by the protected attributed, and it is often called *disparate impact* [50, 76]. Disparate impact measures the total disparity in outcomes between protected classes. By considering the maximum over all thresholds t, this definition ensures fairness in real-world scenarios where the score $\delta(x)$ itself is used for decision-making [76].

This above measure of fairness can be too strict in some applications, as there may be unavoidable correlation between the classifier output and the protected label. For such cases, [50] proposes *equalized odds* as an alternative measure of fairness that instead constrains disparity in outcomes conditional on some informative variable. In the setting of binary classification, one possible informative variable is $Y \in \{\pm 1\}$ itself. This choice leads to the following quantitative measure of equalized odds fairness $EO = \max_{y\in\{\pm1\}} \max_{t\in\mathbb{R}} |\mathbb{P}[\widehat{Y}(X,t) = +1|Z=+1, Y=y] - \mathbb{P}[\widehat{Y}(X,t)=+1|Z=-1, Y=y]|$. Restated, EO measures the disparity in *error rates* between the protected classes. An additional benefit is that a classifier with zero training error will also be fair with respect to this measure of fairness [50].

At an initial glance, the above fairness measures do not look like independence. Yet note event $\{\widehat{Y}(X,t)=+1\}$ is equivalent to event $\{\delta(X) \le t\}$ since $\widehat{Y}(x,t) = \text{sign}(t-\delta(x))$, implying that KS is the Kolmogorov-Smirnov (KS) distance between the distributions of $\delta(X)|Z=+1$ and $\delta(X)|Z=-1$. Since EO has a very similar interpretation, we will focus our discussion on KS. Thus when $KS=0$, we have that $G(t) := \mathbb{P}[\delta(X) \le t|Z=+1] = \mathbb{P}[\delta(X)\le t|Z=-1]$. Hence the joint distribution factorizes as $\mathbb{P}(\delta(X)\le t, Z=z) = \mathbb{P}[\delta(X)\le t|Z=z]\cdot\mathbb{P}(Z=z) = G(t)\cdot\mathbb{P}(Z=z)$, which means the two random variables are independent. Summarizing, we have $KS=0$ if and only if $\delta(X)$ is independent of Z.

1.2. *Existing Approaches to Fairness.* Pre-processing approaches transform the data before estimation, to remove information that could cause unfairness. (Non-)parametric approaches [23, 51, 110], dimensionality reduction approaches based on semidefinite programming (SDP) [75] and the Hilbert Schmidt Independence Criterion (HSIC) [81], and adversarial approaches based on autoencoders [14, 39, 66, 111] have been proposed.

Post-processing approaches process the output of a model to improve its fairness. A canonical example is [50], which designs a method for post-processing an arbitrary classifier in order to ensure fairness; however, it makes a poor tradeoff between accuracy and fairness [104]. Other approaches [28, 45] apply projections based on the Wasserstein distance to explicitly solve an empirical risk minimization problem with a fairness constraint. Unfortunately, these three approaches violate a principle called *individual fairness* [37], which says that individuals with similar predictor variables should have similar model predictions. One of the goals of this paper is to develop methods that can build models that satisfy individual fairness.

Pre- and post-processing approaches usually unlink the estimation process from ensuring fairness. This has motivated regularization approaches to fairness, which generally achieve lower generalization error while improving fairness. Empirical risk minimization formulations for classification [2, 8, 35, 44, 53, 76, 108], regression [3, 13, 22, 61, 81], and general problems [77] have been proposed. A now common approach is to control the correlation between model predictions and the protected attribute [22, 108], which can be generalized by further considering second-order deviations [76]; however, these approaches can only be used with categorical protected attributes. The approach of [53] works for more general protected attributes, but it requires a heuristic to approximate a *mutual information* (MI) measure of fairness as a constraint. Recent work extends these ideas to fair decision-making [40, 65].

1.3. *Contributions and Outline.* Though some of the above regularization approaches can provide strong theoretical guarantees about statistical and computational properties, most of the above approaches can only be used in specific settings. Many cannot be used with

more than one fair variable at a time, or cannot be used for models that produce more than one response variable. Some approaches only apply to categorical fair variables, while some approaches only apply to continuous fair variables. Similarly, some approaches can only be used for classification, while some approaches can only be used for regression.

The goal of this paper is to develop a single methodology for incorporating fairness into statistical decision problems that can be used under a broad number of settings. We first generalize in Section 3 the framework of statistical decision problems [60] to include fairness. This provides a systematic approach for developing and studying fair versions of hypothesis testing, decision-making, estimation, regression, and classification. In Section 4, we propose an optimization hierarchy, for fair statistical decision problems, that lends itself to numerical computation. Tools from variational analysis and random set theory are used to prove in Section 5 that higher levels of this hierarchy lead to consistency in the sense that it asymptotically imposes independence as a constraint in corresponding statistical decision problems for bounded random variables. Section 6 generalizes these results to unbounded random variables, namely sub-Gaussian random variables and random variables with finite moments. In Section 7, we demonstrate numerical effectiveness of our hierarchy using several data sets, and we conclude by using our hierarchy to fairly perform automated dosing of morphine. We note that all of our proofs can be found in the Supplementary Materials [6].

Our approach is to perform empirical risk minimization with constraints that ensure approximate independence between the output of a model and the variables of protected attributes. In fact, the broader idea of quantifying independence using an empirical estimate has a long history in statistics [18, 24, 42, 79, 97, 96, 73, 46]. The distinguishing feature of our approach to ensuring independence is to use a moment-based characterization of independence that generalizes Kac's theorem [16, 52] to multivariate random variables. Some fairness approaches [53] require solving bilevel programs [30, 78], which can be difficult to numerically solve. In contrast, our approach results in an optimization problem with constraints that are smooth polynomial functions, which allows us to leverage advances in convex optimization [59] and related heuristics such as the constrained convex-concave procedure [93, 101, 107] for the purpose of numerically solving the resulting optimization problem. The tradeoff is that we have to include multiple (but a finite number of) constraints.

Our framework differs from recent work on fair empirical risk minimization in several important ways. Some approaches use HSIC [61, 81] or Renyi correlation [8] as a fairness constraint, as opposed to the moment approach we take. (There is in fact a history of using HSIC as a component of optimization problems for tasks such as feature selection [95] and clustering [94].) Fairness is defined in [2, 35] using conditional probabilities, which because of the classification setup considered can be exactly rewritten as a conditional expectation. This allows the fairness constraints to be represented by a finite number of inequalities using sample averages in place of the conditional expectations. In contrast, our framework applies to problems such as regression where fairness as defined by statistical independence cannot be exactly rewritten as a conditional expectation. The work in [3] extends these ideas to regression by a performing a discretization that results in approximation of regression by a classification problem. The fairness constraints in this approach require discrete protected classes, whereas our framework is also able to handle continuous and vector-valued (consisting of both discrete and continuous) protected attributes. The formulation in [77] applies to general risk minimization problems, defines fairness in terms of conditional expectation, and proposes an approximation to ensure convexity of the resulting optimization problem. Our framework uses a different definition of fairness in terms of statistical independence.

Our framework also builds on preliminary work on the use of moment-based constraints for fair statistical methods [76, 75, 108]. These approaches were restricted to binary classification with binary protected classes, made use of only first- or second-order moments of

only the classifier, were based on ad-hoc arguments and justifications, and lacked theoretical analysis of the resulting statistical methods. The past papers [76, 75, 108] leave open the larger question of how moment-based approaches to fairness can be generalized to continuous protected classes, multivariate protected classes, multivariate statistical decisions, and other classes of statistical problems beyond classification. Our work in this paper unifies these past approaches into a broader theoretical framework, proves this framework provides asymptotic and finite-sample guarantees on fairness, and successfully generalizes moment-based methods in order to handle continuous protected classes, multivariate protected classes, multivariate statistical decisions, and multiple classes of statistical decision problems, including fair versions of hypothesis testing, decision-making, estimation, regression, and classification.

Because we have to include multiple constraints, this significantly complicates the theoretical analysis of our optimization hierarchy. The limiting behavior of our framework requires a statistical analysis on the solution to an optimization problem in the limit of a countably-infinite number of random constraints involving empirical moments. Traditional results in statistics do not apply to set-valued functions [5], which are one way to interpret constraints in an optimization problem [84]. In fact, most attention in statistics on sets has been focused on estimating a single set under different measurement models [34, 47, 57, 80, 91]. The traditional theoretical argument is to use the Pompeiu–Hausdorff distance to metricize the set of sets, but this approach is too difficult for use in our setting which has random sets defined using (in the limit) an infinite number of non-convex constraints. Instead, we build on our past work on statistics with set-valued functions [5]: We develop new theoretical arguments for statistics with random sets and set-valued functions, using variational analysis [84, 85] and random sets [70, 71]. These techniques are of potential interest to other set-based statistical problems where empirically-successfully approaches without theoretical guarantees have been used [109]. Examples of such statistical problems include estimation tasks where the predictor variables are a set and the response variable is a scalar, such as galaxy red-shift estimation in cosmology [86] and point-cloud classification in computer vision [105].

2. Preliminaries. This section presents our notation. We also describe some notation and definitions from variational analysis and random sets. Most of the variational analysis definitions are from [84], and the stochastic set convergence notation is originally from [5].

2.1. *Notation.* Let $M : \mathbb{R}^{dp} \to \mathbb{R}^{d\times p}$ be the function that reshapes a vector into a matrix by placing elements into the matrix columnwise from the vector. Similarly, we define $W := M^{-1} : \mathbb{R}^{d\times p} \to \mathbb{R}^{dp}$ to be its inverse.

We use $\mathbb{E}_n(\cdot)$ to denote expectation with respect to the empirical distribution. Recall this is the sample average of the random variable inside parenthesis. As examples, $\mathbb{E}_n(Z) = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\mathbb{E}_n(ZX) = \frac{1}{n}\sum_{i=1}^n Z_i X_i$.

Consider a tensor $\varphi \in \mathbb{R}^{r_1\times\cdots\times r_q}$, and let $[r] = \{1, \ldots, r\}$. The norm $\|\varphi\|$ is the ℓ_∞ vector norm for the tensor considered as a vector. For two tensors $\varphi, \nu \in \mathbb{R}^{r_1\times\cdots\times r_q}$, we define their inner product $\langle\varphi, \nu\rangle$ to be the usual dot product for the tensors interpreted as vectors.

For a tensor interpreted as a multilinear operator $\varphi(u_1, \ldots, u_q)$, we define the two subordinate norms: first that $\|\varphi\|_\circ = \max\left\{\|\varphi(u, \ldots, u)\| \mid \|u\|_2 = 1\right\}$ and second that $\|\varphi\|_* = \max\{\|\varphi(u_1, \ldots, u_q)\| \mid \|u_k\|_2 = 1 \text{ for } k \in [q]\}$, where $\|\cdot\|_2$ is the Euclidean norm for vectors. These are subordinate norms since $\|\varphi(u, \ldots, u)\| \le \|\varphi\|_\circ \big(\|u\|_2\big)^q$ and $\|\varphi(u_1, \ldots, u_q)\| \le \|\varphi\|_* \prod_{k=1}^q \|u_k\|_2$. When $\varphi(\cdot, \ldots, \cdot)$ is symmetric in its arguments, then $\|\varphi\|_\circ = \|\varphi\|_*$ [9, 17].

2.2. *Variational Analysis.* Let $\overline{\mathbb{R}} = [-\infty, \infty]$ denote the extended real line. We define $\Gamma(\cdot, \mathcal{S}) : E \to \overline{\mathbb{R}}$ to be the indicator function defined as: $\Gamma(u, \mathcal{S}) = 0$ if $u \in \mathcal{S}$ and $\Gamma(u, \mathcal{S}) = +\infty$ otherwise, where E is some Euclidean space that will be clear from the context.

For a sequence of sets C_n, the outer limit is $\limsup_n C_n = \{x : \exists n_k \text{ s.t. } x_{n_k} \to x \text{ with } x_{n_k} \in C_{n_k}\}$, and the inner limit is $\liminf_n C_n = \{x : \exists x_n \to x \text{ with } x_n \in C_n\}$. The outer limit consists of all the cluster points of C_n, whereas the inner limit consists of all limit points of C_n. The limit of the sequence of sets C_n exists if the outer and inner limits are equal, and when it exists we write $\lim_n C_n := \limsup_n C_n = \liminf_n C_n$.

A sequence of extended-real-valued functions $f_n : X \to \overline{\mathbb{R}}$ is said to epi-converge to f if at each $x \in X$ we have: $\liminf_n f_n(x_n) \geq f(x)$ for every sequence $x_n \to x$ and $\limsup_n f_n(x_n) \leq f(x)$ for some sequence $x_n \to x$. Epi-convergence is so-named because it is equivalent to set convergence of the epigraphs of f_n, meaning that epi-convergence is equivalent to the condition $\lim_n \{(x,\alpha) \in X \times \mathbb{R} : f_n(x) \leq \alpha\} = \{(x,\alpha) \in X \times \mathbb{R} : f(x) \leq \alpha\}$. We use the notation $\text{e-}\lim_n f_n = f$ to denote epi-convergence relative to X.

A sequence of extended-real-valued functions $f_n : X \to \overline{\mathbb{R}}$ is said to converge pointwise to f if at each $x \in X$ we have that $\lim_n f_n(x) = f(x)$. We abbreviate pointwise convergence relative to X using the notation $\lim_n f_n = f$.

2.3. *Specific Distributions.* We define a multivariate random variable $U \in \mathbb{R}^p$ to be sub-Gaussian with variance parameter σ^2 if we have that $\mathbb{E}\exp(s \cdot \langle t, U - \mathbb{E}(U)\rangle) \leq \exp(\sigma^2 s^2/2)$ for all $t \in \mathbb{S}^{p-1}$, which is the unit sphere in p-dimensions. Thus a sub-Gaussian random variable also satisfies $\mathbb{E}\exp(s \cdot \langle t, U\rangle) \leq M\exp(\sigma^2 s^2)$ for all $t \in \mathbb{S}^{p-1}$, where $M \geq 1$ and $\sigma^2 \geq 0$ are constants; we will use this as our primary characterization. An important implication of this characterization is that $\mathbb{E}\big(\langle t, U\rangle^{2k}\big) \leq M\sigma^{2k} \cdot (2k)!/k!$ for all $t \in \mathbb{S}^{p-1}$.

Sub-Gaussian distributions are ubiquitous. Important examples of sub-Gaussian distributions include: A Gaussian distribution X with mean μ and variance σ^2 is denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, a Bernoulli random variable X with success probability $x \in [0,1]$ is denoted $X \sim \text{Ber}(x)$, and a uniform random variable X with support $[a,b]$ is denoted $X \sim \text{Uni}(a,b)$.

2.4. *Random Sets.* Let $(\mathcal{U}, \mathfrak{F}, \mathbb{P})$ be a complete probability space, where $\mathcal{U}$ is the sample space, $\mathfrak{F}$ is the set of events, and $\mathbb{P}$ is the probability measure. A map $S : \mathcal{U} \to \mathcal{F}$ is a random set if $\{u : S(u) \in \mathcal{X}\} \in \mathfrak{F}$ for each $\mathcal{X}$ in the Borel σ-algebra on $\mathcal{F}$ [71]. Like the usual convention for random variables, we notationally drop the argument for a random set.

When discussing stochastic convergence, we indicate a limit occurs almost surely by appending "as-" to the limit notation. For instance, notation $\text{as-}\limsup_n C_n \subseteq C$ denotes $\mathbb{P}(\limsup_n C_n \subseteq C) = 1$, and notation $\text{as-}\liminf_n C_n \supseteq C$ denotes $\mathbb{P}(\liminf_n C_n \supseteq C) = 1$.

3. Fair Statistical Decision Problems. We use the setting of statistical decision problems: Consider the random variables (X, Y, Z) that have a joint distribution $\mathcal{P} \in \mathcal{D}$ where $\mathcal{D}$ is some fixed family of distributions. The interpretation is that X gives descriptive information, Y has information about some target, and Z encodes protected information which we would like to be fair with respect to. We will not explicitly use Y in this paper, but we note that it is implicitly included within other terms that we discuss.

The goal is to construct a function $\delta(\cdot,\cdot)$ called a *decision rule*, which provides a decision $d = \delta(x,z)$. To evaluate the quality of a decision rule δ, we define a *risk function* $R(\delta)$. In this setup, an optimal decision rule is any function from $\arg\min_{\delta(\cdot,\cdot)} R(\delta)$. However, we can define a related optimization problem that chooses an optimal fair decision rule by solving

$$\delta^*(x,z) \in \arg\min_{\delta(\cdot,\cdot)} \big\{R(\delta) \;\big|\; \delta(X,Z) \perp\!\!\!\perp Z\big\}, \tag{1}$$

where the notation $\delta(X,Z) \perp\!\!\!\perp Z$ indicates independence of $\delta(X,Z)$ and Z.

The above abstract setup is useful because it allows us to reason about fairness for a wide class of problems using a single theoretical framework. This is demonstrated by the following (which is the first to our knowledge) example of a procedure for fair hypothesis testing:

EXAMPLE 1. We consider a parametric hypothesis test over distributions $\mathcal{D}$ given by

$$\begin{bmatrix} \Xi \\ \Psi \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \tag{2}$$

where ρ is known and fixed and $\mu \in \mathbb{R}$. Define the null hypothesis to be $H_0 : \mathbb{E}(\Xi) = 0$ under $\mathcal{D}$. Suppose $X = (\Xi_1, \ldots, \Xi_n)$ and $Z = (\Psi_1, \ldots, \Psi_n)$ consist of i.i.d. random variables. Let d_0 be the decision to accept the null, and let d_1 be the decision to reject the null. The traditional hypothesis test with a significance level of a corresponds to a decision rule δ that minimizes the risk function $R(\delta) = \mathbb{P}_{H_1}(\delta = d_0) + \Gamma(\mathbb{P}_{H_0}(\delta = d_1) - a, \mathbb{R}_{\leq 0})$, where $H_1 = \{\mathcal{P} \in \mathcal{D} : \mathcal{P} \neq H_0\}$ [60] and $\Gamma(\cdot, \cdot)$ is the indicator function that was defined in Section 2.2. An optimal decision rule for this risk is $\delta^* = d_0$ if $p \geq a$ and $\delta^* = d_1$ if $p < a$, where p is a p-value [60]. An optimal decision rule that depends only upon X corresponds to the use of a traditional p-value: $p_o = 2\Phi(-\sqrt{n}|\frac{1}{n}\sum_{i=1}^n \Xi_i|)$, with $\Phi(\cdot)$ being the standard normal c.d.f. Using the above framework, we can compute an optimal *fair* decision rule for this risk. This corresponds to $p_f = 2\Phi(-\sqrt{n/(1-\rho^2)}|\frac{1}{n}\sum_{i=1}^n (\Xi_i - \rho\Psi_i)|)$, which we can interpret as a *fair* p-value. Using p_f results in a test with greater *power* than using p_o, meaning the risk of the above decision rule with p_o is higher than the risk of the decision rule with p_f. This example is interesting because it shows that trying to achieve fairness by removing a protected attribute may lead to worse performance than a fair decision rule that uses this variable.

In many statistical contexts, $\mathcal{D}$ is singleton but unknown. We then instead choose the decision rule using a sample (X_i, Y_i, Z_i) for $i = 1, \ldots, n$, which is i.i.d. from the distribution $\mathcal{P}$. Towards this aim, we approximate the risk function $R(\delta)$ using an (random) approximate risk function $R_n(\delta)$ that depends upon the sample. However, computing a sample-based fair decision rule is not obvious because a statistically well-behaved, sample-based analog of the constraint $\delta(X, Z) \perp\!\!\!\perp Z$ from (1) has not been studied previously.

4. Fair Optimization Hierarchy. We next propose a framework for computing a fair decision rule by solving a sample-based analog of (1). We first describe our assumptions about the statistical and numerical properties of the problem. Next we present our framework and provide some intuition to justify the structure of our formulation. We conclude by discussing some of the favorable computational properties of our framework.

4.1. *Problem Setup.* We make assumptions about the decision rule to be constructed and about the random variables present in our setup. The first two assumptions say that we are solving (1) while constraining the decision rule to belong to a particular parametric class.

ASSUMPTION 1. The decision rule $\delta(x, z) = B \cdot \omega(x, z)$ belongs to a parametric polynomial family, where $B \in \mathcal{B}$ is a matrix, $\mathcal{B} \subset \mathbb{R}^{d \times p}$ is a compact set, and $\omega(x, z) \in \mathbb{R}^p$ is a vector of monomials of the entries of the vectors x, z. Here, B parametrizes the decision rule $\delta(x, z)$, and the function $\omega(x, z)$ is assumed to be known and fixed by our design such as through feature engineering. We define the random variable $\Omega = \omega(X, Z)$, so $\delta(X, Z) = B\Omega$.

REMARK 1. In some settings, it may be desirable to have the fair decision rule depend upon only X and not Z. The above includes this case by noting that $\omega(x, z)$ is free to be chosen to include only monomials of the entries of x.

REMARK 2. This assumption says the decision rules are linear with respect to some polynomial transformation of the X and Z. Such a linear decision rule may not be competitive in terms of risk minimization as compared to more sophisticated models, but linear decision rules are commonly used in many application domains such as health care or economics and as such are important to theoretically study in the setting of fairness.

ASSUMPTION 2. Assume $\mathcal{B} \subseteq \{B \in \mathbb{R}^{d\times p} : \|W(B)\|_2 \leq \sqrt{\lambda}\}$ for $\lambda \geq 1$.

REMARK 3. For certain distributions, it is possible that only constant decision rules are feasible for (1) when constrained by the first two assumptions. In some settings, this may in fact be the most desirable fair decision rule. But in settings where constant decision rules are too trivial for the corresponding application, we note that our framework handles a generalization of (1) where the independence constraint is replaced with a constraint that ensures approximate independence. This generalization is discussed in Sections 5.5, 6.2, and 7.1, and it ensures that our framework is able to generate nontrivial decision rules in all settings.

Our next assumption is about statistical properties of the approximate risk function. Since our primary interest in this paper is studying independence constraints, we directly make assumptions about the convergence of the approximate risk function. Showing that such convergence holds typically involves a separate statistical analysis for the problem at hand.

ASSUMPTION 3. Note the function $R_n(B \cdot \omega(x,z))$ is the approximate risk function composed with the parametric decision rule in Assumption 1. We assume that this function can be written in the form $h_n(B) := R_n(B \cdot \omega(x,z)) = f_n(B) + \Gamma(g_n(B), \{\mathbb{R}_{\leq 0}\}^\eta)$, where $f_n : \mathbb{R}^{d\times p} \to \mathbb{R}$ and $g_n : \mathbb{R}^{d\times p} \to \mathbb{R}^\eta$. Moreover, define the notation $h(B) = R(B \cdot \omega(x,z))$. We assume $\text{as-e-lim}\, h_n = \text{as-lim}\, h_n = h$ relative to $\mathcal{B}$.

REMARK 4. We should interpret the notation of $h_n(B)$ as simultaneously specifying an objective function $f_n(B)$ and a set of constraints $g_n(B) \leq 0$.

REMARK 5. This convergence assumption may look unfamiliar, but we note that it is weaker than the convergence results that are usually shown when proving consistency of estimators. In particular, almost sure uniform convergence of h_n to h implies the above.

The first three assumptions are primarily related to statistical properties. It is instructive to consider examples that show how linear regression and classification problems match these.

EXAMPLE 2. Linear regression with $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ in our setup would mean we choose a linear decision rule $\delta(x) = Bx$ with $B \in \mathbb{R}^{1\times p}$. We could use a squared loss $R_n(B \cdot x) = \frac{1}{n}\sum_{i=1}^n (Y_i - BX_i)^2$ or the least absolute deviation loss $R_n(B \cdot x) = \frac{1}{n}\sum_{i=1}^n |Y_i - BX_i|$ for our regression. The nondifferentiability of the latter can be managed by introducing the variables s_i and noting $R_n(B \cdot x) = \frac{1}{n}\sum_{i=1}^n s_i$ subject to the constraints $-s_i \leq Y_i - BX_i \leq s_i$. This matches the decomposition of $R_n(\delta)$ into an objective with constraints. These loss functions can hence be minimized by many algorithms.

EXAMPLE 3. Linear classification with $(X_i, Y_i) \in \mathbb{R}^p \times \{-1, +1\}$ in our setup would mean we choose a linear decision rule $\delta(x) = Bx$ with $B \in \mathbb{R}^{1\times p}$. We could use typical classification losses: Logistic regression uses $R_n(B \cdot x) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-Y_i \cdot BX_i))$. Because this logistic loss is convex and differentiable, it can be easily optimized. Support vector machine uses the hinge loss $R_n(B \cdot x) = \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - Y_i \cdot BX_i\}$. Its nondifferentiability is handled by introducing variables s_i and noting $R_n(B \cdot x) = \frac{1}{n}\sum_{i=1}^n s_i$ subject to constraints $s_i \geq 0$ and $s_i \geq 1 - Y_i \cdot BX_i$. This matches the decomposition of $R_n(\delta)$ into an objective with constraints, and this can be minimized by many algorithms.

The linear classifier makes binary predictions $\widehat{Y}(x,t) = \text{sign}(t - Bx)$ by applying a threshold t to the decision rule $\delta(x) = Bx$. This interpretation of using the score function as the decision rule is justified by the general principal of empirical risk minimization, and

since independence of $\delta(x) = Bx$ and Z implies independence between $\widehat{Y}(X,t)$ and Z. Furthermore, the hinge loss is classification-calibrated under certain conditions [55], meaning the hinge loss composed with the score function is statistically consistent with respect to the 0-1 classification loss composed with the thresholded binary predictions $\widehat{Y}(x,0)$.

EXAMPLE 4. We could consider the above linear classification setup using the 0-1 classification loss $R_n(B \cdot x) = \frac{1}{n}\sum_{i=1}^{n} H(-Y_i \cdot BX_i)$, where $H(\cdot) : \mathbb{R} \to \{0,1\}$ is the step function defined as: $H(u) = 0$ if $u \leq 0$ and $H(u) = 1$ otherwise. This loss is supported by our setup because Assumption 3 follows by applying standard uniform convergence results [103]. (Uniform convergence is technically stronger than the type of convergence required in Assumption 3.) However, the resulting optimization problem is an integer program [63]. The idea behind the integer programming formulation is that it uses binary variables to keep track of whether or not each $Y_i \cdot BX_i$ is nonnegative.

4.2. *Formulation.* We are now ready to present our framework. Given the above assumptions, we study use of the following sample-based optimal fair decision rule: The level-$(\mathfrak{g},\mathfrak{h})$ fair optimization (FO) is

$$\begin{aligned} &\min_{B \in \mathcal{B}} R_n(B \cdot \omega(x,z)) \\ &\text{s.t. } \left\| \mathbb{E}_n\left(Z^{\otimes m}(B\Omega)^{\otimes q}\right) - \mathbb{E}_n\left(Z^{\otimes m}\right) \otimes \mathbb{E}_n\left((B\Omega)^{\otimes q}\right) \right\| \leq \Delta_{m,q}, \\ &\qquad \text{for } (m,q) \in [\mathfrak{g}] \times [\mathfrak{h}] \end{aligned} \tag{3}$$

where $\mathfrak{g}, \mathfrak{h} \geq 1$ are integers and $\Delta_{m,q} \geq 0$ are nonnegative real numbers. We note that $\mathfrak{g}$, $\mathfrak{h}$, and $\Delta_{m,q}$ will generally be chosen to depend on n, but for simplicity we will not make this n-dependence explicit in our notation. Our optimization hierarchy for fair statistical decision problems is defined by the above formulation given in (3), with the increasing number of constraints in the hierarchy parametrized by increasing values of $\mathfrak{g}, \mathfrak{h}$. We will study the constraints of the above problem and show that they are statistically well-behaved analogs of the independence constraint in (1).

REMARK 6. The above formulation considers fairness in the sense of disparate impact. When the protected attributes are categorical, meaning $Z \in \mathcal{Z}$ for some finite-cardinality set $\mathcal{Z}$, then our formulation can be modified to consider fairness in the sense of equalized odds by replacing the constraints in the above formulation with the constraints $\|\mathbb{E}_n[Z^{\otimes m}(B\Omega)^{\otimes q}|Z=z] - \mathbb{E}_n[Z^{\otimes m}|Z=z] \otimes \mathbb{E}_n[(B\Omega)^{\otimes q}|Z=z]\| \leq \Delta_{m,q}$, for $(m,q) \in [\mathfrak{g}] \times [\mathfrak{h}]$ and $z \in \mathcal{Z}$. Compared to the above formulation, here we take expectations with respect to the empirical distribution conditioned on each possible value in $\mathcal{Z}$.

Our first result provides intuition about the constraints in the FO optimization problem (3). This result generalizes Kac's theorem [16, 52], which characterizes independence of random variables using moment conditions, to the setting of random vectors. This generalization is novel to the best of our knowledge, and so we include its proof for completeness.

THEOREM 1. *Let $M_{(U,V)}(s,t) = \mathbb{E}\exp(\langle s,U\rangle + \langle t,V\rangle)$ be the moment generating function for the multivariate random variable (U,V) where we have $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^d$. If $M_{(U,V)}(s,t)$ is finite in a neighborhood of the origin, then U and V are independent if and only if $\mathbb{E}\left(U^{\otimes m}V^{\otimes q}\right) = \mathbb{E}\left(U^{\otimes m}\right) \otimes \mathbb{E}\left(V^{\otimes q}\right)$ for $m,q \geq 1$.*

REMARK 7. This result requires that $M_{(U,V)}(s,t)$ exists in a neighborhood of the origin. Examples of distributions that satisfy this condition are those with a bounded support (almost surely), as well as those belonging to the sub-Gaussian, sub-exponential, or sub-gamma families of distributions. This encompasses a large number of the most common distributions.

Next, we show a similar result that characterizes approximate independence of random variables using moment conditions. The benefits of this next result are that: it holds for (possibly unbounded) distributions that have finite moments, and it does not require the existence of $M_{(U,V)}(s,t)$ in a neighborhood of the origin. Thus it applies to a larger class of distributions. Our characterization relating moment conditions to approximate independence is the first result of its kind, to our knowledge.

However, we have to specify how independence is quantified. A natural idea is to consider a distance between the joint distribution of $(Z, \widehat{B}_n\Omega)$ and the product distribution of Z and $\widehat{B}_n\Omega$. This idea is natural because independence means that the joint distribution equals the product distribution. Thus the pertinent detail is choosing a distance between distributions to use. Our next example shows a subtle issue in making this choice.

EXAMPLE 5. Consider a setting where $B \in \mathbb{R}$, where $\omega(x,z) = x$, and the distributions are $X \sim \mathrm{Uni}(-1,1)$ and $Z = X$. Then $\Omega = X$. Next let $\mathit{dl}(B) = \sup_{s,t} |\mathbb{P}_{(Z,B\Omega)}(Z \leq s, B\Omega \leq t) - \mathbb{P}_Z(Z \leq s) \cdot \mathbb{P}_{B\Omega}(B\Omega \leq t)|$ be the multivariate Kolmogorov-Smirnov distance between the joint and product distributions of Z and $B\Omega$. Now note $\mathit{dl}(0) = 0$ because Z is trivially independent of the constant $0 \cdot \Omega \equiv 0$. For any $B \neq 0$ we have $\mathit{dl}(B) = \mathit{dl}(1)$, but $\mathit{dl}(1) > 0$ since $Z = \Omega$. Hence for the sequence $B_n = n^{-1}$, we have that $B_n\Omega$ is asymptotically independent of Z but $\mathit{dl}(\lim_n B_n) = \mathit{dl}(0) = 0 \neq \lim_n \mathit{dl}(B_n) = \mathit{dl}(1) > 0$. As a result, the multivariate Kolmogorov-Smirnov distance cannot quantify independence here.

REMARK 8. Because the total variation distance is greater than or equal to the value of the multivariate Kolmogorov-Smirnov distance, the above example also applies to the total variation distance. Thus Pinsker's inequality implies the above example applies to the Kullback–Leibler (KL) divergence. So the above example also applies to mutual information, which is defined as the KL divergence between the joint and product distributions.

REMARK 9. A multivariate version of this example can be constructed where the same issue occurs for a $B \neq 0$, where the issue occurs because B does not have full column rank.

The above examples show that several popular distances between distributions cannot be used for quantifying the degree of independence in our setting of fair optimization. This is perhaps not surprising given that the notion of convergence in distribution is weaker than many popular distances. Consequently, we need to consider topologically-weaker metrics on probability distributions, that are able to metricize convergence in distribution.

One such distance is the Zolotarev metric defined using characteristic functions [115, 56, 82], and we will use this distance to quantify the degree of independence between two random variables. Let $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^d$ be random vectors, and define $\mathfrak{i} = \sqrt{-1}$. Then for $s \in \mathbb{R}^p$, $t \in \mathbb{R}^d$, and $\zeta \in \mathbb{R}$; let $J(s,t,\zeta) = \mathbb{E}\exp(\mathfrak{i}\zeta\langle s, U\rangle + \mathfrak{i}\zeta\langle t, V\rangle)$ and $P(s,t,\zeta) = \mathbb{E}\exp(\mathfrak{i}\zeta\langle s, U\rangle) \cdot \mathbb{E}\exp(\mathfrak{i}\zeta\langle t, V\rangle)$ be the characteristic functions corresponding to the joint and product distributions, respectively, of U and V. The Zolotarev metric between the joint and product distributions is given by

$$\mathbb{H}(U;V) = \sup_{(s,t)\in\mathbb{S}^{p+d-1}} \left[\inf_{T>0} \max\left\{ \frac{1}{2} \sup_{|\zeta|\leq T} \left| J(s,t,\zeta) - P(s,t,\zeta) \right|, \frac{1}{T} \right\} \right]. \tag{4}$$

We call the quantity $\mathbb{H}(U;V)$ the *mutual characteristic* of U and V, and the choice of this name is meant to draw a direct analogy to mutual information.

THEOREM 2. *Consider the random variable (U,V) where $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^d$. If $J_{\mathfrak{g},\mathfrak{h}} = \sup_{(s,t)\in\mathbb{S}^{p+d-1}} \mathbb{E}(\langle s,U\rangle^{\mathfrak{g}+1}\langle t,V\rangle^{\mathfrak{h}+1})$ is finite and $\mathbb{E}\big(U^{\otimes m}V^{\otimes q}\big) = \mathbb{E}\big(U^{\otimes m}\big) \otimes \mathbb{E}\big(V^{\otimes q}\big)$ for $m,q \in [\mathfrak{g}] \times [\mathfrak{h}]$, then we have that*

$$\mathbb{H}(U;V) \leq \left[\frac{J_{\mathfrak{g},\mathfrak{h}}+P_{\mathfrak{g},\mathfrak{h}}}{(\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!}\right]^{1/(\mathfrak{g}+\mathfrak{h}+3)} \tag{5}$$

where $P_{\mathfrak{g},\mathfrak{h}} = \sup_{s\in\mathbb{S}^{p-1}} \mathbb{E}(\langle s,U\rangle^{\mathfrak{g}+1}) \cdot \sup_{t\in\mathbb{S}^{d-1}} \mathbb{E}(\langle t,V\rangle^{\mathfrak{h}+1})$.

These two generalizations of Kac's theorem allow us to interpret the constraints of the FO problem (3). Using Theorem 1, we can interpret the constraints as a finite number ($\mathfrak{g}\cdot\mathfrak{h}$ many, for a level-$(\mathfrak{g},\mathfrak{h})$ FO problem) of sample-based analogs of the corresponding moment conditions from Theorem 1 for independence. Using Theorem 2, we can also interpret the constraints as sample-based analogs of the corresponding finite number of moment conditions that achieve approximate independence in the sense of (5).

4.3. *Computational Properties.* We next discuss some favorable computational properties of the FO problem (3). A key advantage of our framework is that the moment constraints are polynomials. This leads to three general approaches to numerically solve the FO problem.

The first approach when the relevant functions are polynomials, which allow us to draw upon powerful tools for polynomial optimization [59]:

THEOREM 3 (Theorems 5.6, 5.7 of [59]). *Suppose Assumptions 1–3 hold. If, in the notation of Assumption 3, we assume that the functions $f_n : \mathbb{R}^{d\times p} \to \mathbb{R}$ and $g_n : \mathbb{R}^{d\times p} \to \mathbb{R}^\eta$ are polynomials on the set $\mathcal{B}$, then the level-$(\mathfrak{g},\mathfrak{h})$ FO problem (3) can be solved to any desired accuracy by solving a convex optimization problem that can be explicitly constructed.*

REMARK 10. The polynomial assumption is not restrictive since the Stone-Weierstrass theorem shows that if f_n and g_n are continuous then they can be approximated to arbitrary accuracy by polynomials, since the domain of the optimization problem is a compact set $\mathcal{B}$. So this approach can be used for the squared loss, logistic loss, hinge loss (after the earlier reformulation), least absolute deviation loss (after the earlier reformulation), and other losses.

Though the convex optimization problems resulting from the explicit construction of [59] are often large, these resulting optimization problems can be numerically solved for many interesting instances [67, 112]. We provide some intuition. The first insight is that any polynomial optimization problem $\min\{f_n(B) \mid g_n(B) \leq 0, B \in \mathcal{B}\}$ can be written as maximizing a scalar subject to nonnegative polynomial constraints $\max\{s \mid f_n(B) - s \geq 0, -g_n(B) \geq 0, s \in \mathbb{R}, B \in \mathcal{B}\}$. The second insight is that nonnegative polynomials can be approximated on a bounded domain to arbitrary accuracy using sum-of-squares (SOS) polynomials [11, 59]. Since our problems involve optimizing a vector that belongs to Euclidean space, SOS polynomials are literally the set of polynomials that are generated by squaring arbitrary polynomials and then adding them up. Specifically, the nonnegative polynomial constraints can be approximated by instead asking for the polynomials to equal a linear combination of a finite number of SOS polynomials. This is a tractable approximation because the resulting optimization problem is a convex semidefinite program, and the following solution can be made arbitrarily accurate by increasing the finite number of SOS polynomials used in the approximation.

The second approach applies to cases where the relevant functions are differentiable (but not necessarily polynomial), which allow us to use standard optimization algorithms. Specifically, the moment constraints for low levels of our FO hierarchy have structures that enable numerical solution using algorithms like the constrained convex-concave procedure [93, 101, 107]. We can say more about the FO problem for specific levels of the hierarchy, and we omit the proofs since they follow from the definition of the constraint:

PROPOSITION 1. *The constraints in the FO problem (3) for $q = 1$ can be written as the following linear inequality constraints:*

$$\begin{aligned} B\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \Omega_i \otimes (Z_i)^{\otimes m} - \tfrac{1}{n}\sum_{i=1}^n \Omega_i \otimes \tfrac{1}{n}\sum_{i=1}^n (Z_i)^{\otimes m}\right) &\leq \Delta_{m,1} \\ -B\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \Omega_i \otimes (Z_i)^{\otimes m} - \tfrac{1}{n}\sum_{i=1}^n \Omega_i \otimes \tfrac{1}{n}\sum_{i=1}^n (Z_i)^{\otimes m}\right) &\leq \Delta_{m,1} \end{aligned} \tag{6}$$

where the inequality should be interpreted as being elementwise of the left (which is a tensor) with respect to the scalar $\Delta_{m,1}$ on the right.

This result says constraints with $q = 1$ are always convex. Thus the FO problem (3) with $\mathfrak{h} = 1$ is a convex optimization problem when R_n is convex in B. Convexity of R_n occurs in many interesting problems, including linear regression and support vector machines.

PROPOSITION 2. *The constraints in the FO problem (3) for $q = 2$ are inequalities that each involve a difference of two convex quadratic functions.*

This says constraints with $q = 2$ are always a difference of convex functions. This means that stationary points of the FO problem (3) with $\mathfrak{h} = 2$ can be found using the constrained convex-concave procedure [93, 101, 107] whenever R_n is convex in B. Recall that R_n is convex in many interesting problems like linear regression and support vector machines.

PROPOSITION 3. *If Z is a binary random variable, which is coded as either $Z \in \{0, 1\}$ or $Z \in \{\pm 1\}$, then the constraints in the FO problem (3) for $m \geq 2$ are redundant with the corresponding constraint for $m = 1$.*

This result says that when Z is binary, then the hierarchy simplifies and we only need to consider applying the level-$(1, \mathfrak{h})$ FO problems. We will use this simplification when conducting numerical experiments in Section 7.

The third approach applies when the relevant functions are mixed-integer non-convex quadratic-representable, which means the objective and constraints can be represented by non-convex quadratic functions with some variables constrained to be integer-valued. As described in Example 4, this case holds for linear classification using the 0-1 classification loss.

PROPOSITION 4. *Suppose Assumptions 1–3 hold. If, in the notation of Assumption 3, we assume that the functions $f_n : \mathbb{R}^{d\times p} \to \mathbb{R}$ and $g_n : \mathbb{R}^{d\times p} \to \mathbb{R}^\eta$ are mixed-integer non-convex quadratic-representable, then the level-$(\mathfrak{g}, \mathfrak{h})$ FO problem (3) can be solved using a non-convex mixed-integer quadratically constrained program (non-convex MIQCP).*

The development of numerical algorithms to solve non-convex MIQCP problems is an active research area [20, 54, 25], and software packages [89, 1, 19, 64, 102, 48] are already available for solving such problems. The proof of the above result is omitted because it follows immediately from the facts that the moment constraints are polynomials and that any polynomial inequality constraint can be represented by quadratic constraints and a set of new variables. To understand the intuition behind this second fact, consider as an example the constraint ${B_1}^3 \leq 0$. We can represent this by two non-convex quadratic constraints: $B_2 = {B_1}^2$ and $B_1 \cdot B_2 \leq 0$, where we have introduced a new variable B_2.

5. Statistical Consistency of FO Hierarchy. We prove in this section that the sample-based constraints of the FO problem (3) are in fact statistically well-behaved analogs of the independence constraint in (1). Here, we consider the case of bounded random variables:

ASSUMPTION 4. The entries of the random variables X, Z are almost surely bounded by $\alpha \geq 1$. Moreover, the maximal monomial degree of entries in $\omega(x, z)$ is $\rho \geq 1$, and the random variable Z has dimensions $Z \in \mathbb{R}^r$.

5.1. *Concentration of Tensor Moment Estimates.* We begin by defining several multilinear operators. We define the empirical operators $\widehat{\varphi}_{m,q}(B_1, \ldots, B_q) = \mathbb{E}_n(Z^{\otimes m} \bigotimes_{k=1}^{q}(B_k\Omega))$ and $\widehat{\nu}_{m,q}(B_1, \ldots, B_q) = \mathbb{E}_n(Z^{\otimes m}) \otimes \mathbb{E}_n(\bigotimes_{k=1}^{q}(B_i\Omega))$, and we also define the expected operators $\varphi_{m,q}(B_1, \ldots, B_q) = \mathbb{E}(Z^{\otimes m} \bigotimes_{k=1}^{q}(B_k\Omega))$ and $\nu_{m,q}(B_1, \ldots, B_q) = \mathbb{E}(Z^{\otimes m}) \otimes \mathbb{E}(\bigotimes_{k=1}^{q}(B_i\Omega))$. To simply the notation, when the argument of these multilinear operators is (B) we take that to mean the argument is $(B, \ldots, B)$. We can thus identify these operators with terms in the FO problem (3): The $\widehat{\varphi}_{m,q}(B)$ and $\widehat{\nu}_{m,q}(B)$ terms appear in the constraints.

PROPOSITION 5. *If Assumptions 1, 4 hold, then we have*

$$\begin{aligned} \mathbb{P}\big(\|\widehat{\varphi}_{m,q} - \varphi_{m,q}\|_\circ > \ \ \mathcal{R}_{m,q}[n] + \ \ \gamma\big) \ &\leq 2\exp\big(-\tfrac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}\big) \\ \mathbb{P}\big(\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ > 2\mathcal{R}_{m,q}[n] + 2\gamma\big) \ &\leq 4\exp\big(-\tfrac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}\big) \end{aligned} \tag{7}$$

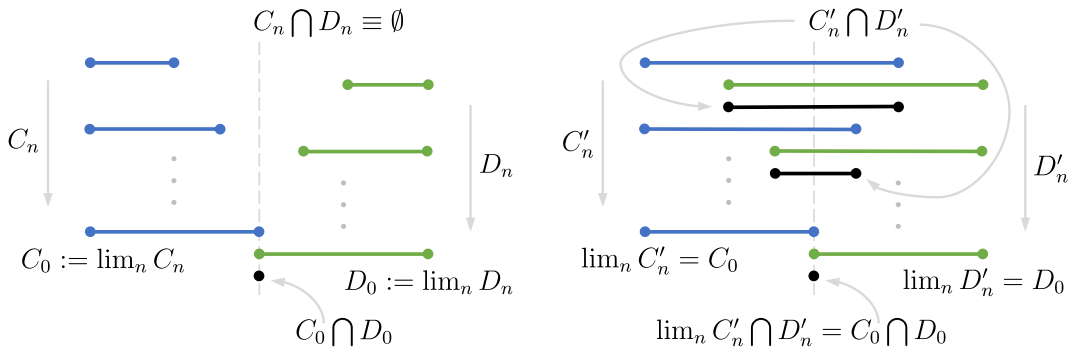
for $\mathcal{R}_{m,q}[n] = 8\alpha^{m+\rho q}p^{q/2}\sqrt{\frac{dp\log(1+4q)+m\log r+q\log d}{n}}$.

REMARK 11. Though a similar proof was used in [103] for random matrices and in [99] for random tensors, we use a stronger argument that is adapted to our setup and results in a faster convergence rate where some terms are logarithmic that would otherwise be polynomial with a weaker argument. We use a stronger chaining argument than [99, 103] by using a telescoping sum that reduces cross terms. We use a tensor symmetrization construction that allows us to exploit Banach's theorem [9, 17]. We achieve better constants than [103] by more carefully bounding our moment series expansion.

5.2. *Feasible Set Consistency.* We are now in a position to study the constraints of the FO problem (3). Towards this goal, we first define $\mathcal{S} = \{B \in \mathcal{B} : B\Omega \perp\!\!\!\perp Z\}$. This is the feasible set of (1), which chooses an optimal fair decision rule when the underlying distributions are exactly known, for a decision rule that satisfies Assumption 1. We next define the family of random sets $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \{B \in \mathcal{B} : \|\widehat{\varphi}_{m,q}(B) - \widehat{\nu}_{m,q}(B)\| \leq \Delta_{m,q}, \text{for } (m,q) \in [\mathfrak{g}] \times [\mathfrak{h}]\}$. This is simply the feasible set of the level-$(\mathfrak{g}, \mathfrak{h})$ FO problem (3).

PROPOSITION 6. *$\mathcal{S}$ and $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ are closed, under Assumptions 1, 4.*

The sequence of random sets $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ is difficult to study because each random set is defined by the intersection of many random constraint inequalities, with the number of these random constraints increasing towards infinity. A more subtle technical difficulty also needs to be addressed: The issue is that when intersecting a sequence of sets, the intersection of the sequence terms generally does not converge to the intersection of the limiting sets [5, 70]. The next example demonstrates this phenomenon in a deterministic setting, and it provides insight into how to address this situation through a carefully designed regularization approach.



(a) Unregularized Set Intersections (b) Regularized Set Intersections

Fig 1: The left shows how the intersection of a sequence of sets may not converge to the intersection of their limits. The right shows how regularization of the sequence of sets can ensure that the intersection of the regularized sets converges to the intersection of their limits.

EXAMPLE 6. Fig. 1 provides a visualization of this example. Let us first define $C_n = [-1, -\frac{1}{n}]$ and $D_n = [\frac{1}{n}, 1]$, which each specify a deterministic sequence of compact sets. Then we have that $\lim_n C_n = [-1, 0] =: C_0$ and that $\lim_n D_n = [0, 1] =: D_0$. However, note that $C_n \bigcap D_n = \emptyset$. So $\lim_n C_n \bigcap D_n = \emptyset \neq C_0 \bigcap D_0 = \{0\}$. Now suppose we carefully regularize these sequences of sets. Specifically consider the regularized sequence of deterministic, compact sets $C'_n = [-1, -\frac{1}{n} + \Delta_n]$ and $D'_n = [\frac{1}{n} - \Delta_n, 1]$ for $\Delta_n = \frac{2}{n}$, where we think of the Δ_n as regularizing by inflating the sets. Clearly this choice of regularization goes to zero since $\lim_n \Delta_n = 0$. More importantly, we now have $C'_n \bigcap D'_n = [-\frac{1}{n}, \frac{1}{n}]$. Hence we have $\lim_n C'_n = C_0$ and $\lim_n D'_n = D_0$ with $\lim_n C'_n \bigcap D'_n = \{0\} = C_0 \bigcap D_0$.

The above example was deterministic, and it may be unclear whether such behavior occurs in our random setting. The next example demonstrates such non-convergence for $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$.

EXAMPLE 7. Consider a setting where $B \in \mathbb{R}$ and the distributions are $X \sim \mathrm{Ber}(x)$ and $Z \sim \mathrm{Ber}(z)$ with $X \perp\!\!\!\perp Z$. We assume that $x \in (0, 1)$ and $z \in (0, 1)$ to prevent degeneracies in this example. In this setup $\mathcal{S} = \mathcal{B}$. Now observe that $(Z_i)^m = Z_i$ and $(X_i)^q = X_i$ for $(m, q) \geq 1$ since $X_i, Z_i \in \{0, 1\}$. This means the $(m, q) \geq 1$ constraints in $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ for $\Delta_{m,q} = 0$ are $|(\frac{1}{n}\sum_{i=1}^n (Z_i)^m (X_i)^q - \frac{1}{n}\sum_{i=1}^n (Z_i)^q \cdot \frac{1}{n}\sum_{i=1}^n (X_i)^q) B^q| = |(\frac{1}{n}\sum_{i=1}^n Z_i X_i - \frac{1}{n}\sum_{i=1}^n Z_i \cdot \frac{1}{n}\sum_{i=1}^n X_i) B^q| = 0$. Hence $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{B}$ when $\mathcal{E}_n = \{\frac{1}{n}\sum_{i=1}^n Z_i X_i = \frac{1}{n}\sum_{i=1}^n Z_i \cdot \frac{1}{n}\sum_{i=1}^n X_i\}$ occurs, and that $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \{0\}$ otherwise. And so trivially by the definition of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ we have $\text{as-lim sup}_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{B}$. If we recall the classical setting of a 2×2 contingency table, this event $\mathcal{E}_n$ is equivalent to having exact equality between a marginal and cross-term in the contingency table. As a result, we consider a test statistic inspired by the Pearson test for independence $T_n = n \cdot \big(\mathbb{E}_n(ZX) - \mathbb{E}_n(Z)\mathbb{E}_n(X)\big)^2$. Clearly by its definition, we have that $T_n = 0$ if and only if $\mathcal{E}_n$ holds. Also, a straightforward calculation gives $\mathbb{E}(T_n) = (\frac{n-1}{n})(zx)(1 - z - x - zx)$. Note that $\mathbb{E}(T_n) > 0$ since we assumed $x, z \in (0, 1)$, and note that $\mathbb{E}(T_n)$ is monotonically increasing towards $\lim_n \mathbb{E}(T_n) = (zx)(1 - z - x - zx) > 0$. Now using McDiarmid's inequality we get for any $t > 0$ that $\mathbb{P}(\mathcal{E}_n) \leq \mathbb{P}(T_n \leq \mathbb{E}(T_n) - t) \leq \exp(-nt^2/8)$. Choosing $t = (zx)(1 - z - x - zx)/2$, the Borel-Cantelli lemma implies $\mathcal{E}_n$ cannot occur infinitely often. Hence we must have $\text{as-lim inf}_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \{0\} \not\supseteq \mathcal{S}$.

Example 6 provides the key intuition for how potential non-convergence of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$, as demonstrated in Example 7, can be resolved. If we can regularize the sets $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ by sufficiently inflating them in such a way that the amount of inflation decreases with n, then we may be able to ensure the almost sure stochastic convergence of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ to $\mathcal{S}$. In fact, the notation of Example 6 was chosen to be suggestive of how we will perform this regularization: We will purposefully keep the $\Delta_{m,q} > 0$ while allowing them to shrink towards zero.

More broadly, the FO problem (3) has two types of tuning parameters, namely the $(\mathfrak{g},\mathfrak{h})$ that controls the number of moment constraints and the $\Delta_{m,q}$ that controls the strictness of the moment constraint. This gives us considerable flexibility when studying asymptotic properties. In the following results, we have to choose both of these tuning parameters.

THEOREM 4. *Suppose* $\Delta_{m,q} = 3(1+\log n)\cdot\mathcal{R}_{m,q}[n]$ *and* $\mathfrak{g} = \mathfrak{h} = O(\log n)$*, such that* $\Delta_{\mathfrak{g},\mathfrak{h}} = o(1)$*. If Assumptions 1, 2, 4 hold, then* $\text{as-lim}_n\, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{S}$.

5.3. *Solution Set Consistency.* Next consider the solution set $\widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \arg\min_B\{R_n(B\cdot\omega(x,z)) \mid B\in\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}\}$ for the level-$(\mathfrak{g},\mathfrak{h})$ FO problem (3). Similarly, consider the solution set $\mathcal{O} = \arg\min_B\left\{R(B\cdot\omega(x,z)) \;\middle|\; B\in\mathcal{S}\right\}$ for the optimization problem (1), which chooses an optimal fair decision rule when the underlying distributions are exactly known. Our next result shows that solving the FO problem (3) provides a statistically consistent approximation to solving the optimization problem (1), and we state the result using the above solutions sets.

THEOREM 5. *Suppose* $\Delta_{m,q} = 3(1+\log n)\cdot\mathcal{R}_{m,q}[n]$ *and* $\mathfrak{g} = \mathfrak{h} = O(\log n)$*, so that* $\Delta_{\mathfrak{g},\mathfrak{h}} = o(1)$*. If Assumptions 1–3, 4 hold, then* $\text{as-lim sup}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{O}$.

REMARK 12. If $\mathcal{O}$ consists of a single point, then it can be shown that $\text{as-lim}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{O}$.

The conclusion "$\text{as-lim sup}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{O}$" of the above theorem says all cluster points (i.e., convergent subsequences) as n increases of optimal solutions to the sample-based FO problem (3) belong to the set of optimal solutions to the problem (1) that we initially set out to solve using a sample-based approach. A stronger result is generally not true [84]; however, as mentioned above it can be shown that if $\mathcal{O}$ is singleton then we have $\text{as-lim}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{O}$.

5.4. *Finite Sample Bounds.* The consistency results of the previous subsection are asymptotic, and here we provide finite sample bounds that characterize this consistency. For our FO problem (3), there are two kinds of consistency to study. One kind of consistency is the usual notion of how good the sample-based optimal fair decision rule $\widehat{\delta}_n(x,z) = \widehat{B}_n\cdot\omega(x,z)$ for any $\widehat{B}_n \in \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}}$ is in terms of minimizing the risk $R(\cdot)$. The second kind of consistency is to quantify how close $\widehat{\delta}_n(X,Z) = \widehat{B}_n\Omega$ is in terms of being independent to Z.

To study the first kind of consistency, we have to strengthen Assumption 3 that says the approximate risk function composed with the parametric decision rule epi-converges almost surely. We replace this with a finite sample analog that specifies uniform convergence:

ASSUMPTION 5. Let $h_n(B)$ and $h(B)$ be the functions that are defined in Assumption 3. We assume that $\sup_{B\in\mathcal{B}}|h_n(B)-h(B)| \le r_n$ holds with probability at least $1-c_n$, where we have that $\lim_n r_n = 0$ and $\lim_n c_n = 0$.

We can now prove finite sample bounds for the FO problem (3). Recall that $\widehat{\delta}_n(x,z) = \widehat{B}_n\cdot\omega(x,z)$ for any $\widehat{B}_n \in \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}}$ is a sample-based optimal fair decision rule, and $\delta^*(x,z) = B^*\cdot\omega(x,z)$ for any $B^*\in\mathcal{O}$ is an optimal fair decision rule.

THEOREM 6. *Suppose* $\Delta_{m,q} = 3(1+\log n)\cdot \mathcal{R}_{m,q}[n]$ *and* $\mathfrak{g} = \mathfrak{h} = \kappa_1 \log n$ *(rounded down when non-integer), where* $\kappa_1 = (20p\log\alpha + 5\log p + 1)^{-1}$. *If Assumptions 1, 2, 4, 5 hold, then* $R(\widehat{\delta}_n) \le R(\delta^*) + 2r_n$, *with probability at least* $1 - 6(\kappa_1 \log n/n)^2 - 2c_n$; *and*

$$\mathbb{H}(\widehat{\delta}_n(X,Z),Z) \le e^{1/\kappa_2} n^{\kappa_1/\kappa_2} \Delta_{\mathfrak{g},\mathfrak{h}} + \tfrac{\kappa_2(r+d)}{\kappa_1 \log n + 1} \tag{8}$$

with probability at least $1 - 6(\kappa_1 \log n/n)^2$, *where* $\kappa_2 = e\alpha^\rho \lambda p$.

REMARK 13. The above theorem implies $\mathbb{H}(\widehat{\delta}_n(X,Z);Z) = O(1/\log n)$ with high probability, because $n^{\kappa_1/\kappa_2}\Delta_{\mathfrak{g},\mathfrak{h}} = o(1/\log n)$ under the conditions of the above theorem.

5.5. *Approximate Independence.* Let $U \in \mathbb{R}^p$ and $V \in \mathbb{R}^d$ be random vectors, and define

$$\mathbb{M}(U;V) = \inf\left\{\epsilon \;\middle|\; \left\|\mathbb{E}\left(U^{\otimes m}V^{\otimes q}\right) - \mathbb{E}\left(U^{\otimes m}\right)\otimes\mathbb{E}\left(V^{\otimes q}\right)\right\| \le \epsilon^{m+q}\cdot m!\cdot q!, \text{for } m,q \ge 1\right\}. \tag{9}$$

We call the quantity $\mathbb{M}(U;V)$ the *mutual majorization* of U and V, and the choice of this name is meant to draw a direct analogy to mutual information. The mutual majorization is nonnegative $\mathbb{M}(U;V) \ge 0$ and symmetric $\mathbb{M}(U;V) = \mathbb{M}(V;U)$ by definition. One utility of this definition for the mutual majorization is that it bounds approximate independence.

PROPOSITION 7. *Let* $M_{(U,V)}(s,t) = \mathbb{E}\exp(\langle s,U\rangle + \langle t,V\rangle)$ *be the moment generating function for the multivariate random variable* (U,V) *where* $U \in \mathbb{R}^p$ *and* $V \in \mathbb{R}^d$. *Suppose that* $M_{(U,V)}(s,t)$ *is finite in a neighborhood of the origin. If* $\mathbb{M}(U;V) \le \epsilon$, *then we have that* $\mathbb{H}(U;V) \le 2(\epsilon\cdot(r+d))^{2/3}$ *when* $\epsilon\cdot(r+d) \le 1$.

The implication of this result is we can use mutual majorization as a surrogate for approximate independence. We thus define an optimization problem that chooses an optimal ϵ-approximately-fair decision rule by solving

$$\delta^*(x,z) \in \arg\min_{\delta(\cdot,\cdot)} \left\{R(\delta) \;\middle|\; \mathbb{M}(\delta(X,Z);Z) \le \epsilon\right\}. \tag{10}$$

The level-$(\mathfrak{g},\mathfrak{h})$ FO problem (3) with appropriate choice of $\Delta_{m,q}$ is a statistically well-behaved, sample-based approximation of the above problem. In order to be able to discuss this, we first define the set $\mathcal{S}(\epsilon) = \{B \in \mathcal{B} : \mathbb{M}(B\Omega;Z) \le \epsilon\}$ and the solution set $\mathcal{O}(\epsilon) = \arg\min_B\{R(B\cdot\omega(x,z)) \mid B \in \mathcal{S}(\epsilon)\}$. These are respectively the feasible set and solution set of the optimization problem (10), which chooses an optimal ϵ-approximately-fair decision rule when the underlying distributions are exactly known.

THEOREM 7. *Let* $\Delta_{m,q} = \epsilon^{m+q}\cdot m!\cdot q! + 3(1+\log n)\cdot\mathcal{R}_{m,q}[n]$ *and suppose* $\mathfrak{g} = \mathfrak{h} = O(\log n)$, *such that* $\log n\cdot\mathcal{R}_{\mathfrak{g},\mathfrak{h}}[n] = o(1)$. *If Assumption 1 holds, then* $\mathcal{S}(\epsilon)$ *is closed. If Assumptions 2, 4 also hold, then* $\text{as-lim}_n\,\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{S}(\epsilon)$. *If Assumption 3 also holds, then* $\text{as-lim sup}_n\,\widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{O}(\epsilon)$.

We can also prove a finite sample version of the above result, which shows that consistency holds for sample-based analogs of (10).

THEOREM 8. *Suppose* $\Delta_{m,q} = \epsilon^{m+q}\cdot m!\cdot q! + 3(1+\log n)\cdot\mathcal{R}_{m,q}[n]$ *and* $\mathfrak{g} = \mathfrak{h} = \kappa_1\log n$ *(rounded down when non-integer), where* $\kappa_1 = (20p\log\alpha + 5\log p + 1)^{-1}$. *If Assumptions 1, 2, 4, 5 hold, then:* $R(\widehat{\delta}_n) \le R(\delta^*) + 2r_n$, *with probability at least* $1 - 6(\kappa_1\log n/n)^2 - 2c_n$; *and when* $\epsilon\cdot(r+d) \le 1$ *then we also have that*

$$\mathbb{H}(\widehat{\delta}_n(X,Z),Z) \le 2(\epsilon\cdot(r+d))^{2/3} + \kappa_3\cdot(1+\log n)\cdot\mathcal{R}_{\mathfrak{g},\mathfrak{h}}[n] + \tfrac{1}{(\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!}\cdot\kappa_4^{\mathfrak{g}+\mathfrak{h}+2} \tag{11}$$

with probability at least $1 - 6(\kappa_1\log n/n)^2$, *where* $\kappa_3 = 3\exp(1/\epsilon)$ *and* $\kappa_4 = \alpha^\rho\lambda p/\epsilon$.

REMARK 14. The above theorem implies $\limsup_n \mathbb{H}(\widehat{\delta}_n(X,Z),Z) \leq 2(\epsilon\cdot(r+d))^{2/3}$ because the second and third terms in (11) converge to zero under the specified conditions.

6. Hierarchy Consistency for Unbounded Random Variables. The underlying generalizations of Kac's Theorem, which relate moment conditions to independence, also apply to unbounded random variables whose moment generating function is finite about the origin (Theorem 1) and to unbounded random variables with some number of finite moments but not necessarily with a moment generating function that exists near the origin (Theorem 2). In this section we show that the sample-based constraints of the FO problem (3) are statistically well-behaved analogs of the independence constraint in (1) when the involved random variables are unbounded. We will consider two cases. The first is when the involved random variables are sub-Gaussian, and the second is for random variables with finite moments.

6.1. *Sub-Gaussian Case.* Our first task is to relax Assumption 4, which assumed the involved random variables are bounded. There is a subtlety in relaxing this assumption.

EXAMPLE 8. Let $X \sim \mathcal{N}(0,1)$, and define $U = X^k$ for some $k \in \mathbb{Z}_+$. Then U is sub-Gaussian for $k=1$, but U is not sub-Gaussian for $k \geq 2$. Furthermore, the moment generating function for U is finite in a neighborhood about the origin only for $k \in \{1,2,4\}$, or restated the moment generating function is not well-defined for $k=3$ or $k \geq 5$ [12].

The consequence of this example is that if we want to consider a sub-Gaussian case, then we need to specify that the joint distribution of (Z,Ω) is sub-Gaussian rather than assuming that (X,Z) is sub-Gaussian. Thus, in lieu of Assumption 4 we assume the following:

ASSUMPTION 6. The (joint) random variable (Z,Ω) is sub-Gaussian with $M \geq 1$ and $\sigma^2 \geq 0$, and the random variable Z has dimensions $Z \in \mathbb{R}^r$.

We can now study consistency of the FO problem (3) when the involved random variables are sub-Gaussian. We first prove a result on the convergence of the tensor moment estimates.

PROPOSITION 8. *If Assumptions 1, 6 hold, then we have*

$$\begin{aligned} \mathbb{P}\big(\|\widehat{\varphi}_{m,q} - \varphi_{m,q}\|_\circ > \ \ \mathcal{C}_{m,q}[n]\cdot\gamma\big) \qquad\qquad &\leq \ \ \big(\gamma^6\cdot n^2\big)^{-1} \\ \mathbb{P}\big(\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ > 2\mathcal{C}_{m,q}[n]\cdot\gamma + \mathcal{C}_{m,q}[n]^2\cdot\gamma^2\big) \ &\leq 4\big(\gamma^6\cdot n^2\big)^{-1} \end{aligned} \tag{12}$$

for $\mathcal{C}_{m,q}[n] = [\frac{e^2M^2 2^7 5^3}{\pi n}\cdot(1+4q)^{dp}(rm^3)^m(dq^3)^q(24\sigma^2/e)^{3m+3q}]^{1/6}$.

Using the above result on concentration of the moment tensors in the sub-Gaussian case, we gets results about asymptotic and finite-sample consistency of the FO problem (3).

THEOREM 9. *Suppose* $\Delta_{m,q} = 3\cdot\mathcal{C}_{m,q}[n] + \mathcal{C}_{m,q}[n]^2$, *and suppose* $\mathfrak{g} = \mathfrak{h} = O(\sqrt{\log n})$, *such that* $\Delta_{\mathfrak{g},\mathfrak{h}} = o(1)$. *If Assumption 1 holds, then* $\mathcal{S}(\epsilon)$ *is closed. If Assumptions 2, 6 also hold, then* $\text{as-lim}_n\, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \mathcal{S}(\epsilon)$. *If Assumption 3 also holds, then* $\text{as-lim sup}_n\, \widehat{\mathcal{O}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{O}(\epsilon)$.

THEOREM 10. *Suppose* $\Delta_{m,q} = 3\cdot\mathcal{C}_{m,q}[n] + \mathcal{C}_{m,q}[n]^2$, *and* $\mathfrak{g} = \mathfrak{h} = \sqrt{\kappa_5\log n}$ *(rounded down when non-integer), where* $\kappa_5 = (\max\{5, 20dp + 5\log(rd) + 30\log(24\sigma^2)\})^{-1}$. *If Assumptions 1, 2, 5, 6 hold, then:* $R(\widehat{\delta}_n) \leq R(\delta^*) + 2r_n$, *with probability at least* $1 - 6\kappa_5\log n/n^2 - 2c_n$; *and* $\mathbb{H}(\widehat{\delta}_n(X,Z),Z) \leq e^{1/\kappa_6}n^{\kappa_5/\kappa_6}\Delta_{\mathfrak{g},\mathfrak{h}} + \kappa_6(r+d)\cdot[\sqrt{\kappa_5\log n} + 1]^{-1/2}$ *for* $n \geq 3 > e$ *with probability at least* $1 - 6\kappa_5\log n/n^2$, *where* $\kappa_6 = \max\{4, 2\sqrt{e}\sigma M\}$.

REMARK 15. The above theorem implies $\mathbb{H}(\widehat{\delta}_n(X,Z);Z) = O((\log n)^{-1/4})$ with high probability, because $n^{\kappa_5/\kappa_6}\Delta_{\mathfrak{g},\mathfrak{h}} = o((\log n)^{-1/4})$ under the conditions of the above theorem.

6.2. *Finite Moments Case.* Our last set of results relax Assumption 4 to the case of unbounded random variables with finite moments. Instead of Assumptions 4 or 6, we assume:

ASSUMPTION 7. Consider the (joint) random variable (Z, Ω), and define $M_{m,q} = \sup_{(s,t)\in\mathbb{S}^{p+d-1}} \mathbb{E}(\langle s, Z\rangle^m \langle t, \Omega\rangle^q)$. Assume that any moments used in the results are finite, and the random variable Z has dimensions $Z \in \mathbb{R}^r$.

With this assumption, we can now study approximate consistency of the FO problem (3) when the involved random variables have finite moments. We first prove a result on the convergence of the tensor moment estimates.

PROPOSITION 9. *If Assumptions 1, 7 hold, then we have*

$$\begin{aligned} \mathbb{P}\big(\|\widehat{\varphi}_{m,q} - \varphi_{m,q}\|_\circ > \ \ \mathcal{Y}_{m,q}[n]\cdot\gamma\big) &\leq \ \big(\gamma^2\cdot n\big)^{-1} \\ \mathbb{P}\big(\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ > 2\mathcal{Y}_{m,q}[n]\cdot\gamma + \mathcal{Y}_{m,q}[n]^2\cdot\gamma^2\big) &\leq 4\big(\gamma^2\cdot n\big)^{-1} \end{aligned} \tag{13}$$

for $\mathcal{Y}_{m,q}[n] = (8/n)^{1/2}\cdot(M_{4m,0}\cdot M_{0,4q})^{1/4}$.

We conclude with a result about the finite sample behavior of solutions to the FO problem (3) when the involved random variables have finite moments. The difference in the hypothesis of this result, relative to the results for the cases of bounded or sub-Gaussian random variables, is that here we will characterize solutions when $\mathfrak{g}$ and $\mathfrak{h}$ are held as fixed constants. In the previous results, we assumed $\mathfrak{g}$ and $\mathfrak{h}$ were increasing with n.

THEOREM 11. *Suppose* $\Delta_{m,q} = 3\cdot\mathcal{Y}_{m,q}[n] + \mathcal{Y}_{m,q}[n]^2$, *and that* $\mathfrak{g}$ *and* $\mathfrak{h}$ *are constants. If Assumptions 1, 2, 5, 7 hold, then we have:* $R(\widehat{\delta}_n) \leq R(\delta^*) + 2r_n$, *with probability at least* $1 - 6\cdot\mathfrak{g}\cdot\mathfrak{h}/n - 2c_n$; *and* $\mathbb{H}(\widehat{\delta}_n(X,Z), Z) \leq \exp((r+d)T)\cdot\Delta_{\mathfrak{g},\mathfrak{h}} + \frac{1}{T}$ *with probability at least* $1 - 6\cdot\mathfrak{g}\cdot\mathfrak{h}/n$, *where* $T^{\mathfrak{g}+\mathfrak{h}+3} = (\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!/(\lambda^{(\mathfrak{h}+1)/2}\cdot(M_{\mathfrak{g}+1,\mathfrak{h}+1} + M_{\mathfrak{g}+1,0}\cdot M_{0,\mathfrak{h}+1}))$.

REMARK 16. The above theorem implies $\limsup_n \mathbb{H}(\widehat{\delta}_n(X,Z), Z) \leq 1/T$ with high probability, because $\Delta_{\mathfrak{g},\mathfrak{h}} = o(1)$ under the conditions of the above theorem.

7. Numerical Experiments. In this section, we implement various levels of the FO problem (3) for: classification, regression, and decision-making. In all cases, fairness is measured using disparate impact. Unless otherwise noted, all experiments were carried out using the Mosek 9 optimization package [72]. We first discuss hyperparameter selection for the FO problem, and then we describe the benchmark fairness methods that we compare our approach to. Next, we present classification and regression implementations of FO on a series of datasets from the UC Irvine Machine Learning Repository [62], the full list of which is in Table 1. Finally, we present a case study on the use of FO to perform fair morphine dosing.

7.1. *Hyperparameter Selection.* Applying the FO problem (3) to particular datasets requires choosing several hyperparameters, namely: λ, $(\mathfrak{g},\mathfrak{h})$, and $\Delta_{m,q}$. The parameter λ bounds the Euclidean norm of the model coefficients B, and it can be shown using standard duality arguments that varying λ is equivalent to controlling the amount of ℓ_2 regularization of the model coefficients. The parameters $(\mathfrak{g},\mathfrak{h})$ control the level of the FO problem, and the theory developed in previous sections says that consistency is achieved when $\mathfrak{g}$ and $\mathfrak{h}$ grow at a logarithmic or square-root-logarithmic rate. This implies that in practice small values of $(\mathfrak{g},\mathfrak{h})$ should be used. Last, the formulation in Section 5.5 suggests an approach that makes

TABLE 1
List of Datasets Used in Numerical Experiments

Dataset	p	n	Z Type	Task	Source
Arrhythmia	10	453	Binary	Classification	[49]
Biodeg	40	1055	Categorical	Classification	[69]
Communities	96	1994	Continuous	Regression	[31, 32, 33, 83]
EEG	12	4000	Binary	Regression	[41]
Energy	8	768	Continuous	Regression	[100]
German Credit	49	1000	Continuous	Classification	[62]
Letter	15	20000	Continuous	Classification	[43]
Music	68	1034	Continuous	Regression	[113]
Parkinson's	18	5875	Binary	Both	[62]
Pima Diabetes	7	768	Continuous	Classification	[92]
Recidivism	6	5278	Binary	Classification	[4]
SkillCraft	17	3338	Continuous	Classification	[98]
Statlog	35	3486	Binary	Classification	[62]
Steel	25	1941	Categorical	Classification	[62]
Taiwan Credit	22	29623	Binary	Classification	[106]
Wine Quality	11	6497	Binary	Both	[29]

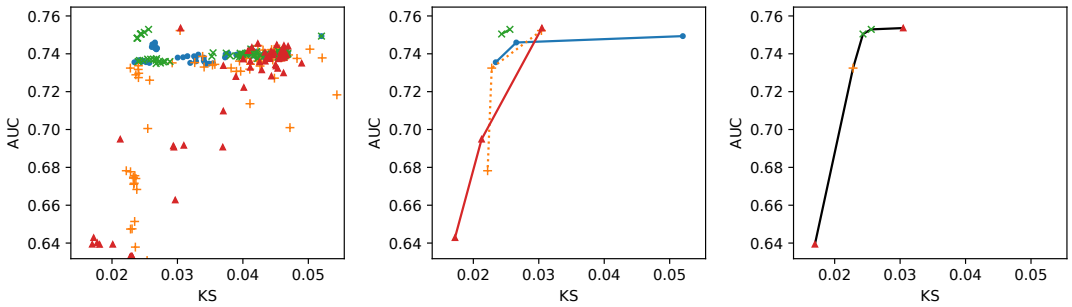
the parameter choices $\Delta_{m,q} = \epsilon^{m+q} \cdot m! \cdot q!$. This is beneficial because it replaces multiple parameters $\Delta_{m,q}$ for $(m,q) \in [\mathfrak{g}] \times [\mathfrak{h}]$ with a single parameter ϵ.

Consequently, applying the FO problem (3) to a particular dataset requires choosing four hyperparameters, which is feasible using cross-validation. However, there is a subtlety because we have two criteria to evaluate the quality of a particular model, and these two criteria are generally (but not always) opposing each other. The first criteria is model accuracy, and the second criteria is model fairness. Because these criteria are generally opposing, cross-validation can only generate a Pareto frontier, which is a curve that for a particular quantitative level of fairness specifies the most accurate model possible at that level of fairness. Choosing a particular model from among that Pareto frontier requires a subjective choice for how much reduction in model accuracy is tolerable for any given increase in model fairness. To make this discussion more concrete, we consider examples of cross-validation for fair linear regression and fair linear classification.

EXAMPLE 9. Consider a classification setup with $(X_i, Y_i) \in \mathbb{R}^p \times \{-1, +1\}$ and $Z_i \in \mathbb{R}$, and suppose we choose a linear decision rule $\delta(x) = Bx$ with $B \in \mathbb{R}^{1\times p}$. Then fair SVM using the level-(2,2) FO problem (3) is given by

$$
\begin{aligned}
&\min_{B\in\mathbb{R}^{1\times p}} \ \tfrac{1}{n}\textstyle\sum_{i=1}^n s_i \\
&\quad \text{s.t. } s_i \geq 0,\ s_i \geq 1 - Y_i \cdot BX_i, \text{ for } i \in [n] \\
&\qquad -\ \epsilon^2 \leq BM_{(1,1)} \ \leq\ \epsilon^2, \quad -2\epsilon^3 \leq BM_{(2,1)} \ \leq 2\epsilon^3 \\
&\qquad -2\epsilon^3 \leq BM_{(1,2)}B^\mathsf{T} \leq 2\epsilon^3, \quad -4\epsilon^4 \leq BM_{(2,2)}B^\mathsf{T} \leq 4\epsilon^4 \\
&\qquad \|B\|_2 \leq \sqrt{\lambda}
\end{aligned}
\tag{14}
$$

where we have the matrices $M_{(1,1)} = \frac{1}{n}\sum_{i=1}^n Z_i \cdot X_i - \frac{1}{n}\sum_{i=1}^n Z_i \cdot \frac{1}{n}\sum_{i=1}^n X_i$, $M_{(2,1)} = \frac{1}{n}\sum_{i=1}^n {Z_i}^2 \cdot X_i - \frac{1}{n}\sum_{i=1}^n {Z_i}^2 \cdot \frac{1}{n}\sum_{i=1}^n X_i$, $M_{(1,2)} = \frac{1}{n}\sum_{i=1}^n Z_i \cdot X_i X_i^\mathsf{T} - \frac{1}{n}\sum_{i=1}^n Z_i \cdot \frac{1}{n}\sum_{i=1}^n X_i X_i^\mathsf{T}$, and $M_{(2,2)} = \frac{1}{n}\sum_{i=1}^n {Z_i}^2 \cdot X_i X_i^\mathsf{T} - \frac{1}{n}\sum_{i=1}^n {Z_i}^2 \cdot \frac{1}{n}\sum_{i=1}^n X_i X_i^\mathsf{T}$. Observe that the constraint in (14) involving the matrix $M_{(m,q)}$ for any value of $(m,q) \in [2] \times [2]$ is precisely the specific form of the (m,q) constraint in (3) for this particular setup. Fair



(a) Cross-Validation (b) Level Pareto Frontiers (c) Pareto Frontier

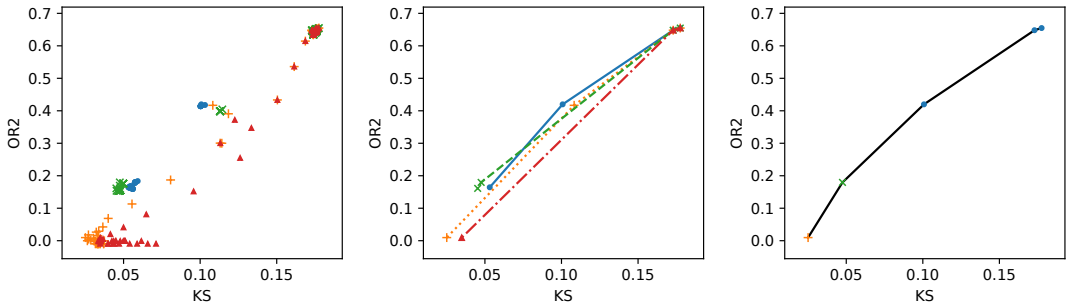
Fig 2: Pareto frontier for fair SVM on the Letter dataset. Cross-validation is used to identify points of possible tradeoff between model accuracy (measured by area under the curve) and fairness (measured by Kolmogorov-Smirnov distance between the joint and product distributions of the model prediction and the protected information) using the FO problem (left), Pareto frontiers can be constructed for each individual level of the FO problem (middle), and a single Pareto frontier can be constructed for all the levels of the FO problem (right). For the points, circles are level-(1,1), pluses are level-(1,2), exes are level-(2,1), and triangles are level-(2,2).

SVM using the level-$(\mathfrak{g},\mathfrak{h})$ FO problem for $1 \leq \mathfrak{g},\mathfrak{h} \leq 2$ is given by (14) with the appropriate constraints involving $M_{(m,q)}$ removed. An example of using five-fold cross-validation to construct a Pareto frontier for fair SVM is shown in Fig 2. For each possible value of the hyperparameters, cross-validation generates a quantitative value for model accuracy and for model fairness. These pairs of values describe points that are plotted in Fig 2a. Fig 2c shows the Pareto frontier. Locations on the Pareto frontier with a point marker can be directly achieved by a model with a given set of hyperparameters, while locations on the Pareto frontier in between two point markers can be achieved by a randomized prediction that randomly chooses from one of two deterministic predictions that arise from the two models corresponding to the two point markers.

EXAMPLE 10. Consider a regression setup with $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ and $Z_i \in \mathbb{R}$, and suppose we choose a linear decision rule $\delta(x) = Bx$ with $B \in \mathbb{R}^{1\times p}$. Then fair regression using the level-(2,2) FO problem (3) is

$$\begin{aligned} &\min_{B\in\mathbb{R}^{1\times p}} \tfrac{1}{n}\textstyle\sum_{i=1}^n (Y_i - BX_i)^2 \\ &\text{s.t. } -\ \epsilon^2 \leq BM_{(1,1)} \leq \epsilon^2, \quad -2\epsilon^3 \leq BM_{(2,1)} \leq 2\epsilon^3 \\ &\qquad -2\epsilon^3 \leq BM_{(1,2)}B^{\mathsf{T}} \leq 2\epsilon^3, \quad -4\epsilon^4 \leq BM_{(2,2)}B^{\mathsf{T}} \leq 4\epsilon^4 \\ &\qquad \|B\|_2 \leq \sqrt{\lambda} \end{aligned} \tag{15}$$

where the matrices are as in Example 9. Observe that the constraint in (15) involving the matrix $M_{(m,q)}$ for any value of $(m,q) \in [2]\times[2]$ is precisely the specific form of the (m,q) constraint in (3) for this particular setup. Fair regression using the level-$(\mathfrak{g},\mathfrak{h})$ FO problem for $1 \leq \mathfrak{g},\mathfrak{h} \leq 2$ is given by (15) with the appropriate constraints involving $M_{(m,q)}$ removed. An example of using five-fold cross-validation to construct a Pareto frontier for fair regression



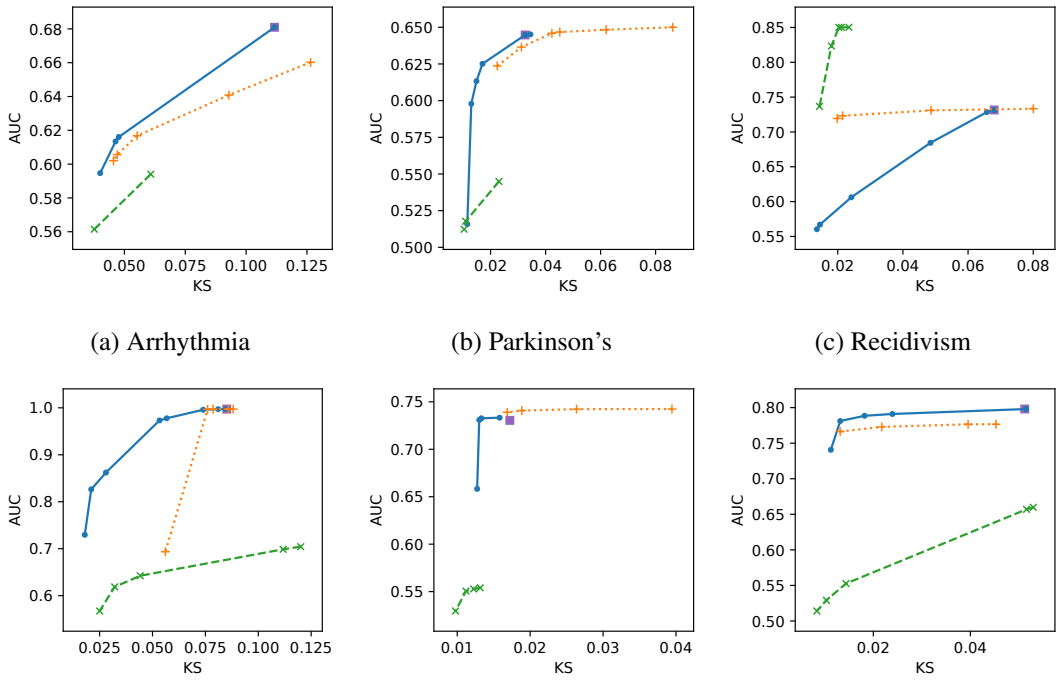
(a) Cross-Validation (b) Level Pareto Frontiers (c) Pareto Frontier

Fig 3: Pareto frontier for fair regression on the Communities dataset. Cross-validation is used to identify points of possible tradeoff between model accuracy (measured by out-of-sample R^2) and fairness (measured by Kolmogorov-Smirnov distance between the joint and product distributions of the model prediction and the protected information) using the FO problem (left), Pareto frontiers can be constructed for each individual level of the FO problem (middle), and a single Pareto frontier can be constructed for all the levels of the FO problem (right). For the points, circles are level-(1,1), pluses are level-(1,2), exes are level-(2,1), and triangles are level-(2,2).

is shown in Fig 3. For each possible value of the hyperparameters, cross-validation generates a quantitative value for model accuracy and for model fairness. These pairs of values describe points that are plotted in Fig 3a. Fig 3c shows the Pareto frontier. Locations on the Pareto frontier with a point marker can be directly achieved by a model with a given set of hyperparameters, while locations on the Pareto frontier in between two point markers can be achieved by a randomized prediction that randomly chooses from one of two deterministic predictions that arise from the two models corresponding to the two point markers.

7.2. *Comparison Methods.* In the following subsections, we compare FO to three other methods. The methods of [13] and [53] are designed for fair classification and fair regression, respectively, and are similar to our method in that they enforce fairness at training time. We also compare FO to the method of [23], although this takes a pre-processing approach. In all comparison methods, we include an ℓ_2 regularization on the model coefficients B. This is done to ensure an equitable comparison to the FO problem (3), which includes a constraint on the Euclidean norm of the model coefficients.

Berk et al. [13]. The method of [13] is one of the few comparable methods for fair regression. They also take an in-training approach, defining two regularization terms that enforce fairness. Let $P_z = \{i \in [n] : Z_i = z\}$, and note $\#P_z$ refers to the cardinality of these sets. Given a binary protected attribute Z, they define a regularizer for group fairness $((\#P_{-1} \cdot \#P_{+1})^{-1} \sum_{i \in P_{-1}} \sum_{j \in P_{+1}} d(Y_i, Y_j) \cdot (X_i^\mathsf{T}\beta - X_j^\mathsf{T}\beta))^2$, for some distance measure $d(\cdot, \cdot)$. Note that this is similar to the term constrained in FO for $(m, q) = (1, 1)$. They also define the following regularizer for individual fairness: $(\#P_{-1} \cdot \#P_1)^{-1} \sum_{i \in P_{-1}} \sum_{j \in P_{+1}} d(Y_i, Y_j) \cdot (X_i^\mathsf{T}\beta - X_j^\mathsf{T}\beta)^2$. This term is similar to a term in FO for $(m, q) = (1, 2)$, although not equivalent. It has the benefit of being convex, although the double-summation term can be computationally prohibitive for large datasets. In our implementation, we estimate this term from a sub-sample (10%) of the data when this issue arises. This method can only accommodate binary-valued protected attributes, and so we cannot provide comparisons to several of the



(a) Arrhythmia (b) Parkinson's (c) Recidivism

(d) Statlog (e) Taiwan Credit (f) Wine Quality

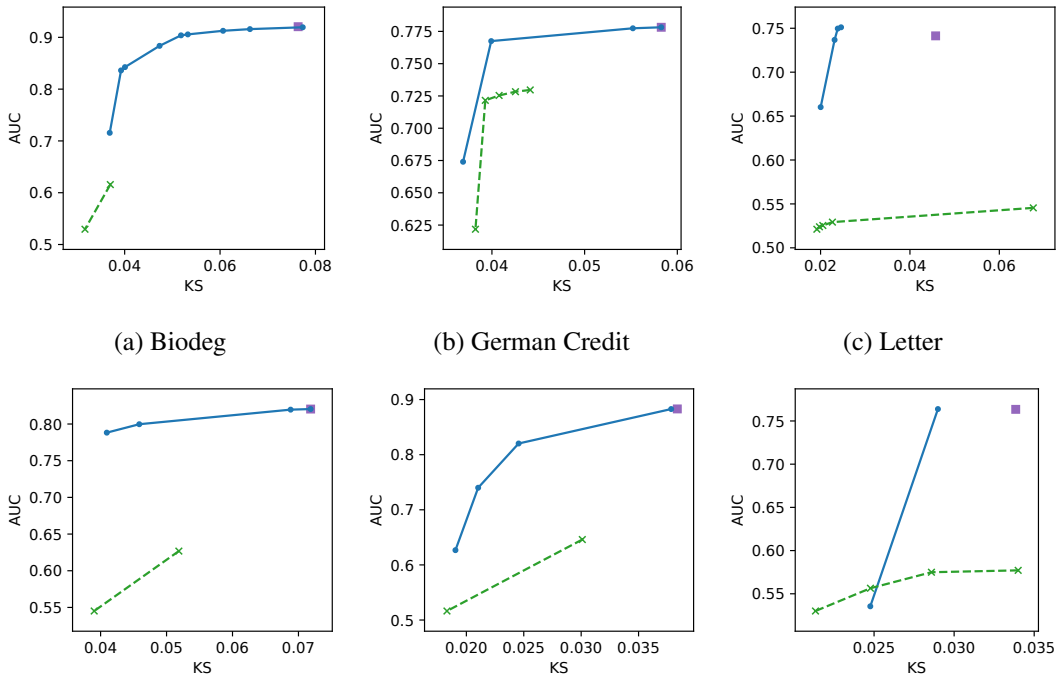
Fig 4: Pareto frontiers for fair SVM on datasets with binary protected attribute. The approaches compared are the FO formulation (solid line), Kamishima et al. [53] (dotted line), and Calmon et al. [23] (dashed line). The square mark denotes linear SVM without any fairness modifications.

datasets for fair regression. For this method, the group fairness and individual fairness terms are implemented as a penalty in the objective.

Calmon et al. [23]. This work is comparable to that of [110]. Both of these works formulate nonparametric optimization problems whose solution yields a conditional distribution $f_{\widehat{X},\widehat{Y}|X,Y,Z}$ that then probabilistically transforms the data. We only compare our method to the approach introduced in [23], since their formulation directly builds on that of [110].

This method minimizes the overall deviation of $f_{\widehat{X},\widehat{Y}}$ from $f_{X,Y}$. The authors chose to minimize $\frac{1}{2}\sum_{x,y}|f_{\widehat{X},\widehat{Y}}(x,y) - f_{X,Y}(x,y)|$, and they included constraints on pointwise distortion $\mathbb{E}_{\widehat{X},\widehat{Y}|X,Y}[\theta((X,Y),(\widehat{X},\widehat{Y})]$ for some user-defined function $\theta : \{\mathbb{R}^p \times \{\pm 1\}\}^2 \to \mathbb{R}_{\geq 0}$. There are also bounds on the dependency of the new main label $\widehat{Y}$ on the original protected label $J(f_{\widehat{Y}|Z}[y|z], f_Y(y))$, where $J(a,b) = |\frac{a}{b} - 1|$ is defined to be the probability ratio measure. Thus, the final formulation is

$$
\begin{aligned}
&\min \tfrac{1}{2}\textstyle\sum_{x,y}|f_{\widehat{X},\widehat{Y}}(x,y) - f_{X,Y}(x,y)| \\
&\text{s.t. } \mathbb{E}_{\widehat{X},\widehat{Y}|X,Y}[\theta((X,Y),(\widehat{X},\widehat{Y})|x,y] \leq c, \quad \text{for all } x,y \\
&\quad |f_Y(y)^{-1} f_{\widehat{Y}|Z}[y|z] - 1| \leq d, \qquad \text{for all } y,z \\
&\quad f_{\widehat{X},\widehat{Y}|X,Y,Z} \text{ are all distributions.}
\end{aligned} \tag{16}
$$



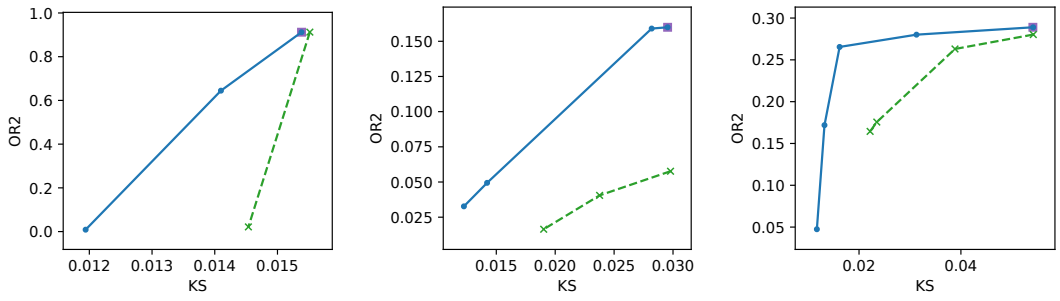
(a) Biodeg (b) German Credit (c) Letter

(d) Pima Diabetes (e) Skillcraft (f) Steel

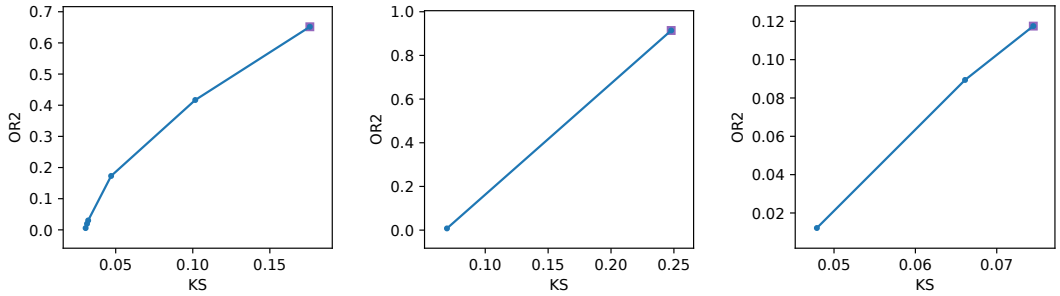
Fig 5: Pareto frontiers for fair SVM on datasets with continuous or categorical protected attributes. The approaches compared are the FO formulation (solid line) and Calmon et al. [23] (dashed line). The square mark denotes linear SVM without any fairness modifications.

Following the procedure used by the authors, we approximate $f_{X,Y,Z}$ with the empirical distribution of the original data, separated into a pre-selected number of bins. The resulting optimization problem will have $8(\#\text{bins})^{2p}$ parameters, which is computationally intractable when the dataset is high-dimensional. To account for this, we follow the original work and choose the 3 features most correlated with the main label Y. Each dimension is split into 8 bins. We choose $\theta((x',y'),(x,y))$ to be 0 if $y=y'$ and $x=x'$, 0.5 if $y=y'$ and x,x vary by at most one in any dimension, and 1 otherwise: This is similar to the θ in the original paper.

Kamishima et al. [53]. Another comparable method is that of [53], which also aims to enforce fairness at training time. As opposed to our approach of bounding interaction moments, they instead regularize with a mutual information term. Also, this method differs from our framework notably in that it imposes different treatments for different protected classes, violating the principle of individual fairness; as a result, it is also unable to handle continuous protected attributes. The authors implement their regularizer in the context of logistic regression. Let σ be a sigmoid function and $g_\beta[y|x,z] = y\sigma(x^\mathsf{T}\beta) + (1-y)(1-\sigma(x^\mathsf{T}\beta))$, and note that the notation β_z indicates that this approach has a different set of coefficients for each possible value of Z. the authors approximate mutual information as $\frac{1}{n}\sum_{i=1}^n \sum_{y\in\{\pm 1\}} g_{\beta_{Z_i}}[y|X_i,Z_i]\log(\widehat{P}[y|Z_i]/\widehat{P}(y))$, with $\widehat{P}[y|z] = (\#P_z)^{-1}\sum_{i\in P_z} g_{\beta_z}[y|X_i,z]$ and $\widehat{P}(y) = \frac{1}{n}\sum_{i=1}^n g_{\beta_{z_i}}[y|X_i,Z_i]$. This is then weighted and added to the objective as a regularizer. We include this method as a comparison to our fair SVM, while noting the core differences mentioned above. All experiments for this method were done using the sequential least squares programming approach of [58].



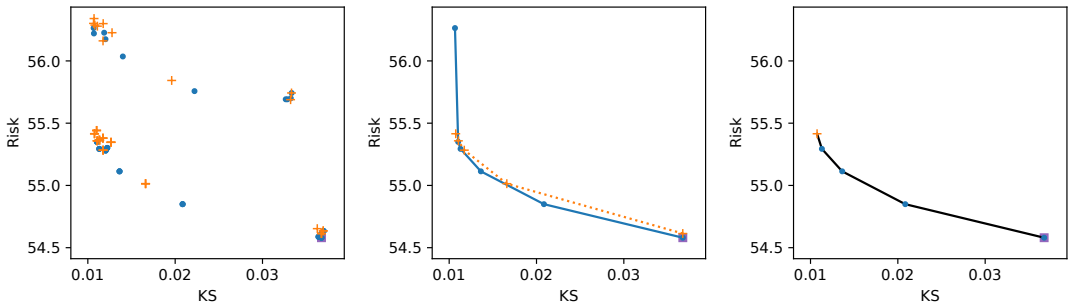
(a) EEG (b) Parkinson's (c) Wine Quality



(d) Communities (e) Energy (f) Music

Fig 6: Pareto frontiers for fair regression. The approaches compared are the FO formulation (solid line) and Berk et al. [13] (dashed line). The square mark denotes ridge regression without any fairness modifications.

7.3. *Fair SVM.* We consider classification problems using a series of datasets. For the FO approach, we consider the formulation in Example 9. We perform five-fold cross validation repeated five times. The Pareto frontiers of different approaches are shown in Fig 4 and Fig 5. Accuracy is measured by the area under the curve (AUC) since classifier models are often used as scores that are then subject to different thresholds. Fairness is measured by the Kolmogorov-Smirnov distance between the joint and product distributions of the model prediction and the protected information. The variance of the results over the five repetitions is low, and so this is not plotted to make the results easier to visualize. The 95% standard error of the results over the five repetitions can be found in the Supplementary Materials [7]. Since the mutual-information-based method of [53] cannot accommodate continuous protected variables, results are not reported for this method for the associated datasets. We note our method often improves fairness with less cost (in terms of accuracy) than the method of [23]. This is to be expected, as such pre-processing approaches do not take into account the downstream task that the transformed data is to be used for. Our method is also able to match or improve the fairness results of the mutual information approach. Recall that this method maintains explicitly different treatments for different protected classes, while ours adheres to the principle of individual fairness. Given this, it is unsurprising that the method of [53] can sometimes achieve fairness at a lower cost to accuracy, although our method even outperforms on this metric for a number of datasets. Further, this feature of disparate treatments can yield fairness values notably worse than even a standard SVM. Interestingly, for the Taiwan Credit, Letter, and Steel datasets our method can do strictly better in terms of both accuracy and fairness than linear SVM without fairness modifications.



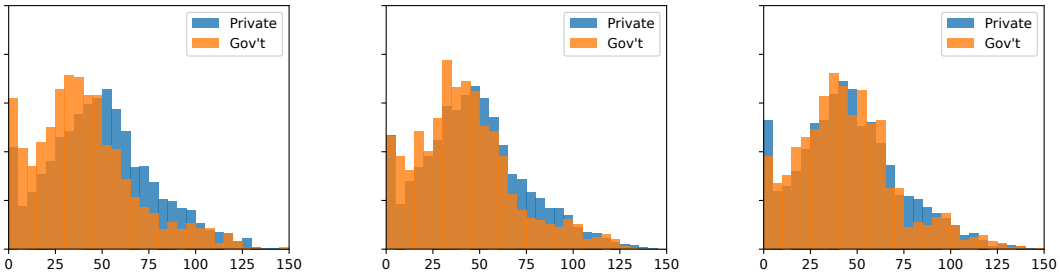
(a) Cross-Validation (b) Level Pareto Frontiers (c) Pareto Frontier

Fig 7: Pareto frontier of learned morphine dosage rules. Five-fold cross-validation repeated five times identifies points of possible tradeoff between model accuracy (measured by risk) and fairness (measured by Kolmogorov-Smirnov distance between the joint and product distributions of the learned dosage and the insurance type) using the FO problem (left), Pareto frontiers can be constructed for each individual level of the FO problem (middle), and a single Pareto frontier can be constructed for all the levels of the FO problem (right). For the points, circles are level-(1,1), pluses are level-(1,2), and the square is quantile regression without fairness modifications.

7.4. *Fair Regression.* We next consider regression problems. For the FO approach, we consider the formulation in Example 10. We perform five-fold cross validation repeated five times. The Pareto frontiers of different approaches are shown in Fig 6. Accuracy is measured by the out-of-sample R^2 (OR2), which means that higher values of OR2 implies better accuracy. Fairness is measured by the Kolmogorov-Smirnov distance between the joint and product distributions of the model prediction and the protected information.The variance of the results over the five repetitions is low, and so this is not plotted to make the results easier to visualize. The 95% standard error of the results over the five repetitions can be found in the Supplementary Materials [7]. As the method of [13] is unable to accommodate non-binary protected attributes, we only provide results for the appropriate datasets. Again, we note that our method can reduce the bias of a typical linear regression problem. Our method generally does better than [13] on datasets where [13] can be applied.

7.5. *Case Study: Morphine Dosing.* Opioid overdoses, including from illicit heroine and synthetic fentanyl, have become the leading cause of death in Americans under 50 [90]. Today, Americans comprise 4.6% of the global population, but 51.2% of global morphine usage. Hence there has been much recent interest in disciplined methods for dosing [68]. At the same time, recent reports suggest women and low-income patients are more likely to be under-diagnosed for pain or made to wait longer for a diagnosis [15, 36]. Thus, we seek to employ FO in order to train an individualized dosing policy that adapts to each patient's measurements and status, but can be made certifiably fair with regards to protected labels.

We extracted data for 7156 morphine prescriptions made to 4612 unique patients extracted from the publicly-available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC III) database [88]. For each patient, we collected age (at the time of prescription), heart rate, breath rate, blood pressure (both systolic and diastolic), weight and temperature. In all cases, measurements are the latest possible within 48 hours of prescription. We also collect, as categorical variables, admission type (ER, urgent care or other), service type (surgery or medical), ethnicity (black, white or other), gender (male or female) and insurance type (private or



(a) Quantile regression (b) level-(1,1) FO (c) level-(1,2) FO

Fig 8: Distributions of morphine dosage, conditional on insurance type, for varying levels of FO. Histograms are shown of morphine dosages recommended by rules generated using quantile regression with no fairness modifications (left), the level-(1,1) FO problem (center), and the level-(1,2) FO problem. These histograms are generated by combining the recommended dosages for the hold-out data when doing five-fold cross-validation repeated five times. These results show that increasing levels of the FO problem can yield more similar distributions. Note that negative dosages recommendations are replaced with zero.

governmental). We also note the presence of embolism or obesity amongst the diagnoses of the patients at admission. We exclude all patients who are not prescribed Morphine Sulfate to be taken intravenously, and all patients for whom the appropriate measurements were not available. Since there are medical justifications for the consideration of gender and ethnicity in opioid dosing, we decide to instead consider insurance type as our protected variable in this analysis. To begin, we conduct a standard linear regression to determine if insurance type does currently play a role in, or is at least highly correlated with, morphine dosage, conditional on all other variables considered. The results found that insurance type had a large magnitude coefficient with $p < 0.001$, which provides some statistical evidence that insurance type is correlated to dosing even after adjusting for the other predictor variables.

One possible risk function for dosing is analogous to the newsvendor problem from the operations research community, where supply must be chosen beforehand to meet random demand and undersupply/oversupply are penalized differently. Recent work formulated a data-driven newsvendor model, where demand is predicted via a quantile regression problem [87]. Similarly, we can treat dosage as a matter of supply, with demand being the amount of medication that a specific patient needs. In our case, we consider decision rules of the form $\delta(x) = Bx$ and impose a linearly increasing cost to both under-prescription and over-prescription $R_n(\delta) = \frac{1}{n}\sum_{i=1}^{n} \max\{0, \delta(X_i) - Y_i\} + 2\max\{0, -(\delta(X_i) - Y_i)\}$, where the cost to over-prescription increases half as quickly as that of under-prescription. The nondifferentiability of this loss is easily handled by introducing the slack variables s_i, t_i and noting that $R_n(\delta) = \frac{1}{n}\sum_{i=1}^{n}(s_i + 2 \cdot t_i)$ subject to the constraints $s_i \geq 0$, $s_i \geq \delta(X_i) - Y_i$, $t_i \geq 0$, and $t_i \geq -(\delta(X_i) - Y_i)$. This reflects the short-term nature of the risks of under-prescription, and the long-term nature of the risks of over-prescription. Given the features described above (excluding insurance payer), we then formulate varying levels of our FO to solve the quantile regression problem for dosing.

Results are displayed in Fig 7 and Fig 8. In Fig 7, the tradeoff between risk and fairness is displayed, as well as the range of best possible dosage rules. Visual evidence of the reduction in disparate impact is shown in Fig 8, which presents the difference in the distribution of dosage levels across insurance types for standard Quantile Regression (QR), the level-(1,1) FO with hyperparameters that provide an intermediate tradeoff between risk and fairness,

and the level-(1,2) FO with hyperparameters that provide the maximum level of fairness achievable. There is a clear disparity between the distributions in Fig 8a, but this difference is significantly reduced in Fig 8b and even more so in Fig 8c. In fact, Fig 7 shows that an intermediate tradeoff between risk and fairness using the level-(1,1) FO problem increases risk by 0.5% while improving fairness by 45%, whereas the maximum fairness achievable by the level-(1,2) FO problem increases the risk by only 1.5% while improving fairness by 70%.

8. Conclusion. We proposed an optimization hierarchy for fair statistical decision problems, which provides a systematic approach to fair versions of hypothesis testing, decision-making, estimation, regression, and classification. We proved that higher levels of this hierarchy asymptotically impose independence between the output of the decision rule and the protected variable as a constraint in corresponding statistical decision problems. We demonstrated numerical effectiveness of our hierarchy using several data sets. An important question that remains to be answered is how to tune the hyperparameters in our hierarchy. Our theoretical results provide some guidance on how to choose the level of the hierarchy and how to reduce the number of tuning parameters to just one. However, further theoretical and empirical study is needed to better understand the tuning process.

Funding. This material is based upon work supported by the NSF under Grant CMMI-1847666, and by the UC Berkeley Center for Long-Term Cybersecurity.

SUPPLEMENTARY MATERIAL

Proofs Supplement
Proofs of all the theoretical results (i.e., propositions and theorems) in this manuscript.

Standard Errors of Pareto Frontiers
Excel spreadsheets containing 95% standard errors of the points shown on the Pareto frontiers in Fig 4, Fig 5, and Fig 6.

REFERENCES

[1] ADJIMAN, C. S., ANDROULAKIS, I. P. and FLOUDAS, C. A. (2000). Global optimization of mixed-integer nonlinear problems. *AIChE Journal* **46** 1769–1797.
[2] AGARWAL, A., BEYGELZIMER, A., DUDÍK, M., LANGFORD, J. and WALLACH, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning* 60–69.
[3] AGARWAL, A., DUDÍK, M. and WU, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*.
[4] ANGWIN, J., LARSON, J., MATTU, S. and KIRCHNER, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* **23**.
[5] ASWANI, A. (2019). Statistics with set-valued functions: applications to inverse approximate optimization. *Mathematical Programming* **174** 225–251.
[6] ASWANI, A. and OLFAT, M. (2022). Proofs Supplement to "Optimization Hierarchy for Fair Statistical Decision Problems".
[7] ASWANI, A. and OLFAT, M. (2022). Standard Errors of Pareto Frontiers Supplement to "Optimization Hierarchy for Fair Statistical Decision Problems".
[8] BAHARLOUEI, S., NOUIEHED, M., BEIRAMI, A. and RAZAVIYAYN, M. (2020). Rényi Fair Inference. In *International Conference on Learning Representations*.
[9] BANACH, S. (1938). Über homogene Polynome in (L^{2}). *Studia Mathematica* **7** 36–44.
[10] BAROCAS, S. and SELBST, A. D. (2016). Big data's disparate impact. *California Law Review* **104**.
[11] BERG, C. (1987). The multidimensional moment problem and semigroups. In *Proc. Symp. Appl. Math* **37** 110–124.
[12] BERG, C. (1988). The cube of a normal distribution is indeterminate. *The Annals of Probability* 910–913.
[13] BERK, R., HEIDARI, H., JABBARI, S., JOSEPH, M., KEARNS, M., MORGENSTERN, J., NEEL, S. and ROTH, A. (2017). A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

[14] BEUTEL, A., CHEN, J., ZHAO, Z. and CHI, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.
[15] BILLOCK, J. (2018). Pain bias: The health inequality rarely discussed. *BBC*.
[16] BISGAARD, T. M. and SASVÁRI, Z. (2006). When does E (Xk· Yl)= E (Xk)· E (Yl) imply independence? *Statistics & probability letters* **76** 1111–1116.
[17] BOCHNAK, J. and SICIAK, J. (1971). Polynomials and multilinear mappings in topological vector-spaces. *Studia Mathematica* **39** 59–76.
[18] BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80** 580–598.
[19] BURER, S. (2009). On the copositive representation of binary and continuous nonconvex quadratic programs. *Mathematical Programming* **120** 479–495.
[20] BURER, S. and LETCHFORD, A. N. (2012). Non-convex mixed-integer nonlinear programming: A survey. *Surveys in Operations Research and Management Science* **17** 97–106.
[21] CALDERS, T., KAMIRAN, F. and PECHENIZKIY, M. (2009). Building classifiers with independency constraints. In *IEEE ICDMW* 13–18.
[22] CALDERS, T., KARIM, A., KAMIRAN, F., ALI, W. and ZHANG, X. (2013). Controlling attribute effect in linear regression. In *IEEE ICDM* 71–80.
[23] CALMON, F., WEI, D., VINZAMURI, B., RAMAMURTHY, K. N. and VARSHNEY, K. R. (2017). Optimized pre-processing for discrimination prevention. In *NeuRIPS* 3992–4001.
[24] CHEN, A. and BICKEL, P. J. (2005). Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing* **53** 3625–3632.
[25] CHEN, C., ATAMTÜRK, A. and OREN, S. S. (2017). A spatial branch-and-cut method for nonconvex QCQP with bounded complex variables. *Mathematical Programming* **165** 549–577.
[26] CHIERICHETTI, F., KUMAR, R., LATTANZI, S. and VASSILVITSKII, S. (2017). Fair Clustering Through Fairlets. In *Advances in Neural Information Processing Systems* 5036–5044.
[27] CHOULDECHOVA, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.
[28] CHZHEN, E., DENIS, C., HEBIRI, M., ONETO, L. and PONTIL, M. (2020). Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems* **33** 7321–7331.
[29] CORTEZ, P., CERDEIRA, A., ALMEIDA, F., MATOS, T. and REIS, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* **47** 547–553.
[30] DEMPE, S. (2002). *Foundations of Bilevel Programming*. Springer.
[31] DEPARTMENT OF COMMERCE, BUREAU OF THE CENSUS (1992). Census of population and housing 1990 United States: Summary tape file 1a and 3a (computer files).
[32] DEPARTMENT OF JUSTICE, BUREAU OF JUSTICE STATISTICS (1992). Law enforcement and administrative statistics (computer file).
[33] DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (1995). Crime in the United States (computer file). *Source: http://www.fbi.gov/ucr/hc2004/openpage.htm*.
[34] DEVROYE, L. and WISE, G. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics* **38** 480–488.
[35] DONINI, M., ONETO, L., BEN-DAVID, S., SHAWE-TAYLOR, J. S. and PONTIL, M. (2018). Empirical risk minimization under fairness constraints. In *NeuRIPS* 2791–2801.
[36] DUSENBURY, M. (2018). "Everbody was telling me there was nothing wrong". *BBC*.
[37] DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O. and ZEMEL, R. (2012). Fairness through awareness. In *ITCS* 214–226.
[38] EDITORIAL (2016). More accountability for big-data algorithms. *Nature* **537**.
[39] EDWARDS, H. and STORKEY, A. (2015). Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.
[40] ENSIGN, D., FRIEDLER, S. A., NEVILLE, S., SCHEIDEGGER, C. and VENKATASUBRAMANIAN, S. (2017). Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*.
[41] FERNANDEZ-FRAGA, S., ACEVES-FERNANDEZ, M., PEDRAZA-ORTEGA, J. and TOVAR-ARRIAGA, S. (2018). Feature Extraction of EEG Signal upon BCI Systems Based on Steady-State Visual Evoked Potentials Using the Ant Colony Optimization Algorithm. *Discrete Dynamics in Nature and Society* **2018**.
[42] FEUERVERGER, A. and MUREIKA, R. A. (1977). The empirical characteristic function and its applications. *The annals of Statistics* **5** 88–97.
[43] FREY, P. W. and SLATE, D. J. (1991). Letter recognition using Holland-style adaptive classifiers. *Machine learning* **6** 161–182.
[44] GOH, G., COTTER, A., GUPTA, M. and FRIEDLANDER, M. P. (2016). Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems* 2415–2423.

[45] GOUIC, T. L., LOUBES, J.-M. and RIGOLLET, P. (2020). Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*.

[46] GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT* 63–77.

[47] GUNTUBOYINA, A. (2012). Optimal rates of convergence for convex set estimation from support functions. *The Annals of Statistics* **40** 385–411.

[48] GUROBI OPTIMIZATION, L. (2020). Gurobi Optimizer Reference Manual.

[49] GUVENIR, H. A., ACAR, B., DEMIROZ, G. and CEKIN, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology 1997* 433–436.

[50] HARDT, M., PRICE, E. and SREBRO, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* 3315–3323.

[51] JOHNSON, K. D., FOSTER, D. P. and STINE, R. A. (2016). Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528*.

[52] KAC, M. (1936). Sur les fonctions indépendantes (I)(Propriétés générales). *Studia Mathematica* **6** 46–58.

[53] KAMISHIMA, T., AKAHO, S., ASOH, H. and SAKUMA, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *ECML-PKDD* 35–50.

[54] KILINÇ-KARZAN, F. and YILDIZ, S. (2015). Two-term disjunctions on the second-order cone. *Mathematical Programming* **154** 463–491.

[55] KITAGAWA, T., SAKAGUCHI, S. and TETENOV, A. (2021). Constrained classification and policy learning. *arXiv preprint arXiv:2106.12886*.

[56] KLEBANOV, L. and MKRTCHYAN, S. (1984). An estimate of the nearness of the distributions in terms of the nearness of their characteristic functions on a finite interval. *Journal of Soviet Mathematics* **25** 1181–1186.

[57] KOROSTELËV, A., SIMAR, L. and TSYBAKOV, A. (1995). Efficient estimation of monotone boundaries. *The Annals of Statistics* 476–489.

[58] KRAFT, D. (1988). A software package for sequential quadratic programming. *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*.

[59] LASSERRE, J.-B. (2010). *Moments, positive polynomials and their applications*. World Scientific.

[60] LEHMANN, E. L. and ROMANO, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

[61] LI, Z., PEREZ-SUAY, A., CAMPS-VALLS, G. and SEJDINOVIC, D. (2019). Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *arXiv preprint arXiv:1911.04322*.

[62] LICHMAN, M. (2013). UCI Machine Learning Repository.

[63] LIITTSCHWAGER, J. M. and WANG, C. (1978). Integer Programming Solution of a Classification Problem. *Management Science* **24** 1515–1525.

[64] LIN, Y. and SCHRAGE, L. (2009). The global solver in the LINDO API. *Optimization Methods & Software* **24** 657–668.

[65] LIU, L. T., DEAN, S., ROLF, E., SIMCHOWITZ, M. and HARDT, M. (2018). Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*.

[66] MADRAS, D., CREAGER, E., PITASSI, T. and ZEMEL, R. (2018). Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

[67] MAJUMDAR, A., VASUDEVAN, R., TOBENKIN, M. M. and TEDRAKE, R. (2014). Convex optimization of nonlinear feedback controllers via occupation measures. *The International Journal of Robotics Research* **33** 1209–1230.

[68] MANCHIKANTI, L., KAYE, A. M., KNEZEVIC, N. N., MCANALLY, H., SLAVIN, K., TRESCOT, A. M. and HIRSCH, J. (2017). Responsible, safe, and effective prescription of opioids for chronic non-cancer pain: American Society of Interventional Pain Physicians (ASIPP) guidelines. *Pain Physician* **20** S3–S92.

[69] MANSOURI, K., RINGSTED, T., BALLABIO, D., TODESCHINI, R. and CONSONNI, V. (2013). Quantitative structure–activity relationship models for ready biodegradability of chemicals. *Journal of chemical information and modeling* **53** 867–878.

[70] MATHERON, G. (1975). *Random sets and integral geometry*. John Wiley & Sons.

[71] MOLCHANOV, I. (2006). *Theory of random sets*. Springer Science & Business Media.

[72] MOSEK, APS (2002). The MOSEK Optimization Tools Version 3.2 (Revision 8) User's Manual and Reference.

[73] MUANDET, K., FUKUMIZU, K., SRIPERUMBUDUR, B. and SCHÖLKOPF, B. (2017). Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning* **10** 1–141. https://doi.org/10.1561/2200000060

[74] EXECUTIVE OFFICE OF THE PRESIDENT (2016). *Big data: A report on algorithmic systems, opportunity, and civil rights*.
[75] OLFAT, M. and ASWANI, A. (2018). Convex Formulations for Fair Principal Component Analysis. In *AAAI* 663–670.
[76] OLFAT, M. and ASWANI, A. (2018). Spectral Algorithms for Computing Fair Support Vector Machines. In *AISTATS* 1933–1942.
[77] ONETO, L., DONINI, M. and PONTIL, M. (2019). General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080*.
[78] OUATTARA, A. and ASWANI, A. (2018). Duality approach to bilevel programs with a convex lower level. In *ACC* 1388–1395.
[79] PÁL, D., PÓCZOS, B. and SZEPESVÁRI, C. (2010). Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *NeuRIPS* 1849–1857.
[80] PATSCHKOWSKI, T. and ROHDE, A. (2016). Adaptation to lowest density regions with application to support recovery. *The Annals of Statistics* **44** 255–287.
[81] PÉREZ-SUAY, A., LAPARRA, V., MATEO-GARCÍA, G., MUÑOZ-MARÍ, J., GÓMEZ-CHOVA, L. and CAMPS-VALLS, G. (2017). Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 339–355. Springer.
[82] RACHEV, S. T., KLEBANOV, L., STOYANOV, S. V. and FABOZZI, F. (2013). *The Methods of Distances in the Theory of Probability and Statistics*. Springer New York.
[83] REDMOND, M. and BAVEJA, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* **141** 660–678.
[84] ROCKAFELLAR, R. J.-B. R TYRRELL & WETS (2009). *Variational analysis* **317**. Springer Science & Business Media.
[85] ROYSET, J. O. and WETS, R. J.-B. (2019). Variational analysis of constrained M-estimators. *Annals of Statistics*. Accepted.
[86] ROZO, E. and RYKOFF, E. S. (2014). redMaPPer II: X-ray and SZ performance benchmarks for the SDSS catalog. *The Astrophysical Journal* **783** 80.
[87] SACHS, A.-L. (2015). The data-driven newsvendor with censored demand observations. In *Retail Analytics* 35–56. Springer.
[88] SAEED, M., VILLARROEL, M., REISNER, A. T., CLIFFORD, G., LEHMAN, L.-W., MOODY, G., HELDT, T., KYAW, T. H., MOODY, B. and MARK, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine* **39** 952.
[89] SAHINIDIS, N. V. (1996). BARON: A general purpose global optimization software package. *Journal of global optimization* **8** 201–205.
[90] SALAM, M. (2017). The opioid epidemic: a crisis years in the making. *The New York Times* **26**.
[91] SCHÖLKOPF, B., PLATT, J., SHAWE-TAYLOR, J., SMOLA, A. and WILLIAMSON, R. (2001). Estimating the support of a high-dimensional distribution. *Neural computation* **13** 1443–1471.
[92] SMITH, J. W., EVERHART, J., DICKSON, W., KNOWLER, W. and JOHANNES, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* 261.
[93] SMOLA, A. J., VISHWANATHAN, S. and HOFMANN, T. (2005). Kernel Methods for Missing Variables. In *AISTATS*.
[94] SONG, L., SMOLA, A., GRETTON, A. and BORGWARDT, K. M. (2007). A dependence maximization view of clustering. In *International Conference on Machine Learning* 815–822.
[95] SONG, L., SMOLA, A., GRETTON, A., BORGWARDT, K. M. and BEDO, J. (2007). Supervised feature selection via dependence estimation. In *International Conference on Machine Learning* 823–830.
[96] SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Annals of Applied Statistics* **3** 1236–1265.
[97] SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769–2794.
[98] THOMPSON, J. J., BLAIR, M. R., CHEN, L. and HENREY, A. J. (2013). Video game telemetry as a critical tool in the study of complex skill learning. *PloS one* **8** e75129.
[99] TOMIOKA, R. and SUZUKI, T. (2014). Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*.
[100] TSANAS, A. and XIFARA, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings* **49** 560–567.
[101] TUY, H. (1995). DC optimization: theory, methods and algorithms. In *Handbook of global optimization* 149–216. Springer.
[102] VIGERSKE, S. and GLEIXNER, A. (2018). SCIP: Global optimization of mixed-integer nonlinear programs in a branch-and-cut framework. *Optimization Methods and Software* **33** 563–593.

[103] WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.

[104] WOODWORTH, B., GUNASEKAR, S., OHANNESSIAN, M. I. and SREBRO, N. (2017). Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*.

[105] WU, Z., SONG, S., KHOSLA, A., YU, F., ZHANG, L., TANG, X. and XIAO, J. (2015). 3d shapenets: A deep representation for volumetric shapes. In *IEEE CVPR* 1912–1920.

[106] YEH, I.-C. and LIEN, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* **36** 2473–2480.

[107] YUILLE, A. L. and RANGARAJAN, A. (2002). The concave-convex procedure (CCCP). In *Advances in neural information processing systems* 1033–1040.

[108] ZAFAR, M. B., VALERA, I., RODRIGUEZ, M. G. and GUMMADI, K. P. (2017). Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

[109] ZAHEER, M., KOTTUR, S., RAVANBAKHSH, S., POCZOS, B., SALAKHUTDINOV, R. R. and SMOLA, A. J. (2017). Deep sets. In *Advances in neural information processing systems* 3391–3401.

[110] ZEMEL, R., WU, Y., SWERSKY, K., PITASSI, T. and DWORK, C. (2013). Learning fair representations. In *ICML* 325–333.

[111] ZHANG, B. H., LEMOINE, B. and MITCHELL, M. (2018). Mitigating unwanted biases with adversarial learning. *arXiv preprint arXiv:1801.07593*.

[112] ZHAO, P., MOHAN, S. and VASUDEVAN, R. (2019). Optimal Control of Polynomial Hybrid Systems via Convex Relaxations. *IEEE Transactions on Automatic Control*.

[113] ZHOU, F., CLAIRE, Q. and KING, R. D. (2014). Predicting the geographical origin of music. In *2014 IEEE International Conference on Data Mining* 1115–1120. IEEE.

[114] ZLIOBAITE, I. (2015). On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

[115] ZOLOTAREV, V. M. (1976). Metric Distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik* **30** 373–401.

Proofs for "Optimization Hierarchy for Fair Statistical Decision Problems"

PROOF OF THEOREM 1. Let $M_U(s) = \mathbb{E}\exp(\langle s, U\rangle)$ and $M_V(t) = \mathbb{E}\exp(\langle t, V\rangle)$ be the moment generating functions for U and V, respectively. Observe that these are defined for s and t in a neighborhood of the origin by the assumption in the hypothesis on $M_{(U,V)}(s,t)$. Our proof begins with the well-known characterization of independence using moment generating functions, that is U and V are independent if and only if $M_{(U,V)}(s,t) = M_U(s)M_V(t)$. In particular, if the "only if" condition holds then we have

$$
\begin{aligned}
M_{(U,V)}(s,t) &= \textstyle\sum_{m=0}^{\infty}\sum_{q=0}^{\infty} \frac{1}{m!\cdot q!}\cdot \mathbb{E}\big(\langle s,U\rangle^m \langle t,V\rangle^q\big) \\
&= \textstyle\sum_{m=0}^{\infty}\sum_{q=0}^{\infty} \frac{1}{m!\cdot q!}\cdot \langle \mathbb{E}\big(U^{\otimes m}V^{\otimes q}\big), s^{\otimes m}t^{\otimes q}\rangle \\
&= \textstyle\sum_{m=0}^{\infty}\sum_{q=0}^{\infty} \frac{1}{m!\cdot q!}\cdot \langle \mathbb{E}\big(U^{\otimes m}\big)\otimes \mathbb{E}\big(V^{\otimes q}\big), s^{\otimes m}t^{\otimes q}\rangle \\
&= \textstyle\sum_{m=0}^{\infty}\sum_{q=0}^{\infty} \frac{1}{m!\cdot q!}\cdot \langle \mathbb{E}\big(U^{\otimes m}\big), s^{\otimes m}\rangle \cdot \langle \mathbb{E}\big(V^{\otimes q}\big), t^{\otimes q}\rangle \\
&= \textstyle\sum_{m=0}^{\infty}\sum_{q=0}^{\infty} \frac{1}{m!\cdot q!}\cdot \mathbb{E}\big(\langle s,U\rangle^m\big)\cdot \mathbb{E}\big(\langle t,V\rangle^q\big) \\
&= \textstyle\sum_{m=0}^{\infty} \frac{1}{m!}\cdot \mathbb{E}\big(\langle s,U\rangle^m\big)\cdot \sum_{q=0}^{\infty}\frac{1}{q!}\cdot \big(\mathbb{E}\langle t,V\rangle^q\big) \\
&= M_U(s)M_V(t)
\end{aligned}
\tag{1}
$$

This proves the reverse direction. To prove the forward direction, we note it follows by applying componentwise for all $\sigma \in [p]^m$ and $\tau \in [d]^q$ the standard result that if U and V are independent, then $\mathbb{E}(\prod_{k=1}^m U_{\sigma_k}\cdot\prod_{k=1}^q V_{\tau_k}) = \mathbb{E}(\prod_{k=1}^m U_{\sigma_k})\cdot\mathbb{E}(\prod_{k=1}^q V_{\tau_k})$ when these expectations exist. Indeed, these expectations exist because of the hypothesis assumption on $M_{(U,V)}(s,t)$. □

PROOF OF THEOREM 2. We need to bound the modulus of $J(s,t,\zeta) - P(s,t,\zeta)$. As a first step, note that the difference of their Taylor polynomials satisfies

$$
\textstyle\sum_{m=0}^{\mathfrak{g}}\sum_{q=0}^{\mathfrak{h}} \frac{1}{m!\cdot q!}\cdot \mathbb{E}\big(\langle s,U\rangle^m\langle t,V\rangle^q\big) + -\sum_{m=0}^{\mathfrak{g}} \frac{1}{m!}\cdot\mathbb{E}\big(\langle s,U\rangle^m\big)\cdot\sum_{q=0}^{\mathfrak{h}}\frac{1}{q!}\cdot\big(\mathbb{E}\langle t,V\rangle^q\big) = 0 \tag{2}
$$

by the same reasoning used to show (1). We note that the above summation is well-defined because of the finiteness assumption on $J_{\mathfrak{g},\mathfrak{h}}$ in the hypothesis of this theorem. Next we apply a standard argument (see for instance Section 26 of Billingsley (1995)) that first uses Jensen's inequality and then uses the elementary inequality $|\exp(i\zeta) - \sum_{m=0}^{\mathfrak{g}}(i\zeta)^m/m!| \le |\zeta|^{\mathfrak{g}+1}/(\mathfrak{g}+1)!$ for the complex exponential. This argument implies that for $|\zeta| \le T$ we have

$$
\big|J(s,t,\zeta) - P(s,t,\zeta)\big| \le \tfrac{J_{\mathfrak{g},\mathfrak{h}}+P_{\mathfrak{g},\mathfrak{h}}}{(\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!}\cdot T^{\mathfrak{g}+\mathfrak{h}+2}. \tag{3}
$$

If we choose $T^{\mathfrak{g}+\mathfrak{h}+3} = (\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!/(J_{\mathfrak{g},\mathfrak{h}}+P_{\mathfrak{g},\mathfrak{h}})$, then the result follows by applying this bound to the definition (4). □

PROOF OF PROPOSITION 5. We use a chaining argument. Suppose $\{t_i\}_{i=1}^N$ is a $\frac{1}{2q}$ covering of $\mathbb{S}^{dp-1}$, and note $N \le (1+4q)^{dp}$ by the volume ratio bound Wainwright (2019). Define $T_i = M(t_i) \in \mathbb{R}^{d\times p}$. Let P_q be the set of all permutations of $[q]$, and let

$$
\Phi(B_1,\ldots,B_q) = \tfrac{1}{q!}\textstyle\sum_{\pi\in P_q}\big(\widehat{\varphi}_{m,q}(B_{\pi_1},\ldots,B_{\pi_q}) - \varphi_{m,q}(B_{\pi_1},\ldots,B_{\pi_q})\big). \tag{4}
$$

Observe that by construction: $\Phi(\cdot,\ldots,\cdot)$ is symmetric, and it satisfies the identity $\Phi(B) = \widehat{\varphi}_{m,q}(B) - \varphi_{m,q}(B)$. Now consider the telescoping sum

$$
\Phi(B) = \Phi(T_i) + \textstyle\sum_{k=1}^{q}\Phi(\overbrace{B,\ldots,B}^{q-k}, B-T_i, \overbrace{T_i,\ldots,T_i}^{k-1}). \tag{5}
$$

Recall $\|W(T_i)\|_2 = 1$ and $\|W(B - T_i)\|_2 \leq \frac{1}{2q}$ for $W(B) \in \mathbb{S}^{dp-1}$. Since $\|\cdot\|_*$ is a subordinate norm, we have $\|\Phi\|_\circ \leq \|\Phi(T_i)\| + \sum_{k=1}^q \frac{1}{2q}\|\Phi\|_*$. But note that $\Phi(\cdot, \ldots, \cdot)$ is symmetric, and so $\|\Phi\|_\circ = \|\Phi\|_*$ Banach (1938); Bochnak and Siciak (1971). Thus we have $\|\Phi\|_\circ \leq 2\|\Phi(T_i)\|$. But by definition of the tensor norm $\|\cdot\|$ we have

$$\|\Phi(T_i)\| = \max_{u_k, v_k} \left|\left\langle \Phi(T_i), \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k \right\rangle\right| \tag{6}$$

for $u_k \in E_r, v_k \in E_d$; where $E_d = \{x \in \{0,1\}^d : \|x\|_1 = 1\}$. So it holds that

$$\|\Phi\|_\circ \leq 2 \max_{i, u_k, v_k} \left|\left\langle \Phi(T_i), \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k \right\rangle\right| \tag{7}$$

for $i \in [N], u_k \in E_r, v_k \in E_d$. Next consider any $s \in \mathbb{R}$, and observe that

$$\begin{aligned} \mathbb{E}\exp\left(s\|\Phi\|_\circ\right) &\leq \mathbb{E}\exp\left(2s \max_{i, u_k, v_k} \left|\left\langle \Phi(T_i), \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k \right\rangle\right|\right) \\ &\leq \sum\nolimits_{\sigma \in \pm 1, i, u_k, v_k} \mathbb{E}\exp\left(2s\sigma \left\langle \Phi(T_i), \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k \right\rangle\right) \end{aligned} \tag{8}$$

We seek to bound the term on the right-hand side. Towards this end, note $\|B\Omega_i\| \leq \sqrt{p}\|W(B)\|_2\|\Omega_i\| \leq \sqrt{p}\alpha^\rho$ by the Cauchy-Schwarz inequality and Assumption 4. This means that for

$$S_i = \sigma \left\langle Z^{\otimes m}(T_i\Omega)^{\otimes q}, \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k \right\rangle \tag{9}$$

we have $\left|S_i\right| \leq \alpha^{m+\rho q} p^{q/2}$. Next observe that

$$\begin{aligned} \mathbb{E}\exp\left(2s\sigma \left\langle \Phi(T_i), \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k \right\rangle\right) &\leq \left(\mathbb{E}\exp\left(\tfrac{4\epsilon s S_i}{n}\right)\right)^n \\ &= \left(\mathbb{E}\sum\nolimits_{k=0}^\infty \tfrac{1}{k!}\left(\tfrac{4\epsilon s S_i}{n}\right)^k\right)^n \\ &= \left(\mathbb{E}\sum\nolimits_{k=0}^\infty \tfrac{1}{(2k)!}\left(\tfrac{4s S_i}{n}\right)^{2k}\right)^n \\ &\leq \left(\sum\nolimits_{k=0}^\infty \tfrac{1}{k!}\left(\tfrac{16 s^2 p^q \alpha^{2m+2\rho q}}{n^2}\right)^k\right)^n \\ &= \exp\left(\tfrac{16 s^2 p^q \alpha^{2m+2\rho q}}{n}\right) \end{aligned} \tag{10}$$

where the first line follows by a stochastic symmetrization step (i.e., Jensen's inequality, multiplication with i.i.d. Rademacher random variables ϵ having distribution $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$, using the triangle inequality, and concluded by Jensen's inequality), the third line follows since ϵ is a symmetric random variable, and the fourth line follows by replacing $(2k!)$ with $k!$ and substituting the absolute bound on $|S_i|$. Combining the above with (8) gives

$$\mathbb{E}\exp\left(s\|\Phi\|_\circ\right) \leq 2(1+4q)^{dp} r^m d^q \exp\left(\tfrac{16 s^2 p^q \alpha^{2m+2\rho q}}{n}\right). \tag{11}$$

Using the Chernoff bound gives

$$\begin{aligned} \mathbb{P}\big(\|\Phi\|_\circ > t\big) &\leq 2(1+4q)^{dp} r^m d^q \inf_{s \in \mathbb{R}} \exp\left(\tfrac{16 s^2 p^q \alpha^{2m+2\rho q}}{n} - st\right) \\ &= 2(1+4q)^{dp} r^m d^q \exp\left(-\tfrac{nt^2}{64 p^q \alpha^{2m+2\rho q}}\right) \end{aligned} \tag{12}$$

The first inequality result now follows by choosing

$$t = \sqrt{\tfrac{64 p^q \alpha^{2m+2\rho q}}{n}\left(dp\log(1+4q) + m\log r + q\log d\right) + \gamma^2} \tag{13}$$

and accordingly simplifying the resulting expression.

We cannot prove the second inequality result directly as above because $\mathbb{E}\widehat{\nu}_{m,q}(B) \neq \nu_{m,q}(B)$, whereas the above proof used the fact that $\mathbb{E}\widehat{\varphi}_{m,q}(B) = \varphi_{m,q}(B)$ in the symmetrization step of (10). We instead have to use an indirect approach to prove this result. We

begin by noting $\widehat{\varphi}_{m,0}(B) = \mathbb{E}_n(Z^{\otimes m})$, $\varphi_{m,0}(B) = \mathbb{E}(Z^{\otimes m})$, $\widehat{\varphi}_{0,q}(B) = \mathbb{E}_n((B\Omega)^{\otimes q})$, and $\varphi_{0,q}(B) = \mathbb{E}((B\Omega)^{\otimes q})$. For any $W(B) \in \mathbb{S}^{dp-1}$ we have that $\|B\Omega_i\| \leq \sqrt{p}\|W(B)\|_2\|\Omega_i\| \leq \sqrt{p}\alpha^\rho$ by the Cauchy-Schwarz inequality and Assumption 4. This means that $\|\widehat{\varphi}_{m,0}\|_\circ \leq \alpha^m$ and $\|\varphi_{0,q}\|_\circ \leq \alpha^{\rho q} p^{q/2}$. Now consider

$$\begin{aligned} \|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ &= \|\widehat{\varphi}_{m,0} \otimes \widehat{\varphi}_{0,q} - \varphi_{m,0} \otimes \varphi_{0,q}\|_\circ \\ &\leq \|\widehat{\varphi}_{m,0}\|_\circ \cdot \|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ + \|\varphi_{0,q}\|_\circ \cdot \|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ \\ &\leq \alpha^m \|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ + \alpha^{\rho q} p^{q/2} \|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ \end{aligned} \tag{14}$$

Then the union bound implies

$$\mathbb{P}\big(\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ \leq 2\mathcal{R}_{m,q}[n] + 2\gamma\big) \geq \tag{15}$$
$$1 - \mathbb{P}\big(\alpha^m \|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ > \mathcal{R}_{m,q}[n] + \gamma\big) - \mathbb{P}\big(\alpha^{\rho q} p^{q/2} \|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ > \mathcal{R}_{m,q}[n] + \gamma\big)$$

for $\mathcal{R}_{m,q}[n] = 8\alpha^{m+\rho q} p^{q/2} \sqrt{\frac{dp\log(1+4q) + m\log r + q\log d}{n}}$, which upon using the first inequality result gives the desired second inequality result. □

PROOF OF PROPOSITION 6. We first prove the result for $\mathcal{S}$. Consider any convergent sequence $B_k \in \mathbb{R}^{d\times p}$ with $B_k \in \mathcal{S}$ and $\lim_k B_k = B_0$. Because of our assumptions, the hypothesis of Theorem 1 is satisfied. This theorem says for all k we have

$$\varphi_{m,q}(B_k) = \nu_{m,q}(B_k), \text{for } m, q \geq 1. \tag{16}$$

But the φ and ν are continuous since they are multilinear operators on Euclidean space. This means $\lim_k \varphi_{m,q}(B_k) = \varphi_{m,q}(B_0)$ and $\lim_k \nu_{m,q}(B_k) = \nu_{m,q}(B_0)$ for $m, q \geq 1$. As a result we have

$$\varphi_{m,q}(B_0) = \nu_{m,q}(B_0), \text{for } m, q \geq 1, \tag{17}$$

which by Theorem 1 implies $B_0 \in \mathcal{S}$. This proves that $\mathcal{S}$ is closed.

The proof for $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ is a simple modification of the above argument. Consider any convergent sequence $B_k \in \mathbb{R}^{d\times p}$ with $B_k \in \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ and $\lim_k B_k = B_0$. By definition of $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ we have for all k that

$$\big\|\widehat{\varphi}_{m,q}(B_k) - \widehat{\nu}_{m,q}(B_k)\big\| \leq \Delta_{m,q}, \text{for } (m,q) \in [\mathfrak{g}] \times [\mathfrak{h}]. \tag{18}$$

But the $\widehat{\varphi}$ and $\widehat{\nu}$ are continuous since they are multilinear operators on Euclidean space, and so the normed function $\big\|\widehat{\varphi}_{m,q}(B) - \widehat{\nu}_{m,q}(B)\big\|$ is also continuous. As a result we have

$$\big\|\widehat{\varphi}_{m,q}(B_0) - \widehat{\nu}_{m,q}(B_0)\big\| = \lim_k \big\|\widehat{\varphi}_{m,q}(B_k) - \widehat{\nu}_{m,q}(B_k)\big\| \leq \Delta_{m,q}, \text{for } m, q \geq 1. \tag{19}$$

This means $B_0 \in \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ by definition. This proves that $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ is closed. □

PROOF OF THEOREM 4. For the first part of the proof we will show $\text{as-lim inf}_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \supseteq \mathcal{S}$. Indeed, suppose this is not true. Then there exists $B_0 \in \mathcal{S}$ and an open neighborhood $\mathcal{N} \subseteq \mathcal{B}$ of B_0 such that $\mathcal{N} \bigcap \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \emptyset$ infinitely often (Theorem 4.5 of Rockafellar (2009)). We can rewrite one of these events as

$$\big\{\mathcal{N} \bigcap \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \emptyset\big\} = \bigcup_{m\in[\mathfrak{g}]} \bigcup_{q\in[\mathfrak{h}]} \big\{ \inf_{B\in\mathcal{N}} \|\widehat{\Xi}_{m,q}(B)\| > \Delta_{m,q} \big\}, \tag{20}$$

where for convenience we define the multilinear operators $\Xi_{m,q} = \varphi_{m,q} - \nu_{m,q}$, $\widehat{\Xi}_{m,q} = \widehat{\varphi}_{m,q} - \widehat{\nu}_{m,q}$, $\Phi_{m,q} = \widehat{\varphi}_{m,q} - \varphi_{m,q}$, and $\Psi_{m,q} = \widehat{\nu}_{m,q} - \nu_{m,q}$. Because Theorem 1 can be rewritten under the assumptions of this theorem as

$$\sup_{B\in\mathcal{S}} \|\varphi_{m,q}(B) - \nu_{m,q}(B)\| = 0 \text{ for } m, q \geq 1, \tag{21}$$

application of the triangle inequality yields

$$\begin{aligned}\|\widehat{\Xi}_{m,q}(B_0)\| &\leq \|\Xi_{m,q}(B_0)\| + \|\Phi_{m,q}(B_0)\| + \|\Psi_{m,q}(B_0)\| \\ &\leq \lambda^{q/2}\|\Phi_{m,q}\|_\circ + \lambda^{q/2}\|\Psi_{m,q}\|_\circ\end{aligned} \tag{22}$$

Let $\mathcal{G}_{m,q}[n] = (1+\log n)\lambda^{q/2}\mathcal{R}_{m,q}[n]$. Note that for all n sufficiently large, the union bound gives us that

$$\begin{aligned}\mathbb{P}\big(\mathcal{N}\bigcap\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \emptyset\big) \leq \textstyle\sum_{m\in[\mathfrak{g}]}\sum_{q\in[\mathfrak{h}]}\mathbb{P}\big(\lambda^{q/2}\|\Phi_{m,q}\|_\circ > \mathcal{G}_{m,q}[n]\big) + \\ \textstyle\sum_{m\in[\mathfrak{g}]}\sum_{q\in[\mathfrak{h}]}\mathbb{P}\big(\lambda^{q/2}\|\Psi_{m,q}\|_\circ > 2\mathcal{G}_{m,q}[n]\big) \\ \leq O((\log n/n)^2)\end{aligned} \tag{23}$$

where the last line used Proposition 5, along with the relation that $\exp(-\frac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}) = O(1/n^2)$ for $\gamma = \log n \cdot \mathcal{R}_{m,q}[n]$. Thus the Borel-Cantelli lemma says $\mathcal{N}\bigcap\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} = \emptyset$ only finitely many times, which is a contradiction. This proves $\text{as-lim}\inf_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \supseteq \mathcal{S}$.

For the second part of the proof we will show $\text{as-lim}\sup_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{S}$. Indeed, suppose this is not true. Then there exists $B_0 \in \limsup_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ and a closed neighborhood $\mathcal{N} \subseteq \mathcal{B}$ of B_0 such that $\mathcal{N}\bigcap\mathcal{S} = \emptyset$ and $\mathcal{N}\bigcap\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \neq \emptyset$ infinitely often (Theorem 4.5 of Rockafellar (2009)). But Theorem 1 implies there exists some $m, q \geq 1$ such that we have

$$\zeta := \inf_{B\in\mathcal{N}} \|\varphi_{m,q}(B) - \nu_{m,q}(B)\| > 0. \tag{24}$$

We will keep m, q fixed at these values for the remainder of the proof. Now note that for one of the events $\mathcal{N}\bigcap\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \neq \emptyset$ we have

$$\big\{\mathcal{N}\bigcap\mathcal{S}_{\mathfrak{g},\mathfrak{h}} \neq \emptyset\big\} \subseteq \big\{\inf_{B\in\mathcal{N}} \|\widehat{\Xi}_{m,q}(B)\| \leq \Delta_{m,q}\big\}. \tag{25}$$

Application of the triangle inequality yields

$$\begin{aligned}\zeta = \inf_{B\in\mathcal{N}}\|\Xi_{m,q}(B)\| &\leq \inf_{B\in\mathcal{N}}\|\widehat{\Xi}_{m,q}(B)\| + \sup_{B\in\mathcal{N}}\|\Phi_{m,q}(B)\| + \sup_{B\in\mathcal{N}}\|\Psi_{m,q}(B)\| \\ &\leq \inf_{B\in\mathcal{N}}\|\widehat{\Xi}_{m,q}(B)\| + \lambda^{q/2}\|\Phi_{m,q}\|_\circ + \lambda^{q/2}\|\Psi_{m,q}\|_\circ.\end{aligned} \tag{26}$$

Let $\mathcal{G}_{m,q}[n] = (1+\log n)\lambda^{q/2}\mathcal{R}_{m,q}[n]$. Note that for all n sufficiently large, we have $\zeta - \Delta_{m,q} \geq \zeta/2 \geq 3\mathcal{G}_{m,q}[n]$. Hence the union bound gives

$$\begin{aligned}\mathbb{P}\big(\mathcal{N}\bigcap\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \neq \emptyset\big) &\leq \mathbb{P}\big(\lambda^{q/2}\|\Phi_{m,q}\|_\circ > \mathcal{G}_{m,q}[n]\big) + \mathbb{P}\big(\lambda^{q/2}\|\Psi_{m,q}\|_\circ > 2\mathcal{G}_{m,q}[n]\big) \\ &\leq O(1/n^2)\end{aligned} \tag{27}$$

where the last line used Proposition 5, along with the relation that $\exp(-\frac{n\gamma^2}{64p^q\alpha^{2m+2\rho q}}) = O(1/n^2)$ for $\gamma = \log n \cdot \mathcal{R}_{m,q}[n]$. Thus the Borel-Cantelli lemma says $\mathcal{N}\bigcap\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \neq \emptyset$ only finitely many times, which is a contradiction. This proves $\text{as-lim}\sup_n \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \subseteq \mathcal{S}$. □

PROOF OF THEOREM 5. First consider the indicator function $\Gamma(B, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}})$. Combining our Theorem 4 with Proposition 7.4 of Rockafellar (2009) gives $\text{as-e-lim}\,\Gamma(\cdot, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}) = \Gamma(\cdot, \mathcal{S})$ relative to $\mathbb{R}^{d\times p}$. Next we claim $\text{as-lim}\,\Gamma(\cdot, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}) = \Gamma(\cdot, \mathcal{S})$ relative to $\mathbb{R}^{d\times p}$. Since Proposition 6 says the $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ are closed, the remark after Theorem 7.10 of Rockafellar (2009) implies it is sufficient to show that for every $B_0 \in \mathcal{S}$ we have $B_0 \notin \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}$ only a finite number of times. A

similar argument to the first part of the proof for Theorem 4 can be used to show this, and so we omit the details.

Next we note that the level-$(\mathfrak{g},\mathfrak{h})$ FO problem (3) can be written as $\min_B h_n(B) + \Gamma(B, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}})$, and the optimization problem (1) can be written as $\min_B h(B) + \Gamma(B, \mathcal{S})$. Now using Theorem 7.46 of Rockafellar (2009) gives us that

$$\text{as-e-lim}\left(h_n(\cdot) + \Gamma(\cdot, \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}})\right) = h(\cdot) + \Gamma(\cdot, \mathcal{S}). \tag{28}$$

The result now follows by direct application of Proposition 7.30 of Rockafellar (2009). □

PROOF OF THEOREM 6. We begin by bounding the probability that $\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \supseteq \mathcal{S}$. Observe that we can rewrite the complement of this event as

$$\{\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \not\supseteq \mathcal{S}\} = \bigcup_{m\in[\mathfrak{g}]} \bigcup_{q\in[\mathfrak{h}]} \{ \sup_{B\in\mathcal{S}} \|\widehat{\Xi}_{m,q}(B)\| > \Delta_{m,q}\}, \tag{29}$$

where for convenience we define the multilinear operators $\Xi_{m,q} = \varphi_{m,q} - \nu_{m,q}$, $\widehat{\Xi}_{m,q} = \widehat{\varphi}_{m,q} - \widehat{\nu}_{m,q}$, $\Phi_{m,q} = \widehat{\varphi}_{m,q} - \varphi_{m,q}$, and $\Psi_{m,q} = \widehat{\nu}_{m,q} - \nu_{m,q}$. Because Theorem 1 can be rewritten under the assumptions of this theorem as

$$\sup_{B\in\mathcal{S}} \|\varphi_{m,q}(B) - \nu_{m,q}(B)\| = 0 \text{ for } m, q \geq 1, \tag{30}$$

then for any $B \in \mathcal{S}$ the application of the triangle inequality yields

$$\begin{aligned} \|\widehat{\Xi}_{m,q}(B)\| &\leq \|\Xi_{m,q}(B)\| + \|\Phi_{m,q}(B)\| + \|\Psi_{m,q}(B)\| \\ &\leq \lambda^{q/2}\|\Phi_{m,q}\|_\circ + \lambda^{q/2}\|\Psi_{m,q}\|_\circ \end{aligned} \tag{31}$$

Let $\mathcal{G}_{m,q}[n] = (1+\log n)\lambda^{q/2}\mathcal{R}_{m,q}[n]$, and note that the union bound gives

$$\begin{aligned} \mathbb{P}\big(\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \not\supseteq \mathcal{S}\big) &\leq \textstyle\sum_{m\in[\mathfrak{g}]} \sum_{q\in[\mathfrak{h}]} \mathbb{P}\big(\lambda^{q/2}\|\Phi_{m,q}\|_\circ > \mathcal{G}_{m,q}[n]\big) + \\ &\qquad \textstyle\sum_{m\in[\mathfrak{g}]} \sum_{q\in[\mathfrak{h}]} \mathbb{P}\big(\lambda^{q/2}\|\Psi_{m,q}\|_\circ > 2\mathcal{G}_{m,q}[n]\big) \\ &\leq 6(\kappa_1 \log n/n)^2 \end{aligned} \tag{32}$$

where the last line used Proposition 5. This implies $\mathbb{P}(\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \supseteq \mathcal{S}) \geq 1 - 6(\kappa_1 \log n/n)^2$, which means that $\mathbb{P}(h_n(\widehat{B}_n) \leq h_n(B^*) \text{ for all } B^* \in \mathcal{O}) \geq \mathbb{P}(\mathcal{O} \subseteq \widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}}) \geq \mathbb{P}(\widehat{\mathcal{S}}_{\mathfrak{g},\mathfrak{h}} \supseteq \mathcal{S}) \geq 1 - 6(\kappa_1 \log n/n)^2$. Combining this with Assumption 5 implies $R(\widehat{\delta}_n) \leq R(\delta^*) + 2r_n$, with probability at least $1 - 6(\kappa_1 \log n/n)^2 - 2c_n$. This proves the first part of the result.

We prove the second part of the result in two steps. As the first step, we consider the event

$$\mathcal{E} = \bigcup_{m\in[\mathfrak{g}]} \bigcup_{q\in[\mathfrak{h}]} \{ \sup\nolimits_{\widehat{B}_n \in \widehat{O}_{\mathfrak{g},\mathfrak{h}}} \|\Xi_{m,q}(\widehat{B}_n)\| > 2\Delta_{m,q}\}, \tag{33}$$

and note that for $\widehat{B}_n \in \widehat{O}_{\mathfrak{g},\mathfrak{h}}$ application of the triangle inequality yields

$$\begin{aligned} \|\Xi_{m,q}(\widehat{B}_n)\| &\leq \|\widehat{\Xi}_{m,q}(\widehat{B}_n)\| + \|\Phi_{m,q}(\widehat{B}_n)\| + \|\Psi_{m,q}(\widehat{B}_n)\| \\ &\leq \Delta_{m,q} + \lambda^{q/2}\|\Phi_{m,q}\|_\circ + \lambda^{q/2}\|\Psi_{m,q}\|_\circ \end{aligned} \tag{34}$$

since $\|\widehat{\Xi}_{m,q}(\widehat{B}_n)\| \leq \Delta_{m,q}$ by definition of $\widehat{O}_{\mathfrak{g},\mathfrak{h}}$. Thus the union bound gives

$$\begin{aligned} \mathbb{P}\big(\mathcal{E}\big) &\leq \textstyle\sum_{m\in[\mathfrak{g}]} \sum_{q\in[\mathfrak{h}]} \mathbb{P}\big(\lambda^{q/2}\|\Phi_{m,q}\|_\circ > \mathcal{G}_{m,q}[n]\big) + \\ &\qquad \textstyle\sum_{m\in[\mathfrak{g}]} \sum_{q\in[\mathfrak{h}]} \mathbb{P}\big(\lambda^{q/2}\|\Psi_{m,q}\|_\circ > 2\mathcal{G}_{m,q}[n]\big) \\ &\leq 6(\kappa_1 \log n/n)^2 \end{aligned} \tag{35}$$

where the last line used Proposition 5. This implies $\mathbb{P}(\|\Xi_{m,q}(\widehat{B}_n)\| \leq 2\Delta_{m,q}$ for $(m,q) \in [\mathfrak{g}] \times [\mathfrak{h}]) \geq 1 - 6(\kappa_1 \log n/n)^2$.

We conclude with the second step for our proof of the second part of the result. Because our random variables are bounded, we can use series expansions to express the characteristic functions in the definition (4) of $\mathbb{H}(\widehat{B}_n\Omega; Z)$. In particular, we have that

$$J(s,t,\zeta) - P(s,t,\zeta) = \textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} \frac{(\mathrm{i}\zeta)^{m+q}}{m!\cdot q!} \cdot \langle \Xi_{m,q}(\widehat{B}_n), s^{\otimes m} t^{\otimes q}\rangle. \tag{36}$$

We need to bound the modulus of the above. Hölder's inequality gives us that

$$|\langle \Xi_{m,q}(\widehat{B}_n), s^{\otimes m} t^{\otimes q}\rangle| \leq (r^m + d^q)^{1/2} \big\|\Xi_{m,q}(\widehat{B}_n)\big\| \leq (r+d)^{(m+q)} \|\Xi_{m,q}(\widehat{B}_n)\|. \tag{37}$$

In the proof of Proposition 5 we showed $\|\psi_{m,q}(\widehat{B}_n)\|_\circ \leq \alpha^{m+\rho q} p^{q/2}$ and $\|\nu_{m,q}(\widehat{B}_n)\|_\circ \leq \alpha^{m+\rho q} p^{q/2}$. Thus $\|\Xi_{m,q}(\widehat{B}_n)\| \leq 2\alpha^{m+\rho q}(\lambda p)^{q/2}$, which we will use for $m = \mathfrak{g}+1$ and $q = \mathfrak{h}+1$. We next use these bounds with a standard argument (see for instance Section 26 of Billingsley (1995)) that first uses Jensen's inequality and then uses the elementary inequality for the complex exponential that $|\exp(i\zeta) - \sum_{m=0}^{\mathfrak{g}} (i\zeta)^m/m!| \leq |\zeta|^{\mathfrak{g}+1}/(\mathfrak{g}+1)!$. This two step argument implies that for $|\zeta| \leq T$ we have

$$\begin{aligned} &\big|J(s,t,\zeta) - P(s,t,\zeta) - \textstyle\sum_{m=1}^{\mathfrak{g}}\sum_{q=1}^{\mathfrak{h}} \frac{(\mathrm{i}\zeta)^{m+q}}{m!\cdot q!} \cdot \langle \Xi_{m,q}(\widehat{B}_n), s^{\otimes m} t^{\otimes q}\rangle\big| \\ &\qquad \leq \tfrac{2}{(\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!} \cdot \alpha^{\mathfrak{g}+1+\rho(\mathfrak{h}+1)} \cdot (\lambda p)^{(\mathfrak{h}+1)/2} \cdot ((r+d)T)^{\mathfrak{g}+\mathfrak{h}+2}. \end{aligned} \tag{38}$$

Using the reverse triangle inequality implies the modulus is bounded by

$$\begin{aligned} &\big|J(s,t,\zeta) - P(s,t,\zeta)\big| \leq \\ &\qquad \textstyle\sum_{m=1}^{\mathfrak{g}}\sum_{q=1}^{\mathfrak{h}} \frac{((r+d)\zeta)^{m+q}}{m!\cdot q!} \cdot \big\|\Xi_{m,q}(\widehat{B}_n)\big\| + \frac{2}{(\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!} \cdot (\alpha^\rho \lambda p (r+d) T)^{\mathfrak{g}+\mathfrak{h}+2} \end{aligned} \tag{39}$$

for all $|\zeta| \leq T$. Combining this with the first step of the proof for the second part of the result implies that with probability at least $1 - 6(\kappa_1 \log n/n)^2$ we have for $|\zeta| \leq T$ that

$$\big|J(s,t,\zeta) - P(s,t,\zeta)\big| \leq 2\exp((r+d)T) \cdot \Delta_{\mathfrak{g},\mathfrak{h}} + \tfrac{2(\alpha^\rho \lambda p (r+d) T)^{\mathfrak{g}+\mathfrak{h}+2}}{(\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!} \tag{40}$$

where the first term follows from the exponential series. If we choose that $T = (\kappa_1 \log n + 1)/(\kappa_2(r+d))$, then using the standard error bound $(\mathfrak{g}+1)! \geq (2\pi(\mathfrak{g}+1))^{1/2}((\mathfrak{g}+1)/e)^{\mathfrak{g}+1}$ for Stirling's approximation leads to

$$\big|J(s,t,\zeta) - P(s,t,\zeta)\big| \leq 2e^{1/\kappa_2} n^{\kappa_1/\kappa_2} \cdot \Delta_{\mathfrak{g},\mathfrak{h}} + \tfrac{1}{\pi(\kappa_1 \log n + 1)}, \tag{41}$$

which holds with probability at least $1 - 6(\kappa_1 \log n/n)^2$. The second result follows by applying this bound and choice of T to the definition (4). □

PROOF OF PROPOSITION 7. We need to bound the modulus of $J(s,t,\zeta) - P(s,t,\zeta)$. Because $M_{(U,V)}(s,t)$ exists in a neighborhood of the origin, this means the characteristic functions can be represented as infinite series. Thus we have

$$\begin{aligned} &\big|J(s,t,\zeta) - P(s,t,\zeta)\big| = \\ &\qquad \big|\textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} \frac{(\mathrm{i}\zeta)^{m+q}}{m!\cdot q!} \cdot \langle \mathbb{E}\big(U^{\otimes m} V^{\otimes q}\big) - \mathbb{E}\big(U^{\otimes m}\big) \otimes \mathbb{E}\big(V^{\otimes q}\big), s^{\otimes m} t^{\otimes q}\rangle\big| \leq \\ &\qquad\qquad \textstyle\sum_{m=1}^{\infty}\sum_{q=1}^{\infty} (\epsilon(r+d)\zeta)^{m+q} = (\tau/(1-\tau))^2. \end{aligned} \tag{42}$$

when $\tau = \epsilon(r+d)\zeta \in [0,1)$. If we choose $T^{-1} = \epsilon(r+d) + (\epsilon(r+d))^{2/3}$, then the result follows by applying this bound to the definition (4). □

PROOF OF THEOREM 7. The proof is omitted because it is a straightforward modification of the proofs for Proposition 6 and Theorems 4 and 5. □

PROOF OF THEOREM 8. The proof is identical to that of Theorem 6, up to (39). (This means the first part of the current result is proved the same way as in Theorem 6.) To complete the proof we first bound (39) using the $\Delta_{m,q}$ in the hypothesis of this theorem. Comparing to (42), we get with probability at least $1 - 6(\kappa_1 \log n/n)^2$ we have that

$$\big|J(s,t,\zeta) - P(s,t,\zeta)\big| \leq \tag{43}$$
$$2(\tau/(1-\tau))^2 + 6\exp((r+d)T)\cdot(1+\log n)\cdot\mathcal{R}_{\mathfrak{g},\mathfrak{h}}[n] + \tfrac{2(\alpha^\rho \lambda p(r+d)T)^{\mathfrak{g}+\mathfrak{h}+2}}{(\mathfrak{g}+1)!\cdot(\mathfrak{h}+1)!}$$

when $\tau = \epsilon(r+d)T \in [0,1)$ and for all $|\zeta| \leq T$. If we choose $T^{-1} = \epsilon(r+d) + (\epsilon(r+d))^{2/3}$, then the second result follows by applying this bound and choice of T to the definition (4). □

PROOF OF PROPOSITION 8. The proof for the first part of this result follows the same steps as the first part of the proof of Proposition 5 up to and including (7). Next observe that

$$\begin{aligned}\mathbb{E}\big(\|\Phi\|_\circ^{\,6}\big) &\leq \mathbb{E}\big(2^6 \max_{i,u_k,v_k} |\langle \Phi(T_i), \textstyle\bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\rangle|^6\big) \\ &\leq 2^6 \cdot \textstyle\sum_{i,u_k,v_k} \mathbb{E}\big(\langle \Phi(T_i), \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\rangle^6\big)\end{aligned} \tag{44}$$

We seek to bound the term on the right-hand side. For convenience, define the terms $S_i = \langle Z^{\otimes m}(T_i\Omega)^{\otimes q}, \bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\rangle$ and $V_i = S_i - \mathbb{E}(S_i)$. Next observe that the Marcinkiewicz-Zygmund inequality Rio (2009) implies that

$$\mathbb{E}\big(\langle \Phi(T_i), \textstyle\bigotimes_{k=1}^m u_k \bigotimes_{k=1}^q v_k\rangle^6\big) \leq 5^3 \cdot \mathbb{E}({V_i}^6)/n^3. \tag{45}$$

We next have to bound the expectation on the right. Consider

$$\begin{aligned}\mathbb{E}\big({V_i}^6\big) &\leq 2\mathbb{E}\big(\epsilon^6 {S_i}^6\big) \\ &\leq 2\cdot\big[\mathbb{E}\big(\langle u_k, Z\rangle^{12m}\big)\cdot\mathbb{E}\big(\langle v_k,\ T_i\Omega\rangle^{12q}\big)\big]^{1/2} \\ &\leq 2\cdot\big[\mathbb{E}\big(\langle u_k, Z\rangle^{12m}\big)\cdot\mathbb{E}\big(\langle T_i^{\mathsf{T}} v_k, \Omega\rangle^{12q}\big)\big]^{1/2} \\ &\leq 2M\sigma^{6m+6q}\cdot\big[\tfrac{(12m)!\cdot(12q)!}{(6m)!\cdot(6q)!}\big]^{1/2} \\ &\leq 2eM\cdot(24\sigma^2/e)^{3m+3q}\cdot m^{3m}\cdot q^{3q}/\sqrt{\pi}\end{aligned} \tag{46}$$

where the first line follows by a stochastic symmetrization step (i.e., Jensen's inequality, multiplication with i.i.d. Rademacher random variables ϵ having distribution $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$, using the triangle inequality, and concluded by Jensen's inequality), the second line follows by the Cauchy-Schwarz inequality, the third line uses a matrix transpose T_i^{T}, the fourth line follows by the $\mathbb{E}\big(\langle t, U\rangle^{2k}\big) \leq M\sigma^{2k}\cdot(2k)!/k!$ for all $t \in \mathbb{S}^{p-1}$ characterization of sub-Gaussian distributions because $\|T_i^{\mathsf{T}} v_k\|_2 \leq \|v_k\|_2$ since $T_i = M(t_i)$ for $t_i \in \mathbb{S}^{dp-1}$, and the fifth line uses Stirling's approximation. Combining the above with (44) gives

$$\mathbb{E}\big(\|\Phi\|_\circ^{\,6}\big) \leq \tfrac{eM2^7 5^3}{\sqrt{\pi}n^3}\cdot(1+4q)^{dp}(rm^3)^m(dq^3)^q(24\sigma^2/e)^{3m+3q}. \tag{47}$$

Let $\kappa = (eM/\sqrt{\pi})^{1/6}$ and note that Markov's inequality implies

$$\mathbb{P}\big(\|\widehat{\varphi}_{m,q} - \varphi_{m,q}\|_\circ > \mathcal{C}_{m,q}[n]\cdot\gamma/\kappa\big) \leq \big(\gamma^6\cdot n^2\big)^{-1}. \tag{48}$$

The first result now follows by nothing that $\kappa > 1$.

The proof for the second part of this result proceeds slightly differently than the second part of the proof of Proposition 5. Recall that we have $\widehat{\varphi}_{m,0}(B) = \mathbb{E}_n(Z^{\otimes m})$, $\varphi_{m,0}(B) = \mathbb{E}(Z^{\otimes m})$, $\widehat{\varphi}_{0,q}(B) = \mathbb{E}_n((B\Omega)^{\otimes q})$, and $\varphi_{0,q}(B) = \mathbb{E}((B\Omega)^{\otimes q})$. Let $\kappa = eM/\sqrt{\pi}$, and observe that Jensen's inequality implies

$$\|\varphi_{m,0}\|_\circ^6 \leq \mathbb{E}\big(\langle u_k, Z\rangle^{6m}\big) \leq eM \cdot (12\sigma^2/e)^{3m} m^{3m}/\sqrt{\pi} \leq \mathcal{C}_{m,0}[n]^6/\kappa. \tag{49}$$

A similar calculation shows that for some $T = M(t)$ with $t \in \mathbb{S}^{dp-1}$ we have

$$\|\varphi_{0,q}\|_\circ^6 \leq \mathbb{E}\big(\langle T^\mathsf{T} v_k, \Omega\rangle^{6m}\big) \leq eM \cdot (12\sigma^2/e)^{3q} q^{3q}/\sqrt{\pi} \leq \mathcal{C}_{0,q}[n]^6/\kappa. \tag{50}$$

Next note that two applications of the triangle inequality imply

$$\begin{aligned}\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ \leq \|\varphi_{m,0}\|_\circ \cdot \|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ + \\ \|\varphi_{0,q}\|_\circ \cdot \|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ + \|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ \cdot \|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ.\end{aligned} \tag{51}$$

Hence the union bound implies

$$\mathbb{P}\big(\|\widehat{\nu}_{m,q} - \nu_{m,q}\|_\circ > 2\mathcal{C}_{m,q}[n] \cdot \gamma + \mathcal{C}_{m,q}[n]^2 \cdot \gamma^2\big) \leq \mathsf{I} + \mathsf{II} + \mathsf{III} + \mathsf{IV} \tag{52}$$

for terms we define next. To bound these terms, we use (48). Observe that $\mathsf{I} = \mathbb{P}\big(\|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ > \mathcal{C}_{m,q}[n] \cdot \gamma\big) \leq \big(\gamma^6 \cdot n^2\big)^{-1}$, that $\mathsf{II} = \mathbb{P}\big(\|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ > \mathcal{C}_{m,q}[n] \cdot \gamma\big) \leq \big(\gamma^6 \cdot n^2\big)^{-1}$, that

$$\begin{aligned}\mathsf{III} &= \mathbb{P}\big(\mathcal{C}_{m,0}[n] \cdot \|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ > \kappa \cdot \mathcal{C}_{m,q}[n] \cdot \gamma\big) \\ &\leq \mathbb{P}\big(\|\widehat{\varphi}_{0,q} - \varphi_{0,q}\|_\circ > \mathcal{C}_{0,q}[n] \cdot \gamma/\kappa\big) \\ &\leq \big(\gamma^6 \cdot n^2\big)^{-1}\end{aligned} \tag{53}$$

and that

$$\begin{aligned}\mathsf{IV} &= \mathbb{P}\big(\mathcal{C}_{0,q}[n] \cdot \|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ > \kappa \cdot \mathcal{C}_{m,q}[n] \cdot \gamma\big) \\ &\leq \mathbb{P}\big(\|\widehat{\varphi}_{m,0} - \varphi_{m,0}\|_\circ > \mathcal{C}_{m,0}[n] \cdot \gamma/\kappa\big) \\ &\leq \big(\gamma^6 \cdot n^2\big)^{-1}\end{aligned} \tag{54}$$

Combining the above with (51) gives the second result. □

PROOF OF THEOREM 9. The proof is omitted because it is a straightforward modification of the proofs for Proposition 6 and Theorems 4 and 5. □

PROOF OF THEOREM 10. The proof is omitted because it is a straightforward modification of the proof for Theorem 6 after noting that Cauchy-Schwarz and Jensen's inequalities imply

$$\|\Xi_{m,q}(\widehat{B}_n)\| \leq 2eM \cdot (\sqrt{4\sigma^2/e})^{m+q} m^{m/2} q^{q/2}/\sqrt{\pi}. \tag{55}$$

□

PROOF OF PROPOSITION 9. The proof is omitted because it is a straightforward modification of the proof for Proposition 8. □

PROOF OF THEOREM 11. The proof is omitted because it is a straightforward modification of the proof for Theorems 2 and 6. □

REFERENCES

BANACH, S. (1938). Über homogene Polynome in (L^{2}). *Studia Mathematica* **7** 36–44.

BILLINGSLEY, P. (1995). *Probability and Measure*, 3 ed. Wiley.

BOCHNAK, J. and SICIAK, J. (1971). Polynomials and multilinear mappings in topological vector-spaces. *Studia Mathematica* **39** 59–76.

RIO, E. (2009). Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability* **22** 146–163.

ROCKAFELLAR, R. J.-B. R TYRRELL & WETS (2009). *Variational analysis* **317**. Springer Science & Business Media.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.