Calibrated Data-Dependent Constraints with Exact Satisfaction Guarantees

Songkai Xue
Department of Statistics
University of Michigan
sxue@umich.edu

Yuekai Sun
Department of Statistics
University of Michigan
yuekai@umich.edu

Mikhail Yurochkin
IBM Research
MIT-IBM Watson AI Lab
mikhail.yurochkin@ibm.com

Abstract

We consider the task of training machine learning models with data-dependent constraints. Such constraints often arise as empirical versions of expected value constraints that enforce fairness or stability goals. We reformulate data-dependent constraints so that they are *calibrated*: enforcing the reformulated constraints guarantees that their expected value counterparts are satisfied with a user-prescribed probability. The resulting optimization problem is amendable to standard stochastic optimization algorithms, and we demonstrate the efficacy of our method on a fairness-sensitive classification task where we wish to guarantee the classifier's fairness (at test time).

1 Motivation

In machine learning (ML) practice, accuracy is often only one of many training objectives. For example, algorithmic fairness considerations may require a credit scoring system to perform comparably on men and women. Here are a few other examples.

Churn rate and stability The churn rate of an ML model compared to another model is the fraction of samples on which the predictions of the two models differ [21, 29]. In ML practice, one may wish to control the churn rate between a new model and its predecessor because a high churn rate can disorient users and downstream system components. One way of training models with small churn is to enforce a churn rate constraint during training.

Precision, recall, *etc.* Classification and information retrieval models must often balance precision and recall. To train such models, practitioners carefully trade off one metric for the other by optimizing for one metric subject to constraints on the other.

Resource constraints Practitioners sometimes wish to control how often a classifier predicts a certain class due to budget or resource constraints. For example, a company that uses ML to select customers for a targeted offer may wish to constrain the fraction of customers selected for the offer. Another prominent example of a stochastic optimization problem with resource constraints is the newsvendor problem, which we come back to in section 4.

Unlike constraints on the structure of model parameters (*e.g.*, sparsity), the constraints encoding the preceding training objectives are *data-dependent*. This leads to the issue of *constraint generalization*: whether the constraints *generalize* out-of-sample. For example, if a classifier is trained to have comparable accuracy on two subpopulations in the training data, will it also have comparable accuracy on samples from the two subpopulations at test time?

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

In this paper, we consider the out-of-sample generalization of *expected-value* constraints. To keep things simple, consider a stochastic optimization problem with a single *expected-value* constraint:

$$\theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}_{P_0}\big[f(\theta;Z)\big] = \int_{\mathcal{Z}} f(\theta;z)dP_0(z) \\ \text{subject to} & \mathbb{E}_{P_0}\big[g(\theta;Z)\big] = \int_{\mathcal{Z}} g(\theta;z)dP_0(z) \leq 0 \end{array} \right\}, \tag{1.1}$$

where Θ is a (finite-dimensional) parameter space, $f, g : \Theta \times \mathcal{Z} \to \mathbb{R}$ are (known) cost and constraint functions, and $Z \in \mathcal{Z}$ is a random variable that represents a sample. The distribution of Z is unknown, so we cannot solve (1.1) directly. Instead, we obtain IID training samples $\{Z_i\}_{i=1}^n$ from the true underlying distribution P_0 and solve the empirical version of (1.1):

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n f(\theta;Z_i) \\ \text{subject to} & \frac{1}{n}\sum_{i=1}^n g(\theta;Z_i) \leq 0 \end{array} \right\}. \tag{1.2}$$

The estimator $\widehat{\theta}_n$ (of $\theta^\star$) is guaranteed to satisfy the empirical constraint (*i.e.*, $\frac{1}{n}\sum_{i=1}^n g(\widehat{\theta}_n;Z_i) \leq 0$), but it is unclear whether $\widehat{\theta}_n$ satisfies the actual (population) constraint $\mathbb{E}_{P_0}\big[g(\theta;Z)\big] \leq 0$. As we shall see, under standard assumptions on (1.1), $\widehat{\theta}_n$ only satisfies the actual constraint with probability approaching $\frac{1}{2}$ (see corollary 2.2). This is especially problematic for constraints that encode algorithmic fairness goals. For example, the 80% rule published by the US Equal Employment Opportunity Commission, interpreted in the machine learning context, requires the rate at which a classifier predicts the advantaged label in minority groups to be at least 80% of the rate at which the classifier predicts the advantaged label in the majority group [3].

In this paper, we propose a distributionally robust version of (1.2) that *guarantees* the actual constraint $\mathbb{E}_{P_0}\big[g(\theta;Z)\big] \leq 0$ will be satisfied with probability $1-\alpha$:

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n f(\theta;Z_i) \\ \text{subject to} & \sup_{P:D_\varphi(P\|\widehat{P}_n)\leq\frac{\rho_\alpha}{n}} \mathbb{E}_P\big[g(\theta;Z)\big] \leq 0 \end{array} \right\}, \tag{1.3}$$

where D_φ is a φ-divergence (see section 2 for details), $\widehat{P}_n$ is the empirical distribution of the training samples, and $\sqrt{\rho_\alpha}$ is the $1-\alpha$ quantile of a standard normal random variable. More concretely, we show that $\widehat{\theta}_n$ achieves *asymptotically exact constraint satisfaction*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\mathbb{E}_{P_0}\big[g(\widehat{\theta}_n;Z)\big] \leq 0\Big\} = 1-\alpha. \tag{1.4}$$

Here the inner expectation is with respect to Z; the outer probability is with respect to the training samples $\{Z_i\}_{i=1}^n$. Three desirable properties of (1.3) are

1. **exact constraint satisfaction:** If the actual probability of constraint satisfaction exceeds $1-\alpha$, then the method is too conservative. This may (unnecessarily) increase the cost of the model. By picking ρ_α in (1.3) carefully, constraints are satisfied with asymptotically exact probability $1-\alpha$.
2. **computationally efficiency:** As we shall see, the computational cost of solving (1.3) is comparable to the cost of solving distributionally robust sample average approximation (SAA) problems.
3. **pivotal:** There are no nuisance parameters to estimate (*e.g.*, asymptotic variances) in (1.3). The user merely needs to look up the correct quantile of the standard normal distribution for their desired level of constraint generalization.

The rest of this paper is organized as follows. In Section 2, we develop method, theory, and algorithm for stochastic optimization problems with single constraint. In Section 3, we extend our method, theory, and algorithm to stochastic optimization problems with multiple constraints. In Section 4, we validate our theory by simulating a resource-constrained newsvendor problem. In Section 5, we demonstrate the efficacy of our method by using it to train an algorithmically fair income classifier. In addition, we show how to apply our method to a fairness constrained learning problem and discuss two practical considerations for fair ML application scenarios. Finally, we summarize our work in Section 6 and point out an interesting avenue of future work.

1.1 Related work

The closest work to our work is [26]. They seek to pick a (data-dependent) *uncertainty set* $\mathcal{U}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{\sup_\theta \left\{\mathbb{E}_{P_0}\big[g(\theta;Z)\big] - \sup_{P\in\mathcal{U}} \mathbb{E}_P\big[g(\theta;Z)\big]\right\} \leq 0\right\} = 1-\alpha. \tag{1.5}$$

This condition is stronger than necessary: we only require

$$\lim_{n\to\infty} \mathbb{P}\left\{\mathbb{E}_{P_0}\big[g(\widehat{\theta}_n; Z)\big] - \sup_{P\in\mathcal{U}} \mathbb{E}_P\big[g(\widehat{\theta}_n; Z)\big] \leq 0\right\} = 1-\alpha \tag{1.6}$$

where $\widehat{\theta}_n$ is a (data-dependent) estimator (not necessarily (1.2) or (1.3)). [26] study (asymptotic) constraint satisfaction (1.4) for all deterministic objective functions (see [26], §1.1 for details). They advocate picking a KL divergence ball with radius that depends on the excursion probability of a certain χ^2 process.

Another closely related line of work is on data-splitting approaches for ensuring constraint generalization [36, 7]. At a high level, they split the training data into a training and validation subsets and use the validation subset to tune models trained on the training subset so that they satisfy the constraints. Although (computationally) simple and intuitive, their approach does not allow users to precisely control the constraint violation probability.

[26] is the latest in a line of work on distributionally robust optimization (DRO) that show the optimal value of DRO problems

$$\min_{\theta\in\Theta} \sup_{P\in\mathcal{U}} \mathbb{E}_P\big[g(\theta; Z)\big], \tag{1.7}$$

where $\mathcal{U}$ is a (data-dependent) uncertainty set of probability distributions, are upper confidence bounds for the optimal values of stochastic optimization problems. Common choices of uncertainty sets in DRO include uncertainty sets defined by moment or support constraints [6, 12, 22], φ-divergences [4, 25, 30], and Wasserstein distances [33, 5, 18, 27, 34]. This line of work is motivated by Owen's seminal work on empirical likelihood [31]. In recent work, [25, 15] show that the optimal value of DRO problems with empirical likelihood uncertainty sets leads to asymptotically exact upper confidence bounds for the optimal value of stochastic optimization problems ([15] consider more general φ-divergence uncertainty sets). [5] establish similar coverage results for Wasserstein uncertainty sets.

Our work is also closely related to the work on the variance regularization properties of DRO [30], which uses DRO to approximate the variance regularization cost function (see (2.4)). [20] establish similar results for Wasserstein DRO.

2 Single expected value constraint

We motivate (1.3) by considering a few alternatives. First, we note that the results later in this section show that (1.2) violates the actual constraint in (1.1) approximately half the time (see corollary 2.2). The most straightforward modification of (1.2) to ensure $\widehat{\theta}_n$ satisfies the (actual) constraint $\mathbb{E}_{P_0}\big[g(\theta; Z)\big] \leq 0$ is to add a "margin" in (1.3); *i.e.* enforce the constraint

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n g(\theta; Z_i) + \epsilon_n \leq 0 \tag{2.1}$$

in (1.2). If we pick the slack term ϵ_n such that

$$\mathbb{P}\left\{\sup_{\theta\in\Theta}\left\{\mathbb{E}_{P_0}\big[g(\theta; Z)\big] - \tfrac{1}{n}\textstyle\sum_{i=1}^n g(\theta; Z_i)\right\} > \epsilon_n\right\} \leq \alpha,$$

then it is not hard to check that the resulting $\widehat{\theta}_n$ satisfies the (actual) constraint with probability greater than $1-\alpha$ [35, 28]. However, this approach is most likely conservative because the constraint is unnecessarily stringent for θ's such that $\frac{1}{n}\sum_{i=1}^n g(\theta; Z_i)$ is less variable. It is also not pivotal: ϵ_n is often set using bounds from (uniform) concentration inequalities, which typically depend on unknown problem parameters.

To relax the empirical constraint in a way that adapts to the variability of the empirical constraints, we replace the uniform margin in (2.1) with a parameter-dependent margin:

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n g(\theta; Z_i) + z_\alpha \frac{\widehat{\boldsymbol{\sigma}}(\theta)}{\sqrt{n}} \leq 0, \tag{2.2}$$

where z_α is the $1-\alpha$ quantile of a standard normal random variable and $\widehat{\boldsymbol{\sigma}}^2(\theta)$ is an estimate of the asymptotic variance of $g(\theta; Z)$. We recognize the (parameter-dependent) margin as (a multiple of) the standard error of the empirical constraint. It is possible to show that enforcing (2.2) achieves asymptotically exact constraint generalization (1.4) [26].

The main issue with this method is it is not amenable to standard stochastic optimization algorithms. In particular, even if the original constraint in (1.2) is convex, (2.2) is generally non-convex. Another issue is that it is not pivotal: the user must estimate the asymptotic variance of $g(\theta; Z)$.

To overcome these two issues, we consider a distributionally robust version of (1.2); *i.e.* enforcing

$$\sup_{P: D_\varphi(P\|\widehat{P}_n) \leq \frac{\rho_\alpha}{n}} \mathbb{E}_P\big[g(\theta; Z)\big] \leq 0, \tag{2.3}$$

where $D_\varphi(P\|Q) \triangleq \int \varphi(\frac{dP}{dQ})dQ$ is a φ-divergence. Common choices of φ include $\varphi(t) = (t-1)^2$ (which leads to the χ^2-divergence) and $\varphi(t) = -\log t + t - 1$ (which leads to the Kullback-Leibler divergence). Although there are many other choices for the uncertainty set in (2.3), we pick an φ-divergence ball because (i) (2.3) with an φ-divergence ball is asymptotically equivalent to (2.2):

$$\sup_{P: D_\varphi(P\|\widehat{P}_n) \leq \frac{\rho_\alpha}{n}} \mathbb{E}_P\big[g(\theta; Z)\big] \approx \tfrac{1}{n}\textstyle\sum_{i=1}^n g(\theta; Z_i) + z_\alpha \frac{\widehat{\boldsymbol{\sigma}}(\theta)}{\sqrt{n}}, \tag{2.4}$$

and (ii) it leads to pivotal uncertainty sets. For theoretical analysis, we always use $\varphi(t) = (t-1)^2$ and χ^2-divergence in the remainder of this paper.

Before we state the asymptotically exact constraint satisfaction property of (1.3) rigorously, we describe our assumptions on the problem.

1. **smoothness and concentration:** f and g are twice continuously differentiable with respect to θ, and $f(\theta^\star; Z)$, $\nabla f(\theta^\star; Z)$, $g(\theta^\star; Z)$, $\nabla g(\theta^\star; Z)$ are sub-Gaussian random variables.
2. **uniqueness:** the stochastic optimization problem with a single expected value constraint (1.1) has a unique optimal primal-dual pair $(\theta^\star, \lambda^\star)$, and $\theta^\star$ belongs to the interior of the compact set Θ.
3. **strict complementarity:** $\lambda^\star > 0$.
4. **positive definiteness:** The Hessian of the Lagrangian evaluated at $(\theta^\star, \lambda^\star)$ is positive definite.

The preceding assumptions are not the most general, but they are easy to interpret. The smoothness conditions on f and g with respect to θ, the concentration conditions of $f(\theta^\star; Z)$ and $g(\theta^\star; Z)$, and the uniqueness condition facilitate the use of standard tools from asymptotic statistics to study the large sample properties of the constraint value. The strict complementarity condition rules out problems in which the constraint is extraneous; *i.e.* problems in which the unconstrained minimum coincides with the constrained minimum.

We are ready to state the asymptotically exact constraint satisfaction property of (1.3) rigorously. The main technical result characterizes the limiting distribution of the constraint value.

Theorem 2.1. *Let $\widehat{\theta}_n$ be an optimal solution of* (1.3) *converging in probability as $n \to \infty$ to $\theta^\star$. Under the standing assumptions, we have*

$$\sqrt{n}\left(\mathbb{E}_{P_0}\big[g(\widehat{\theta}_n; Z)\big] - \cancel{\mathbb{E}_{P_0}\big[g(\theta^\star; Z)\big]}\right) \xrightarrow{d} \mathcal{N}\left(-\sqrt{\rho_\alpha \operatorname{Var}_{P_0}[g(\theta^\star; Z)]}, \operatorname{Var}_{P_0}[g(\theta^\star; Z)]\right).$$

We translate this result on the constraint value to a result on constraint generalization.

Corollary 2.2. *Let $\sqrt{\rho_\alpha}$ be the $1-\alpha$ quantile of a standard normal random variable. Under the conditions of theorem 2.1, we have*

$$\lim_{n\to\infty} \mathbb{P}\left\{\mathbb{E}_{P_0}\big[g(\widehat{\theta}_n; Z)\big] \leq 0\right\} = \mathbb{P}\left\{U \leq \sqrt{\rho_\alpha}\right\} = 1 - \alpha,$$

where $U \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable.

From theorem 2.1 and corollary 2.2 (see proofs in Appendix A), we find that

1. picking $\rho_\alpha = 0$ (*i.e.*, equivalently solving (1.2)) leads to a constraint violation probability that approaches $\frac{1}{2}$ in the large sample limit.
2. the relation between the mean and variance of the limiting distribution of the constraint value in Theorem 2.1 allows us to pick ρ_α in a pivotal way (*i.e.* does not depend on nuisance parameters).

2.1 Stochastic approximation for (1.3)

In the rest of this section, we derive a stochastic optimization algorithm to solve (1.3) efficiently. As we shall see, the computational cost of this algorithm is comparable to the cost of solving a DRO problem. The key insight is that the robust constraint function has a dual form (see Appendix J):

$$\sup_{P: D_\varphi(P\|\widehat{P}_n) \leq \rho} \mathbb{E}_P\big[g(\theta; Z)\big] = \inf_{\mu \geq 0, \nu \in \mathbb{R}} \left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n \mu\varphi^*\big(\frac{g(\theta; Z_i) - \nu}{\mu}\big) + \mu\rho + \nu\right\}, \tag{2.5}$$

where $\varphi^*(s) \triangleq \sup_t\{st - \varphi(t)\}$ is the convex conjugate of φ. As we use χ^2-squared divergence and $\varphi(t) = (t-1)^2$, the corresponding $\varphi^*(s) = \frac{s^2}{4} + s$. The Lagrangian of (1.3) is

$$\begin{aligned} L(\theta, \lambda) &\triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n f(\theta; Z_i) + \lambda \sup_{P: D_\varphi(P\|\widehat{P}_n) \le \frac{\rho_\alpha}{n}} \mathbb{E}_P\big[g(\theta; Z)\big] \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n f(\theta; Z_i) + \lambda \inf_{\mu \ge 0, \nu \in \mathbb{R}} \left\{ \tfrac{1}{n}\sum_{i=1}^n \mu\varphi^*\big(\tfrac{g(\theta;Z_i)-\nu}{\mu}\big) + \mu\tfrac{\rho_\alpha}{n} + \nu \right\}. \end{aligned}$$

We see that evaluating the dual function $\inf_\theta L(\theta, \lambda)$ (at a fixed λ) entails solving a stochastic optimization problem that is suitable for stochastic approximation. This suggests a dual ascent algorithm for solving (1.3):

1. evaluate the dual function at λ_t by solving a stochastic optimization problem
2. update λ_t with a dual ascent step.

We summarize this algorithm in Algorithm 1. The main cost of Algorithm 1 is incurred in the third line: evaluating the dual function. Fortunately, this step is suitable for stochastic approximation, so we can leverage recent advances in the literature to reduce the (computational) cost of this step. The total cost of this algorithm is comparable to that of distributionally robust optimization.

Algorithm 1 Dual ascent algorithm for (1.3)

1: **Input:** starting dual iterate $\lambda_0 \ge 0$
2: **repeat**
3: Evaluate dual function:

$$(\theta_t, \mu_t, \nu_t) \leftarrow \arg\min_{\theta, \mu \ge 0, \nu} \tfrac{1}{n}\textstyle\sum_{i=1}^n f(\theta; Z_i) + \lambda_t \left\{ \tfrac{1}{n}\sum_{i=1}^n \mu\varphi^*\big(\tfrac{g(\theta;Z_i)-\nu}{\mu}\big) + \mu\tfrac{\rho_\alpha}{n} + \nu \right\}$$

4: Dual ascent update: $\lambda_{t+1} \leftarrow \left[\lambda_t + \eta_t \left\{ \tfrac{1}{n}\sum_{i=1}^n \mu_t\varphi^*\big(\tfrac{g(\theta_t;Z_i)-\nu_t}{\mu_t}\big) + \mu_t\tfrac{\rho_\alpha}{n} + \nu_t \right\}\right]_+$
5: **until** converged

3 Multiple expected value constraints

In this section, we extend the results from the preceding section to stochastic optimization problems with multiple data-dependent constraints. Consider a stochastic optimization problem with K expected value constraints

$$\theta^\star \in \begin{Bmatrix} \arg\min_{\theta\in\Theta} & \mathbb{E}_{P_0}\big[f(\theta; Z)\big] \\ \text{subject to} & \big\{\mathbb{E}_{P_0}\big[g_k(\theta; Z)\big] \le 0\big\}_{k=1}^K \end{Bmatrix}, \tag{3.1}$$

Following the development in Section 2, we enforce the expected value constraints with robust versions of the sample average constraints:

$$\widehat{\theta}_n \in \begin{Bmatrix} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n f(\theta; Z_i) \\ \text{subject to} & \left\{\sup_{P: D_\varphi(P\|\widehat{P}_n) \le \frac{\rho_k}{n}} \mathbb{E}_P\big[g_k(\theta; Z)\big] \le 0\right\}_{k=1}^K \end{Bmatrix}, \tag{3.2}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)^\top$ are uncertainty set radii for the constraints. There are other approaches to enforcing multiple constraints that result in constraint generalization; we focus on (3.2) here because it allows the user to adjust the constraint generalization probability for different constraints.

First, we extend theorem 2.1 and corollary 2.2 to problems with multiple (expected value) constraints. We assume

1. **smoothness and concentration:** for $k \in [K]$, f, g_k are twice continuously differentiable with respect to θ, and $f(\theta^\star; Z), \nabla f(\theta^\star; Z), g_k(\theta^\star; Z), \nabla g_k(\theta^\star; Z)$ are sub-Gaussian random variables.
2. **uniqueness:** the stochastic optimization problem with K expected value constraints (3.1) has a unique optimal primal-dual pair $(\theta^\star, \boldsymbol{\lambda}^\star)$, and $\theta^\star$ belongs to the interior of the compact set Θ.
3. **strict complementarity:** $\boldsymbol{\lambda}^\star \in \operatorname{int}(\mathbb{R}_+^K)$, *i.e.*, each component of $\boldsymbol{\lambda}^\star$ is strictly positive.
4. **positive definiteness:** The Hessian of the Lagrangian evaluated at $(\theta^\star, \boldsymbol{\lambda}^\star)$ is positive definite.

The strict complementarity constraint seems especially strong here because it requires all the constraints to be active. It is possible (with extra notational overhead) to state the result in terms of just the active constraints. We refer to Section 5.1 for more information about the unknown active set. Further, as long as the sample size is large enough, the active constraints in (3.2) coincide with the active constraints in (3.1). To keep things simple, we assume all the constraints are active.

Theorem 3.1. *Let $\widehat{\theta}_n$ be an optimal solution of* (3.2) *converging in probability as $n \to \infty$ to $\theta^\star$. Under the standing assumptions, we have*

$$\sqrt{n}\begin{bmatrix}\mathbb{E}_{P_0}\big[g_1(\widehat{\theta}_n;Z)\big]\\ \vdots \\ \mathbb{E}_{P_0}\big[g_K(\widehat{\theta}_n;Z)\big]\end{bmatrix} \xrightarrow{d} \mathcal{N}\left(-\begin{bmatrix}\sqrt{\rho_1 \operatorname{Var}_{P_0}[g_1(\theta^\star;Z)]}\\ \vdots \\ \sqrt{\rho_K \operatorname{Var}_{P_0}[g_K(\theta^\star;Z)]}\end{bmatrix}, \operatorname{Var}_{P_0}\begin{bmatrix}g_1(\theta^\star;Z)\\ \vdots \\ g_K(\theta^\star;Z)\end{bmatrix}\right).$$

Corollary 3.2. *Under the conditions of theorem 3.1, we have*

$$\lim_{n\to\infty} \mathbb{P}\left\{\begin{bmatrix}\mathbb{E}_{P_0}\big[g_1(\widehat{\theta}_n;Z)\big]\\ \vdots \\ \mathbb{E}_{P_0}\big[g_K(\widehat{\theta}_n;Z)\big]\end{bmatrix} \in -\mathbb{R}^K_+\right\} = \mathbb{P}\{\boldsymbol{U} \le \sqrt{\boldsymbol{\rho}}\},$$

where $\sqrt{\boldsymbol{\rho}} = (\sqrt{\rho_1},\dots,\sqrt{\rho_K})^\top$, and $\boldsymbol{U}$ is a Gaussian random vector with mean zero and covariance

$$\operatorname{Corr}_{P_0}\begin{bmatrix}g_1(\theta^\star;Z)\\ \vdots \\ g_K(\theta^\star;Z)\end{bmatrix} \triangleq D^{-\frac{1}{2}} \operatorname{Cov}_{P_0}\begin{bmatrix}g_1(\theta^\star;Z)\\ \vdots \\ g_K(\theta^\star;Z)\end{bmatrix} D^{-\frac{1}{2}}, \qquad D \triangleq \operatorname{diag}\left(\{\operatorname{Var}_{P_0}[g_k(\theta^\star,Z)]\}_{k=1}^K\right). \tag{3.3}$$

From theorem 3.1 and corollary 3.2 (see proofs in Appendix B and C), we find that the probability of constraint satisfaction decreases *exponentially* as the number of constraints increases. We also see that our method is no longer pivotal for multiple expected value constraints: the uncertainty set radii depends on the (unknown) correlation structure among the constraint values. Fortunately, it is not hard to estimate this correlation structure. The most straightforward way is with the empirical correlation matrix. Let $\widehat{\Sigma}_n$ be the empirical covariance matrix of the constraint values. The empirical correlation matrix is then given by $\widehat{R}_n \triangleq \operatorname{diag}(\widehat{\Sigma}_n)^{-\frac{1}{2}}\widehat{\Sigma}_n \operatorname{diag}(\widehat{\Sigma}_n)^{-\frac{1}{2}}$.

Finally, it is straightforward to extend the algorithm for solving (1.3) to (3.2). The Lagrangian of (3.2) is

$$\begin{aligned} L(\theta,\boldsymbol{\lambda}) &\triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n f(\theta;Z_i) + \sum_{k=1}^K \lambda_k \sup_{P:D_\varphi(P\|P_n)\le\frac{\rho_k}{n}} \mathbb{E}_P\big[g_k(\theta;Z)\big] \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n f(\theta;Z_i) + \sum_{k=1}^K \lambda_k \inf_{\mu_k\ge 0,\nu_k\in\mathbb{R}}\left\{\tfrac{1}{n}\sum_{i=1}^n \mu_k\varphi^*\big(\tfrac{g_k(\theta;Z_i)-\nu_k}{\mu_k}\big) + \mu_k\tfrac{\rho_k}{n} + \nu_k\right\}, \end{aligned}$$

where we recalled the dual form of the robust constraint function (2.5) in the second step. We see that evaluating the dual function $\inf_\theta L(\theta,\boldsymbol{\lambda})$ (at a fixed $\boldsymbol{\lambda}$) entails solving a stochastic optimization problem that is suitable for stochastic approximation. This suggests a similar dual ascent algorithm for solving (1.3); we skip the details here (see Algorithm 2 in Appendix D).

4 Simulations

We simulate the frequency of constraint satisfaction for the following multi-item newsvendor problem:

$$\begin{array}{ll} \max_{\theta\in\Theta} & \mathbb{E}_{P_0}\big[p^\top \min\{Z,\theta\} - c^\top\theta\big] \\ \text{subject to} & \mathbb{E}_{P_0}[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+] \le \varepsilon_1 \\ & \mathbb{E}_{P_0}[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+] \le \varepsilon_2 \end{array} \tag{4.1}$$

where $c \in \mathbb{R}^d_+$ is the manufacturing cost, $p \in \mathbb{R}^d_+$ is the sell price, $\theta \in \Theta = [0,100]^d$ is the number of items in stock, $Z \in \mathbb{R}^d$ is a random variable with probability distribution P_0 representing the demand, and there are d items in total. The distribution P_0 is unknown but we observe IID samples $Z_1,\dots,Z_n$ from P_0. All of the items have been partitioned into two groups so that the corresponding

demand and stock can be written as $Z = (Z^{(1)}, Z^{(2)})$ and $\theta = (\theta^{(1)}, \theta^{(2)})$. The constraints in the problem exclude stock levels that underestimate the demand too much for each group of items, where $\varepsilon_1, \varepsilon_2 > 0$ indicate tolerance level of such underestimation. The target of the problem is to maximize the profit while satisfying the constraints. It is easy to rewrite the maximization problem (4.1) as a minimization problem with expected value constraints in the form of (3.1) so that we can apply our method (3.2). We pick P_0 as multivariate Gaussian with independent components so that the two constraints are generally uncorrelated with each other (see Appendix E for details).

Throughout the simulations, we solve (3.2) with $\boldsymbol{\rho} = (z_\alpha, z_\alpha)^\top$ for $\alpha \in \{0.4, 0.25, 0.1, 0.05, 0.005\}$. As suggested by our asymptotic theory in Section 3, the nominal probability of constraint satisfaction is $1 - \alpha$ for each constraint and $(1 - \alpha)^2$ for both constraints due to the independence setup.

In Figure 1, we plot frequencies of constraint satisfaction for each constraint and both constraints, all of which are averaged over 1000 replicates. As the sample size n grows, the frequency versus probability curve converges to the theoretical dashed line of limiting probability of constraint satisfaction, validating our theory in the large sample regime. For more simulations (*e.g.*, single constraint, two dependent constraints) we refer to Appendix E.

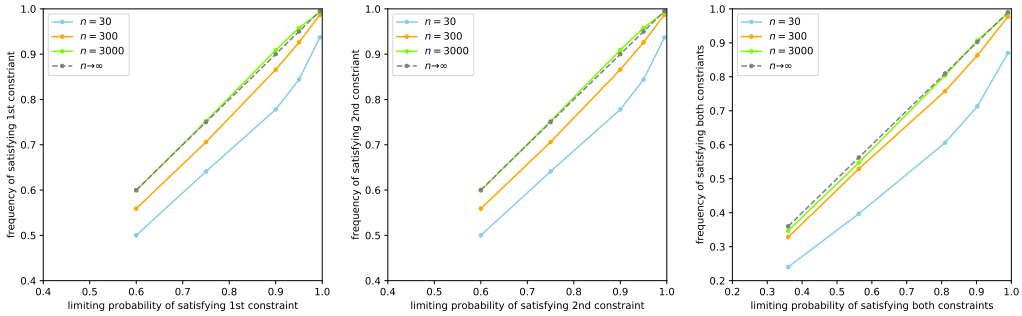

Figure 1: Frequency versus limiting probability of constraint satisfaction of the first constraint (left), the second constraint (middle), and both of the constraints (right).

5 Application to fair machine learning

As ML models are deployed in high-stakes decision making and decision support roles, the fairness of the models has come under increased scrutiny. In response, there is a flurry of recent work on mathematical definitions of algorithmic fairness [16, 23, 24] and algorithms to enforce the definitions [1, 10, 37].

A prominent class of fairness definitions is *group fairness*; such definitions require equality of certain metrics (*e.g.* false/true positive rates) among demographic groups. For example, consider a fair binary classification problem. Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space, $\{0, 1\}$ be the set of possible labels, and $\mathcal{A}$ be the set of possible values of the protected/sensitive attribute. In this setup, training and test examples are tuples of the form $(X, A, Y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, and a classifier is a map $f : \mathcal{X} \to \{0, 1\}$. A popular definition of algorithmic fairness for binary classification is *equality of opportunity* [23].

Definition 5.1 (equality of opportunity)**.** *Let $Y = 1$ be the advantaged label that is associated with a positive outcome and $\widehat{Y} \triangleq f(X)$ be the output of the classifier. Equality of opportunity entails $\mathbb{P}\{\widehat{Y} = 1 \mid A = a, Y = 1\} = \mathbb{P}\{\widehat{Y} = 1 \mid A = a', Y = 1\}$ for all $a, a' \in \mathcal{A}$.*

Equality of opportunity, or true positive rate parity, means that the prediction $\widehat{Y} = h(X)$ conditioned on the advantaged label $Y = 1$ is statistically independent of the protected attribute A. Furthermore, an approximate version of equality of opportunity can be readily defined. We say that $\widehat{Y} = h(X)$ satisfies ε-*equality of opportunity* if $\mathbb{P}\{\widehat{Y} = 1 \mid A = a, Y = 1\} - \mathbb{P}\{\widehat{Y} = 1 \mid A = a', Y = 1\} \leq \varepsilon$ for for all $a, a' \in \mathcal{A}$. In this case, $\varepsilon > 0$ represents a practitioner's *tolerance* for fairness violations.

Given a parametric model space $\mathcal{H} = \{f_\theta(\cdot) : \theta \in \Theta\}$ and loss function $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, an in-processing fair ML routine is to minimize the (empirical) risk $\mathbb{E}\left[\ell(\theta; X, Y)\right]$ while satisfying some fairness constraints. Most commonly, definitions of group fairness (including equality of opportunity,

demographic parity, and more) can be written as a special example of a general set of linear constraints [1, 2] of the form $\mathbf{M}\boldsymbol{\mu}(\theta) \leq \mathbf{c}$, where matrix $\mathbf{M} \in \mathbb{R}^{K\times T}$ and vector $\mathbf{c} \in \mathbb{R}^K$ encode the constraints; $\boldsymbol{\mu}(\theta) : \Theta \to \mathbb{R}^T$ is a vector of (conditional) moments $\mu_t(\theta) = \mathbb{E}\left[h_t(X, A, Y, \theta) \mid \mathcal{E}_t\right]$ for $t \in [T]$; $g_t : \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \times \Theta \to \mathbb{R}$; event $\mathcal{E}_t$ is defined with respect to (X, A, Y).

This framework fits to our methodology if we note that each (conditional) moment can be written as

$$\mu_t(\theta) = \frac{\mathbb{E}_{(X,A,Y)\sim P_0}\left[h_t(X, A, Y, \theta) \times \mathbf{1}\left\{\mathcal{E}_t(X, Y, A)\right\}\right]}{\mathbb{E}_{(X,A,Y)\sim P_0}\left[\mathbf{1}\left\{\mathcal{E}_t(X, Y, A)\right\}\right]}. \tag{5.1}$$

Here the indicator $\mathbf{1}\{\mathcal{E}_t\}$ takes value 1 if the event $\mathcal{E}_t$ happens, and 0 otherwise. Moreover, we use $\mathcal{E}_t(X, A, Y)$ to emphasize that $\mathcal{E}_t$ only depends on (X, Y, A) but not on θ in any way.

Note that (5.1) is a ratio of expected values, which is a non-linear statistical functional of P_0. To use our method, we first replace the denominator of $\mu_t(\theta)$ with an estimator, such as the unbiased estimator $\widehat{\mathbb{P}}(\mathcal{E}_t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\mathcal{E}_t(X_i, A_i, Y_i)\}$. The resulting plug-in estimation of $\mu_t(\theta)$ then becomes linear in P_0, allowing us to apply our method (see similar tricks in [8]). We describe the application of our method to ε-equality of opportunity in Appendix F.

5.1 A two-stage method for unknown active set

In practice, it is probable that only a subset of the constraints are active. Furthermore, we do not know beforehand whether or not a constraint is active in the true population problem. To handle this scenario, we propose a two-stage method:

1. At the first stage, we solve the sample average approximation (SAA) problem (3.2) with $\boldsymbol{\rho} = \mathbf{0}_K$. By doing so, we identify the active set of the SAA problem.
2. At the second stage, we solve (3.2) with $\boldsymbol{\rho}$ such that ρ_k is a positive number only if the k-th constraint, $k \in [K]$, was identified as active at the first stage.

In Appendix G, we show that the two-stage method also enjoys the calibration property (similar to Theorem 3.1 and Corollary 3.2) under standard assumptions (*i.e.*, strict complementarity). At a high level, the limiting probability of satisfying the true constraints depends solely on the correlation structure between active constraints and the uncertainty set radii for active constraints, as long as the SAA problem identifies active constraints with probability tending to 1.

5.2 Proxy dual function for non-differentiable constraints

Constraint functions in fair ML are often non-differentiable. For instance, fairness metrics are typically linear combinations of indicators that result in non-differentiable rate constraints [8–10]. This prevents the use of any gradient-based optimization algorithms. Fortunately, only the dual function evaluation step in Algorithm 1 requires access to gradients. Therefore, we can modify the algorithm by: (1) introducing proxy dual function, which uses a differentiable surrogate $\tilde{g}$ instead of the non-differentiable g in the dual function evaluation step; (2) keeping g in the dual ascent step. For an indicator function $h(t) = \mathbf{1}\{t > 0\}$, one can replace it by sigmoidal function $h_1(t) = (1+e^{-at})^{-1}$ or hinge upper bound $h_2(t) = \max\{0, t + 1\}$ to produce smooth surrogates for non-differentiable rate constraints [11, 17, 9]. We summarize the proxy dual ascent algorithm in Appendix H.

5.3 Adult experiments

We compare the frequency of constraint satisfaction (at test time) of the sample average approximation and our methods with nominal probability $0.60, 0.75, 0.90, 0.95$ using the Adult dataset from UCI [13]. The classification task is to predict whether an individual's income per year is higher than \$50K. The fairness goal is ε-demographic parity (ε-DP): $|\mathbb{P}(\widehat{Y} = 1 \mid A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid A = 0)| \leq \varepsilon$, where $A = 1$ for male is the advantaged group and $A = 0$ for female is the disadvantaged group. We use a logistic regression model for classification and techniques in this section for implementation.

In Figure 2, we have line plots for frequency of constraint satisfaction and box plots for classification error rate, all of which are summarized over 100 replicates. The left panel shows that solving (3.1) directly leads to one half chance of constraint violation, while our method's constraint satisfaction frequency matches its nominal value. The price of a higher chance of test-time fairness satisfaction is

an increase in classification error rate as shown in the right panel. From the baseline to 95% chance of fairness satisfaction, we basically trade off 2% increase in error rate. We refer to Appendix I and K for details and more experiments.

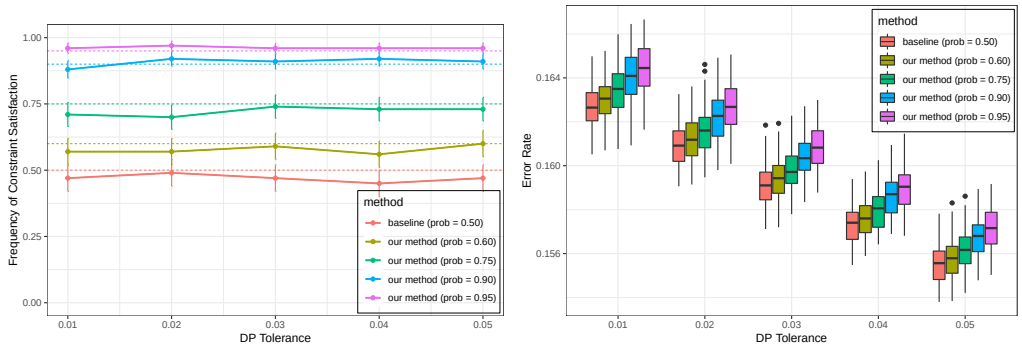

Figure 2: Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA) and our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$) are compared.

6 Summary and discussion

We explore the problem of exact constraint satisfaction probability in stochastic optimization with expected-value constraints. We propose a distributionally robust reformulation of data-dependent constraints and provide a theoretical guarantee of constraint satisfaction with an asymptotically exact probability specified by the user. For solving the reformulated problem, a scalable dual ascent algorithm and its variants are proposed. The computational cost of our algorithm is comparable to that of a standard distributionally robust optimization problem. Our theory on exact constraint satisfaction probability is validated via simulations on the resource-constrained newsvendor problem. The efficacy of our methods is empirically demonstrated on fair machine learning applications.

Some data-dependent constraints are by nature *non-linear* in the underlying probability measure. For example, (5.1) is a ratio of expected values. An intriguing direction for future research is to generalize the methods and theory developed in this work to constraints on non-linear functions of expected values. Such forms of constraints are known as *statistical functionals* in statistics literature [19]. The non-linear dependence of the constraint function on the probability measure precludes the stochastic approximation as a general way of evaluating the dual function, as the constraint function no longer admits a dual form (2.5), calling for the development of a new algorithm.

Acknowledgments and Disclosure of Funding

This paper is based upon work supported by the National Science Foundation (NSF) under grants no. 1916271, 2027737, and 2113373.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A Reductions Approach to Fair Classification. *arXiv:1803.02453 [cs]*, July 2018.

[2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair Regression: Quantitative Definitions and Reduction-based Algorithms. *arXiv:1905.12843 [cs, stat]*, May 2019.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

[4] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, November 2012. ISSN 0025-1909. doi: 10.1287/mnsc.1120.1641.

[5] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein Profile Inference and Applications to Machine Learning. *arXiv:1610.05627 [math, stat]*, October 2016.

[6] Xin Chen, Melvyn Sim, and Peng Sun. A Robust Optimization Perspective on Stochastic Programming. *Operations Research*, 55:1058–1071, 2007. doi: 10.1287/opre.1070.0441.

[7] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. *arXiv:1807.00028 [cs, stat]*, September 2018.

[8] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR, 2019.

[9] Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172): 1–59, 2019. ISSN 1533-7928.

[10] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-Player Games for Efficient Non-Convex Constrained Optimization. In *Algorithmic Learning Theory*, pages 300–332. PMLR, March 2019.

[11] Mark A Davenport, Richard G Baraniuk, and Clayton D Scott. Tuning support vector machines for minimax and neyman-pearson classification. *IEEE transactions on pattern analysis and machine intelligence*, 32(10):1888–1898, 2010.

[12] Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58:595–612, 2010. doi: 10.1287/opre.1090.0741.

[13] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[14] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. October 2016.

[15] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *arXiv:1610.03425 [stat]*, October 2016.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness Through Awareness. *arXiv:1104.3913 [cs]*, April 2011.

[17] Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Ryan Rifkin, and Gal Elidan. Scalable learning of non-decomposable objectives. In *Artificial intelligence and statistics*, pages 832–840. PMLR, 2017.

[18] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. May 2015.

[19] Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

[20] Chao Gao. Robust Regression via Mutivariate Regression Depth. *arXiv:1702.04656 [math, stat]*, February 2017.

[21] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P. Friedlander. Satisfying Real-world Goals with Dataset Constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

[22] Joel Goh and Melvyn Sim. Distributionally Robust Optimization and Its Tractable Approximations. *Operations Research*, 58(4-part-1):902–917, August 2010. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.1090.0795.

[23] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]*, October 2016.

[24] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. *arXiv:1703.06856 [cs, stat]*, March 2018.

[25] H. Lam and Enlu Zhou. Quantifying uncertainty in sample average approximation. In *2015 Winter Simulation Conference (WSC)*, pages 3846–3857, December 2015. doi: 10.1109/WSC.2015.7408541.

[26] Henry Lam. Recovering Best Statistical Guarantees via the Empirical Divergence-Based Distributionally Robust Optimization. *Operations Research*, 67(4):1090–1105, July 2019. ISSN 0030-364X. doi: 10.1287/opre.2018.1786.

[27] Jaeho Lee and Maxim Raginsky. Minimax Statistical Learning with Wasserstein Distances. *arXiv:1705.07815 [cs]*, May 2017.

[28] James Luedtke and Shabbir Ahmed. A Sample Approximation Approach for Optimization with Probabilistic Constraints. *SIAM Journal on Optimization*, 19(2):674–699, January 2008. ISSN 1052-6234. doi: 10.1137/070702928.

[29] Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and Iterate: Reducing Prediction Churn. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[30] Hongseok Namkoong and John C. Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with F-divergences. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2216–2224, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

[31] Art Owen. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18(1): 90–120, March 1990. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176347494.

[32] Reuven Y Rubinstein and Alexander Shapiro. *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*, volume 13. Wiley, 1993.

[33] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. September 2015.

[34] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. *arXiv:1710.10571 [cs, stat]*, October 2017.

[35] Wei Wang and Shabbir Ahmed. Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36(5):515–519, September 2008. ISSN 0167-6377. doi: 10.1016/j.orl.2008.05.003.

[36] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, June 2017.

[37] Mikhail Yurochkin and Yuekai Sun. SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness. In *International Conference on Learning Representations*, September 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Materials for Calibrated Data-Dependent Constraints with Exact Satisfaction Guarantees

Songkai Xue
Department of Statistics
University of Michigan
sxue@umich.edu

Yuekai Sun
Department of Statistics
University of Michigan
yuekai@umich.edu

Mikhail Yurochkin
IBM Research
MIT-IBM Watson AI Lab
mikhail.yurochkin@ibm.com

Abstract

In Section A, B, C, we present proofs of theoretical results. In Section D, we summarize the dual ascent algorithm for solving (3.2). In Section E, we provide details for simulations on the multi-item newsvendor problem with independent constraints and run additional simulations. In Section F, we demonstrate how our method can be applied to ε-equality of opportunity. In Section G, we show the theoretical properties of the two-stage method for unknown active set. In Section H, we summarize the proxy dual ascent algorithm for handling non-differentiable constraints. In Section I, we provide details for Adult experiments. In Section J, we provide a standard derivation for the dual form of the robust constraint function (2.5). In Section K, experiments on additional baseline and dataset are conducted.

A Proofs of Theorem 2.1 and Corollary 2.2

Note that Theorem 3.1 implies Theorem 2.1 and Corollary 3.2 implies Corollary 2.2 by letting $K = 1$. Therefore, it is sufficient to prove Theorem 3.1 and Corollary 3.2, whose proofs can be found in Appendix B and C respectively. □

B Proof of Theorem 3.1

Consider a stochastic optimization problem with K expected value constraints

$$(\mathcal{P}_0): \quad \theta^\star \in \arg\min_{\theta\in\Theta} \left\{\mathbb{E}_{Z\sim P_0}\left[f(\theta;Z)\right] : \mathbb{E}_{Z\sim P_0}\left[g_k(\theta;Z)\right] \le 0, k\in[K]\right\}.$$

Our proposed robust constraint method solves

$$(\mathcal{P}_n): \quad \widehat{\theta}_{n,\boldsymbol{\rho}} \in \arg\min_{\theta\in\Theta} \left\{\mathbb{E}_{Z\sim P_n}\left[f(\theta;Z)\right] : \sup_{D_\varphi(Q\|P_n)\le\rho_k/n} \mathbb{E}_{Z\sim Q}\left[g_k(\theta;Z)\right] \le 0, k\in[K]\right\},$$

where $\boldsymbol{\rho} = (\rho_1,\ldots,\rho_K)^\top$ is the collection of critical radii of uncertainty sets. Here we denote the empirical distribution $\widehat{P}_n$ by P_n for notation simplicity.

As a special case of our robust method, the sample average approximation (SAA) or empirical risk minimization (ERM) solves

$$\widehat{\theta}_{n,\mathbf{0}_K} \in \arg\min_{\theta\in\Theta} \left\{\mathbb{E}_{Z\sim P_n}\left[f(\theta;Z)\right] : \mathbb{E}_{Z\sim P_n}\left[g_k(\theta;Z)\right] \le 0, k\in[K]\right\}.$$

We denote

$$F(\theta) = \mathbb{E}_{Z\sim P_0}\left[f(\theta;Z)\right] \quad \text{and} \quad \widehat{F}_n(\theta) = \mathbb{E}_{Z\sim P_n}\left[f(\theta;Z)\right],$$

$$G_k(\theta) = \mathbb{E}_{Z\sim P_n}\left[g_k(\theta; Z)\right] \quad \text{and} \quad \widehat{G}_{kn}(\theta) = \sup_{D_\varphi(Q\|P_n)\le \rho_k/n} \mathbb{E}_{Z\sim Q}\left[g_k(\theta; Z)\right] \quad \text{for} \quad k \in [K],$$

and

$$\boldsymbol{G}(\theta) = \begin{pmatrix} G_1(\theta) \\ \vdots \\ G_K(\theta) \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{G}}_n(\theta) = \begin{pmatrix} \widehat{G}_{1n}(\theta) \\ \vdots \\ \widehat{G}_{Kn}(\theta) \end{pmatrix}.$$

Note that $\widehat{F}_n(\cdot)$ and $\widehat{G}_{kn}(\cdot)$'s are random functions serving as approximations to $F(\cdot)$ and $G_k(\cdot)$'s. Consider the Lagrangian functions

$$L(\theta, \boldsymbol{\lambda}) = F(\theta) + \boldsymbol{\lambda}^\top \boldsymbol{G}(\theta) = F(\theta) + \sum_{k=1}^K \lambda_k G_k(\theta)$$

and

$$\widehat{L}_n(\theta, \boldsymbol{\lambda}) = \widehat{F}_n(\theta) + \boldsymbol{\lambda}^\top \widehat{\boldsymbol{G}}_n(\theta) = \widehat{F}_n(\theta) + \sum_{k=1}^K \lambda_k \widehat{G}_{kn}(\theta)$$

of the programs $(\mathcal{P}_0)$ and $(\mathcal{P}_n)$ respectively.

Lemma B.1 (Theorem 6.6.2 in [32])**.** *Suppose that:*

(i) The functions $F(\theta)$ and $G_k(\theta)$, $k \in [K]$, are twice continuously differentiable.

(ii) The true program $(\mathcal{P}_0)$ has a unique optimal solution $\theta^\star$ and a unique vector $\boldsymbol{\lambda}^\star$ of the Lagrange multipliers with $\theta^\star$ being an interior point of Θ.

(iii) The Hessian matrix $\nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star)$ is positive definite.

(iv) The random functions $\widehat{G}_{kn}(\theta)$,$k \in [K]$, are Lipschitz continuous in a neighborhood of $\theta^\star$ and differentiable at $\theta^\star$ with probability 1.

(v)

$$\|\Delta_{in}(\theta^\star)\|_2 = O_p(n^{-1/2}), \quad i = 1, 2, 3$$

and there is a neighborhood U of $\theta^\star$ such that

$$\sup_{\theta\in U} \frac{\|\Delta_{in}(\theta) - \Delta_{in}(\theta^\star)\|_2}{n^{-1/2} + \|\theta - \theta^\star\|_2} = o_p(1), \quad i = 1, 2, 3.$$

Here we define random mappings $\Delta_{1n}(\theta) = \nabla\widehat{F}_n(\theta) - \nabla F(\theta)$, $\Delta_{2n}(\theta) = \widehat{\boldsymbol{G}}_n(\theta) - \boldsymbol{G}(\theta)$, and $\Delta_{3n}(\theta) = \nabla\widehat{\boldsymbol{G}}_n(\theta) - \nabla\boldsymbol{G}(\theta)$.

(vi) Random vectors $\sqrt{n}(\nabla\widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star), \widehat{\boldsymbol{G}}_n(\theta^\star))$ converge in distribution as $n \to \infty$ to a random vector $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$.

Let $\widehat{\theta}_n$ be an optimal solution of $(\mathcal{P}_n)$ converging in probability as $n \to \infty$ to $\theta^\star$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}}(\boldsymbol{Y})$$

where $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}(\boldsymbol{Y})$ is the optimal solution to the quadratic programming problem

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \boldsymbol{x}^\top \boldsymbol{Y}_1 + \frac{1}{2}\boldsymbol{x}^\top \nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star)\boldsymbol{x} \\ \text{subject to} & \nabla\boldsymbol{G}(\theta^\star)^\top \boldsymbol{x} + \boldsymbol{Y}_2 = \boldsymbol{0} \end{array}.$$

From now on, we use $\varphi(t) = (t-1)^2$, which gives the χ^2-divergence. Lemma B.1 is adapted from Theorem 6.6.2 in [32] under the strict complementarity assumption. Recall the standing assumptions, *(i)*, *(iv)*, *(v)* are guaranteed by the smoothness and concentration assumption, *(ii)* is postulated by the uniqueness assumption, and *(iii)* is ensured by the positive definiteness assumption. Now we derive the limiting distribution required in *(vi)*.

Let $\mathbb{B}$ denote the ℓ_2-ball of radius 1 in $\mathbb{R}^d$. According to Lemma 24 in [14], for each $k \in [K]$ there exists $\epsilon_k > 0$ such that, with probability 1, there exists an N_k such that $n \ge N_k$ implies

$$\widehat{G}_{kn}(\theta) = \mathbb{E}_{P_n}[g_k(\theta; Z)] + \sqrt{\frac{\rho_k \operatorname{Var}_{P_n}[g_k(\theta; Z)]}{n}} \text{ for all } \theta \in \theta^\star + \epsilon_k\mathbb{B}.$$

Taking $\epsilon_0 = \min\{\epsilon_k : k \in [K]\}$ and $N_0 = \max\{N_k : k \in [K]\}$, we have the following uniform expansion holds, that is,

$$\widehat{G}_{kn}(\theta) = \mathbb{E}_{P_n}[g_k(\theta; Z)] + \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_n}[g_k(\theta; Z)]} \text{ for all } k \in [K] \text{ and } \theta \in \theta^\star + \epsilon_0 \mathbb{B}$$

P_0-almost surely given $n \geq N_0$.

Therefore, for sufficiently large n, for $k \in [K]$ we have

$$\begin{aligned}
\widehat{G}_{kn}(\theta^\star) &= \mathbb{E}_{P_n}[g_k(\theta^\star; Z)] + \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_n}[g_k(\theta; Z)]} \\
&= \mathbb{E}_P[g_k(\theta^\star; Z)] + \{\mathbb{E}_{P_n}[g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[g_k(\theta^\star; Z)]\} + \sqrt{\frac{\rho_k}{n} \left(\operatorname{Var}_{P_0}[g_k(\theta^\star; Z)] + o_P(1)\right)} \\
&= G_k(\theta^\star) + \{\mathbb{E}_{P_n}[g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[g_k(\theta^\star; Z)]\} + \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_0}[g_k(\theta^\star; Z)]} + o_P(n^{-1/2})
\end{aligned}$$

and

$$\begin{aligned}
\nabla \widehat{G}_{kn}(\theta^\star) =& \mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] + \nabla \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_n}[g_k(\theta; Z)]} \\
=& \mathbb{E}_P[\nabla g_k(\theta^\star; Z)] + \{\mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla g_k(\theta^\star; Z)]\} + \\
& \sqrt{\frac{\rho_k}{n}} \frac{\mathbb{E}_{P_n}[(\nabla g_k(\theta^\star, X) - \mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)])\,(g_k(\theta^\star; Z) - \mathbb{E}_{P_n}[g_k(\theta^\star; Z)])]}{\sqrt{\operatorname{Var}_{P_n}[g_k(\theta^\star; Z)]}} \\
=& \nabla \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] + \{\mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla g_k(\theta^\star; Z)]\} + \\
& \sqrt{\frac{\rho_k}{n}} \left(\frac{\operatorname{Cov}_{P_0}(\nabla g_k(\theta^\star; Z), g_k(\theta^\star; Z))}{\sqrt{\operatorname{Var}_{P_0}[g_k(\theta^\star; Z)]}} + o_P(1) \right) \\
=& \nabla G_k(\theta^\star) + \{\mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla g_k(\theta^\star; Z)]\} + \\
& \sqrt{\frac{\rho_k}{n}} \frac{\operatorname{Cov}_{P_0}(\nabla g_k(\theta^\star; Z), g_k(\theta^\star; Z))}{\sqrt{\operatorname{Var}_{P_0}[g_k(\theta^\star; Z)]}} + o_P(n^{-1/2}).
\end{aligned}$$

For the objective function and its empirical counterpart, we have

$$\widehat{F}(\theta^\star) = \mathbb{E}_{P_n}[f(\theta^\star; Z)] = F(\theta^\star) + \{\mathbb{E}_{P_n}[f(\theta^\star; Z)] - \mathbb{E}_{P_0}[f(\theta^\star; Z)]\}$$

and

$$\nabla \widehat{F}(\theta^\star) = \mathbb{E}_{P_n}[\nabla f(\theta^\star; Z)] = \nabla F(\theta^\star) + \{\mathbb{E}_{P_n}[\nabla f(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla f(\theta^\star; Z)]\}.$$

Now we derive the the limiting distribution of random vectors $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star), \widehat{\boldsymbol{G}}_n(\theta^\star))$. For simplicity of notations, we denote $P_n m = \mathbb{E}_{P_n}[m(\theta^\star; Z)]$ and $P_0 m = \mathbb{E}_{P_0}[m(\theta^\star; Z)]$ for any random function $m(\theta; Z)$, $\operatorname{Cov}(\nabla g_k, g_k) = \operatorname{Cov}_{P_0}(\nabla g_k(\theta^\star; Z), g_k(\theta^\star; Z))$ and $\operatorname{Var}[g_k] =$

$\mathrm{Var}_{P_0}[g_k(\theta^\star; Z)]$. Then we have

$$
\begin{aligned}
&\begin{bmatrix} \nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star) \\ \widehat{\boldsymbol{G}}_n(\theta^\star) \end{bmatrix} \\
&= \begin{bmatrix} \nabla \widehat{F}_n(\theta^\star) + \sum_{k=1}^K \lambda_k^\star \nabla \widehat{G}_{kn}(\theta^\star) \\ \vdots \\ \widehat{G}_{kn}(\theta^\star) \\ \vdots \end{bmatrix} \\
&= \begin{bmatrix} \underbrace{\nabla F(\theta^\star) + \sum_{k=1}^K \lambda_k^\star \nabla G_k(\theta^\star)}_{=0 \text{ due to KKT condition}} + (P_n \nabla f - P_0 \nabla f) + \sum_{k=1}^K \lambda_k^\star (P_n \nabla g_k - P_0 \nabla g_k) + \frac{1}{\sqrt{n}} \sum_{k=1}^K \frac{\lambda_k^\star \sqrt{\rho_k}\, \mathrm{Cov}(\nabla g_k, g_k)}{\sqrt{\mathrm{Var}[g_k]}} \\ \vdots \\ \underbrace{G_k(\theta^\star)}_{=0 \text{ due to active constraint}} + (P_n g_k - P_0 g_k) + \sqrt{\frac{\rho_k}{n}\, \mathrm{Var}[g_k]} \\ \vdots \end{bmatrix} \\
&\quad + o_P\left(\frac{1}{\sqrt{n}}\right) \\
&= \frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{k=1}^K \frac{\lambda_k^\star \sqrt{\rho_k}\, \mathrm{Cov}(\nabla g_k, g_k)}{\sqrt{\mathrm{Var}[g_k]}} \\ \vdots \\ \sqrt{\rho_k\, \mathrm{Var}[g_k]} \\ \vdots \end{bmatrix} + \left\{ P_n \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} - P_0 \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} \right\} + o_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

By central limit theorem,

$$
\sqrt{n} \left\{ P_n \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} - P_0 \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} \right\} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),
$$

where

$$
\begin{aligned}
&\Sigma_{11} = \mathrm{Var}_{P_0}\left[\nabla f(\theta^\star; Z) + \sum_{k=1}^K \lambda_k^\star \nabla g_k(\theta^\star; Z)\right] \in \mathbb{R}^{d \times d}, \\
&\Sigma_{12} = \mathrm{Cov}_{P_0}\left(\nabla f(\theta^\star; Z) + \sum_{k=1}^K \lambda_k^\star \nabla g_k(\theta^\star; Z), \boldsymbol{G}(\theta^\star; Z)\right) \in \mathbb{R}^{d \times K}, \\
&\Sigma_{21} = \Sigma_{12}^\top, \\
&\Sigma_{22} = \mathrm{Var}_{P_0}[\boldsymbol{G}(\theta^\star; Z)] \in \mathbb{R}^{K \times K}.
\end{aligned}
$$

By Slutsky's theorem,

$$\begin{aligned}
&\sqrt{n}\begin{bmatrix}\nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star)\\ \widehat{\boldsymbol{G}}_n(\theta^\star)\end{bmatrix}\\
&= \begin{bmatrix}\sum_{k=1}^K \frac{\lambda_k^\star \sqrt{\rho_k}\,\mathrm{Cov}(\nabla g_k, g_k)}{\sqrt{\mathrm{Var}[g_k]}}\\ \vdots \\ \sqrt{\rho_k \,\mathrm{Var}[g_k]}\\ \vdots\end{bmatrix} + \sqrt{n}\left\{ P_n \begin{bmatrix}\nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k\\ \vdots \\ g_k \\ \vdots\end{bmatrix} - P_0 \begin{bmatrix}\nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k\\ \vdots \\ g_k \\ \vdots\end{bmatrix}\right\} + o_P(1)\\
&\xrightarrow{d} \mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right),
\end{aligned}$$

where

$$\begin{aligned}
\mu_1 &= \sum_{k=1}^K \frac{\lambda_k^\star \sqrt{\rho_k}\,\mathrm{Cov}_{P_0}(\nabla g_k(\theta^\star; Z), g_k(\theta^\star; Z))}{\sqrt{\mathrm{Var}_{P_0}[g_k(\theta^\star; Z)]}} \in \mathbb{R}^d,\\
\mu_2 &= \begin{bmatrix}\sqrt{\rho_1 \,\mathrm{Var}_{P_0}[g_1(\theta^\star; Z)]}\\ \vdots \\ \sqrt{\rho_K \,\mathrm{Var}_{P_0}[g_K(\theta^\star; Z)]}\end{bmatrix} \in \mathbb{R}^K.
\end{aligned}$$

Therefore, we conclude that the limiting distribution of $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star), \widehat{\boldsymbol{G}}_n(\theta^\star))$ is

$$(\boldsymbol{Y}_1, \boldsymbol{Y}_2) \sim \mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right).$$

By Lemma (B.1), we have

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}},$$

where $\bar{\boldsymbol{x}}$ is given by the linear system

$$\underbrace{\begin{bmatrix}\nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star) & \nabla \boldsymbol{G}(\theta^\star)\\ \nabla \boldsymbol{G}(\theta^\star)^\top & \boldsymbol{0}\end{bmatrix}}_{\triangleq B} \begin{bmatrix}\bar{\boldsymbol{x}}\\ \bar{\boldsymbol{\lambda}}\end{bmatrix} = -\begin{bmatrix}\boldsymbol{Y}_1\\ \boldsymbol{Y}_2\end{bmatrix} \sim \mathcal{N}\left(-\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right),$$

or

$$\begin{bmatrix}\bar{\boldsymbol{x}}\\ \bar{\boldsymbol{\lambda}}\end{bmatrix} \sim \mathcal{N}\left(-B^{-1}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, B^{-1}\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} B^{-1}\right), \tag{B.1}$$

which implies $\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}} \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ for some $\bar{\mu}$ and $\bar{\Sigma}$ determined by (B.1).

By delta method, we have

$$\sqrt{n}\begin{bmatrix}\mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)]\\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)]\end{bmatrix} = \sqrt{n}\boldsymbol{G}(\widehat{\theta}_n) = \sqrt{n}\{\boldsymbol{G}(\widehat{\theta}_n) - \underbrace{\boldsymbol{G}(\theta^\star)}_{=0}\} \xrightarrow{d} \mathcal{N}(\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\mu}, \nabla \boldsymbol{G}(\theta^\star)^\top \bar{\Sigma} \nabla \boldsymbol{G}(\theta^\star)).$$

Now we calculate $\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\mu}$ and $\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\Sigma} \nabla \boldsymbol{G}(\theta^\star))$.

For notation simplicity, we denote $\nabla^2 L = \nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star), \nabla \boldsymbol{G} = \nabla \boldsymbol{G}(\theta^\star)$ and $H = (\nabla^2 L)^{-1} \nabla \boldsymbol{G}[\nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} \nabla \boldsymbol{G}]^{-1}$. By block matrix inversion, we have

$$B^{-1} = \begin{bmatrix}(\nabla^2 L)^{-1} - H \nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} & H\\ H^\top & -[\nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} \nabla \boldsymbol{G}]^{-1}.\end{bmatrix}$$

By (B.1), we have

$$\bar{\mu} = -\left\{(\nabla^2 L)^{-1} - H \nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1}\right\} \mu_1 - H \mu_2.$$

Note that $\nabla \boldsymbol{G}^\top H = \boldsymbol{I}_K$ and $\nabla \boldsymbol{G}^\top \left\{(\nabla^2 L)^{-1} - H \nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1}\right\} = \boldsymbol{0}_{K \times K}$. We have

$$\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\mu} = -\nabla \boldsymbol{G}^\top \left\{(\nabla^2 L)^{-1} - H \nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1}\right\} \mu_1 - \nabla \boldsymbol{G} H \mu_2 = -\mu_2$$

and

$$
\begin{aligned}
&\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\Sigma} \nabla \boldsymbol{G}(\theta^\star) \\
=&\nabla \boldsymbol{G}^\top \left[\left\{ (\nabla^2 L)^{-1} - H \nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} \right\} \Sigma_{11} + H \Sigma_{21} \right] \underbrace{\left\{ (\nabla^2 L)^{-1} - (\nabla^2 L)^{-1} \nabla \boldsymbol{G} H^\top \right\} \nabla \boldsymbol{G}}_{=\mathbf{0}_{K\times K}} \\
&+ \underbrace{\nabla \boldsymbol{G}^\top \left\{ (\nabla^2 L)^{-1} - H \nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} \right\}}_{=\mathbf{0}_{K\times K}} \Sigma_{12} H^\top \nabla \boldsymbol{G} + \nabla \boldsymbol{G}^\top H \Sigma_{22} H^\top \nabla \boldsymbol{G} \\
=&\Sigma_{22}.
\end{aligned}
$$

Therefore, we conclude that

$$
\sqrt{n} \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \xrightarrow{d} \mathcal{N}(-\mu_2, \Sigma_{22}) \overset{d}{=\!=} \mathcal{N}\left(- \begin{bmatrix} \sqrt{\rho_1 \operatorname{Var}_{P_0}[g_1(\theta^\star; Z)]} \\ \vdots \\ \sqrt{\rho_K \operatorname{Var}_{P_0}[g_K(\theta^\star; Z)]} \end{bmatrix}, \operatorname{Var}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right).
$$

Hence we complete the proof of Theorem 3.1. □

C Proof of Corollary 3.2

Recall that $D = \operatorname{diag}(\operatorname{Var}_{P_0}[g_1(\theta^\star; Z)], \ldots, \operatorname{Var}_{P_0}[g_K(\theta^\star; Z)])$. According to Theorem 3.1, we have

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left\{ \widehat{\theta}_n \text{ is feasible} \right\} \\
=&\lim_{n\to\infty} \mathbb{P}\left\{ \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \in -\mathbb{R}_+^K \right\} \\
=&\lim_{n\to\infty} \mathbb{P}\left\{ \sqrt{n} \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \leq \mathbf{0}_K \right\} \\
=&\mathbb{P}\left\{ \mathcal{N}\left(- \begin{bmatrix} (\rho_1 \operatorname{Var}_{P_0}[g_1(\theta^\star; Z)])^{\frac{1}{2}} \\ \vdots \\ (\rho_K \operatorname{Var}_{P_0}[g_K(\theta^\star; Z)])^{\frac{1}{2}} \end{bmatrix}, \operatorname{Var}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right) \leq \mathbf{0}_K \right\} \\
=&\mathbb{P}\left\{ \mathcal{N}\left(-D^{-\frac{1}{2}} \begin{bmatrix} (\rho_1 \operatorname{Var}_{P_0}[g_1(\theta^\star; Z)])^{\frac{1}{2}} \\ \vdots \\ (\rho_K \operatorname{Var}_{P_0}[g_K(\theta^\star; Z)])^{\frac{1}{2}} \end{bmatrix}, D^{-\frac{1}{2}} \operatorname{Var} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} D^{-\frac{1}{2}} \right) \leq D^{-\frac{1}{2}} \mathbf{0}_K \right\} \\
=&\mathbb{P}\left\{ \mathcal{N}\left(- \begin{bmatrix} \sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_K} \end{bmatrix}, \operatorname{Corr}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right) \leq \mathbf{0}_K \right\} \\
=&\mathbb{P}\left\{ \mathcal{N}\left(\mathbf{0}_K, \operatorname{Corr}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right) \leq \begin{bmatrix} \sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_K} \end{bmatrix} \right\}.
\end{aligned}
$$

Hence we complete the proof of Corollary 3.2. □

D Dual ascent algorithm for (3.2)

We summarize the dual ascent algorithm for solving (3.2) in Algorithm 2. Similar to Algorithm 1, the main cost of Algorithm 2 is incurred in the evaluation of dual function. The dual function

evaluation step is still suitable for stochastic approximation. Therefore, the total cost of Algorithm 2 is comparable to that of a standard distributionally robust optimization problem.

Algorithm 2 Dual ascent algorithm for (3.2)

1: **Input:** starting dual iterate $\boldsymbol{\lambda}_0 = (\lambda_{01}, \ldots, \lambda_{0K})^\top \in \mathbb{R}_+^K$
2: **repeat**
3: Evaluate dual function:

$$(\theta_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t) \leftarrow \arg\min_{\theta, \boldsymbol{\mu} \in \mathbb{R}_+^K, \boldsymbol{\nu}} \frac{1}{n} \sum_{i=1}^n f(\theta; Z_i) + \sum_{k=1}^K \lambda_{tk} \left\{ \frac{1}{n} \sum_{i=1}^n \mu_{tk} \varphi^*\left(\frac{g_k(\theta; Z_i) - \nu_{tk}}{\mu_{tk}}\right) + \mu_{tk} \frac{\rho_k}{n} + \nu_{tk} \right\}$$

4: Dual ascent update:

$$\lambda_{t+1,k} \leftarrow \left[\lambda_{tk} + \eta_t \left\{ \frac{1}{n} \sum_{i=1}^n \mu_{tk} \varphi^*\left(\frac{g_k(\theta_t; Z_i) - \nu_{tk}}{\mu_{tk}}\right) + \mu_{tk} \frac{\rho_\alpha}{n} + \nu_{tk} \right\} \right]_+, k \in [K]$$

5: **until** converged

E Simulations: details and more

In this section, we provide details for simulations on the multi-item newsvendor problem with independent constraints (which we present in the main text), and conduct more simulations on: (1) multi-item newsvendor problem with dependent constraints, and (2) single-item newsvendor problem with a single constraint.

E.1 Multi-item newsvendor problem

First recall that in Section 4, we simulate the frequency of constraint satisfaction for the following multi-item newsvendor problem:

$$\begin{array}{ll} \max_{\theta \in \Theta} & \mathbb{E}_{P_0}\left[p^\top \min\{Z, \theta\} - c^\top \theta\right] \\ \text{subject to} & \mathbb{E}_{P_0}[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+] \le \varepsilon_1 \\ & \mathbb{E}_{P_0}[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+] \le \varepsilon_2 \end{array} \tag{E.1}$$

where $c \in \mathbb{R}_+^d$ is the manufacturing cost, $p \in \mathbb{R}_+^d$ is the sell price, $\theta \in \Theta = [0, 100]^d$ is the number of items in stock, $Z \in \mathbb{R}^d$ is a random variable with probability distribution P_0 representing the demand, and there are d items in total. The distribution P_0 is unknown but we observe IID samples $Z_1, \ldots, Z_n$ from P_0. All of the items have been partitioned into two groups so that the corresponding demand and stock can be written as $Z = (Z^{(1)}, Z^{(2)})$ and $\theta = (\theta^{(1)}, \theta^{(2)})$. The constraints in the problem exclude stock levels that underestimate the demand too much for each group of items, where $\varepsilon_1, \varepsilon_2 > 0$ indicate tolerance level of such underestimation. The target of the problem is to maximize the profit while satisfying the constraints.

We can rewrite the maximization problem (E.1) as a minimization problem with expected value constraints in the form of (3.1), that is,

$$\begin{array}{ll} \min_{\theta \in \Theta} & \mathbb{E}_{P_0}\left[c^\top \theta - p^\top \min\{Z, \theta\}\right] \\ \text{subject to} & \mathbb{E}_{P_0}[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+ - \varepsilon_1] \le 0 \\ & \mathbb{E}_{P_0}[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+ - \varepsilon_2] \le 0 \end{array} \tag{E.2}$$

so that we can apply our method (3.2).

We generate $Z_1, \ldots, Z_n$ IID from multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. In addition, we set the number of items $d = 4$, the mean of the normal distribution $\mu = (10, 10, 10, 10)^\top$, the cost $c = (1, 1, 1, 1)^\top$, the price $p = (2, 2, 2, 2)^\top$, and the tolerance level of underestimation $(\varepsilon_1, \varepsilon_2) = (1, 1)$. Moreover, we partition the items into the group of the first two items and the group of the last two items. We solve the empirical problem with robust constraint:

$$\begin{array}{ll} \min_{\theta \in \Theta} & \mathbb{E}_{\widehat{P}_n}\left[c^\top \theta - p^\top \min\{Z, \theta\}\right] \\ \text{subject to} & \sup_{D_\varphi(P \| \widehat{P}_n) \le \rho/n} \mathbb{E}_P[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+ - \varepsilon_1] \le 0 \\ & \sup_{D_\varphi(P \| \widehat{P}_n) \le \rho/n} \mathbb{E}_P[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+ - \varepsilon_2] \le 0 \end{array} \tag{E.3}$$

using samples of size $n \in \{30, 300, 3000\}$.

For the covariance of the multivariate normal distribution, we consider a exchangeable correlation structure

$$\Sigma = 9 \times \begin{bmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{bmatrix}$$

We solve (E.3) with $\boldsymbol{\rho} = (\rho, \rho)^\top = (z_\alpha, z_\alpha)^\top$ for $\alpha \in \{0.4, 0.25, 0.1, 0.05, 0.005\}$, of which the corresponding $\sqrt{\rho} \in \{0.253, 0.674, 1.281, 1.644, 2.575\}$.

Independent constraints In Section 4, we consider $r = 0$ so that the two constraints are generally uncorrelated with each other. As suggested by the asymptotic theory in Section 3, the nominal probability of constraint satisfaction is $1 - \alpha$ for each constraint and $(1 - \alpha)^2$ for both constraints due to the independence of two constraints. The results are discussed in Section 4.

Dependent constraints Now we consider $r = 0.6$ so that the two constraints are generally correlated with each other. As suggested by the asymptotic theory in Section 3, the nominal probability of constraint satisfaction is $1 - \alpha$ for each constraint. The nominal probability of constraint satisfaction for both constraints is no longer $(1 - \alpha)^2$ due to the dependence of two constraints, but the probability is given by

$$\mathbb{P}\left\{\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \mathrm{Corr}_{P_0}(g_1(\theta^\star; Z), g_2(\theta^\star; Z)) \\ \mathrm{Corr}_{P_0}(g_1(\theta^\star; Z), g_2(\theta^\star; Z)) & 1\end{bmatrix}\right) \leq_{\mathbb{R}^2} \begin{bmatrix}\sqrt{\rho}\\ \sqrt{\rho}\end{bmatrix}\right\}.$$

In Figure 3, we plot frequencies of constraint satisfaction for each constraint and both constraints, all of which are averaged over 1000 replicates. As the sample size n grows, the frequency versus probability curve converges to the theoretical dashed line of limiting probability of constraint satisfaction, validating our theory in the large sample regime. We note that in this example, the frequency of constraint satisfaction is higher than that of the experiments with independent constraints, for each $\boldsymbol{\rho}$.

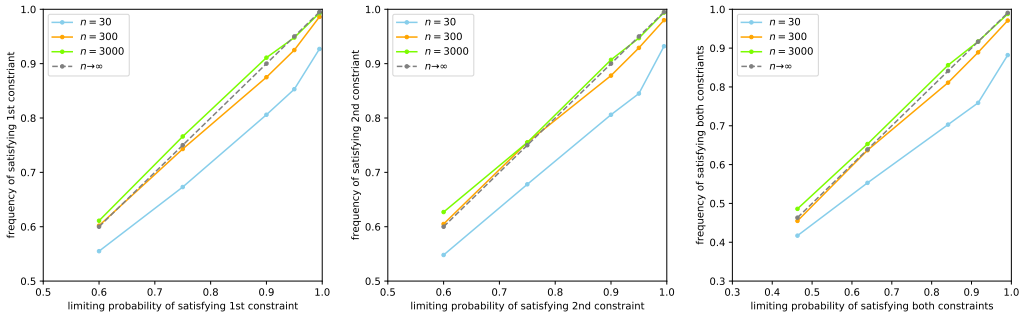

Figure 3: Frequency versus limiting probability of constraint satisfaction of the first constraint (left), the second constraint (middle), and both of the constraints (right).

E.2 Single-item newsvendor problem

In this subsection, we consider the following single-item newsvendor problem:

$$\begin{array}{ll} \max_{\theta \in \Theta} & \mathbb{E}_{P_0}\left[p \min\{Z, \theta\} - c\theta\right] \\ \text{subject to} & \mathbb{E}_{P_0}\left[(Z - \theta)_+\right] \leq \varepsilon \end{array}$$

where $c > 0$ is the manufacturing cost, $p \geq c$ is the sell price, $\theta \in \Theta = [0, 100]$ is the number of items in stock, and Z is a random variable with probability distribution P_0 representing the demand. The distribution P_0 is unknown but instead we observe IID samples $Z_1, \ldots, Z_n$ from P. The constraint in the problem excludes stocking levels that underestimate the demand too much, where $\varepsilon > 0$ indicates tolerance level of such underestimation. The target of the problem is to maximize the profit while

satisfying the constraint. Note that the problem is equivalent to

$$\begin{array}{ll}\min_{\theta\in\Theta} & \mathbb{E}_{P_0}\left[c\theta - p\min\{Z,\theta\}\right] \\ \text{subject to} & \mathbb{E}_{P_0}\left[(Z-\theta)_+ - \varepsilon\right] \le 0\end{array}$$

which is a particular case of (1.1).

We generate $Z_1, \ldots, Z_n$ IID from exponential distribution with mean 10. In addition, we set the cost $c = 1$, the price $p = 2$, and the tolerance level of underestimation $\varepsilon = 1$. We solve the empirical problem with robust constraint:

$$\begin{array}{ll}\min_{\theta\in\Theta} & \mathbb{E}_{\widehat{P}_n}\left[c\theta - p\min\{Z,\theta\}\right] \\ \text{subject to} & \sup_{D_\varphi(P\|\widehat{P}_n)\le\rho/n} \mathbb{E}_P\left[(Z-\theta)_+ - \varepsilon\right] \le 0\end{array}$$

using samples of size $n \in \{300, 3000, 30000\}$.

In Figure 4, we plot frequencies of constraint satisfaction which are averaged over 1000 replicates. As the sample size n grows, the frequency versus probability curve converges to the theoretical dashed line of limiting probability of constraint satisfaction, validating our theory in the large sample regime.

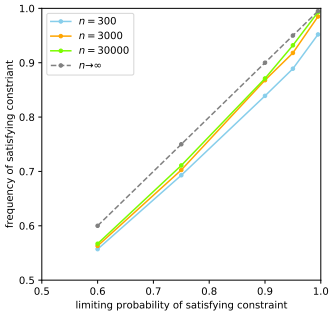

Figure 4: Frequency versus limiting probability of constraint satisfaction.

F Application of our method to ε-equality of opportunity

Continuing with Section 5.1, we demonstrate how to apply our method to enforce ε-equality of opportunity. To keep things simple, we assume there are only two demographic groups; *i.e.* $|\mathcal{A}| = 2$. Without loss of generality, we refer to one group as advantaged ($A = 1$) and the other as disadvantaged ($A = 0$). First we estimate rates $\mathbb{P}\{A = 1, Y = 1\}$ and $\mathbb{P}\{A = 0, Y = 1\}$ consistently by

$$\widehat{p}_1 = \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbf{1}\{A_i = 1, Y_i = 1\} \quad \text{and} \quad \widehat{p}_0 = \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbf{1}\{A_i = 0, Y_i = 1\}.$$

Then, we construct a robust constraint

$$\sup_{P: D_\varphi(P\|\widehat{P}_n)\le\rho/n} \mathbb{E}_{(X,A,Y)\sim P}\left[\frac{\mathbf{1}\{\widehat{Y}=1, A=1, Y=1\}}{\widehat{p}_1} - \frac{\mathbf{1}\{\widehat{Y}=1, A=0, Y=1\}}{\widehat{p}_0} - \varepsilon\right] \le 0,$$

or equivalently

$$\sup_{\boldsymbol{p}:\sum_{i=1}^n \varphi(np_i)\le\rho} \sum_{i=1}^n p_i\left[\frac{\mathbf{1}\{f_\theta(X_i)=1, A_i=1, Y_i=1\}}{\widehat{p}_1} - \frac{\mathbf{1}\{f_\theta(X_i)=1, A_i=0, Y_i=1\}}{\widehat{p}_0} - \varepsilon\right] \le 0.$$

G Two-stage method for unknown active set

In this section, we show that the two-stage method in Section 5.1 also has the calibration property (similar to Theorem 3.1 and Corollary 3.2) if the true program $(\mathcal{P}_0)$, *i.e.* (3.1), is not ill-behaved.

First we recall the two-stage method:

1. At the first stage, we solve the program $(\mathcal{P}_n)$ with $\boldsymbol{\rho} = \mathbf{0}_K$, that is, the sample average approximation (SAA) problem. We identify that $\widehat{J}_{+,n} \subset [K]$ is the active set of the SAA problem, that is, the j-th constraint of the SAA problem is active if and only if $j \in \widehat{J}_{+,n}$.

2. At the second stage, we solve the program $(\mathcal{P}_n)$ with $\boldsymbol{\rho}$ such that ρ_j is some positive number if $j \in \widehat{J}_{+,n}$. This means that at the second stage we replace the sample mean approximation to the constraint by its distributionally robust counterpart if such constraint was identified as active at the first stage.

The index set $[K]$ can be partitioned into three parts:

$$\begin{aligned} J_+(\theta^\star, \boldsymbol{\lambda}^\star) &= \{k \in [K] : \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] = 0, \lambda_k^\star > 0\}, \\ J_0(\theta^\star, \boldsymbol{\lambda}^\star) &= \{k \in [K] : \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] = 0, \lambda_k^\star = 0\}, \\ J_-(\theta^\star, \boldsymbol{\lambda}^\star) &= \{k \in [K] : \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] < 0, \lambda_k^\star = 0\}, \end{aligned}$$

where J_+ is the active set with positive Lagrange multipliers, J_0 is the active set with zero Lagrange multipliers, and J_- is the inactive set. We assume the *strict complementarity* holds in the sense that $J_0(\theta^\star, \boldsymbol{\lambda}^\star) = \varnothing$.

Proposition G.1 (Preservance of active constraints)**.** *Assume the strict complementarity holds. We have $\widehat{J}_{+,n} = J_+(\theta^\star, \boldsymbol{\lambda}^\star)$ with probability converging to 1 as $n \to \infty$.*

Proof of Proposition G.1. Let $\widehat{\theta}_n^{(\mathrm{SAA})}$ be a solution to the sample average approximation problem at the first stage. Let $\widehat{\boldsymbol{\lambda}}_n$ be the associated Lagrange multiplier. It is known by the consistency of SAA [32] that

$$\widehat{\lambda}_{k,n} = \lambda_k^\star + o_p(1) \text{ for } k \in J_+(\theta^\star, \boldsymbol{\lambda}^\star)$$

and

$$\mathbb{E}_P[g_k(\widehat{\theta}_n^{(\mathrm{SAA})}; Z)] = \mathbb{E}_P[g_k(\theta^\star; Z)] + o_p(1) \text{ for } k \in J_-(\theta^\star, \boldsymbol{\lambda}^\star)$$

Therefore, with probability converging to 1, $\widehat{\lambda}_{k,n} > 0$, $k \in J_+(\theta^\star, \boldsymbol{\lambda}^\star)$ and $\mathbb{E}_P[g_k(\widehat{\theta}_n^{(\mathrm{SAA})}; Z)] < 0$, $k \in J_-(\theta^\star, \boldsymbol{\lambda}^\star)$. Hence we complete the proof of Proposition G.1. □

Theorem G.2. *Suppose the true program $(\mathcal{P}_0)$ has m active constraints with positive Lagrange multipliers and $K - m$ inactive constraints (without loss of generality we let $J_+ = \{1, \ldots, m\}$ and $J_- = \{m+1, \ldots, K\}$). Let $\widehat{\theta}_n$ be the two-stage estimator. Under the standing assumptions, we have*

$$\lim_{n\to\infty} \mathbb{P}\left\{ \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \in -\mathbb{R}_+^K \right\} = \mathbb{P}\left\{ \mathcal{N}\left(\mathbf{0}_m, \mathrm{Corr}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_m(\theta^\star; Z) \end{bmatrix} \right) \le \begin{bmatrix} \sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_m} \end{bmatrix} \right\}.$$

This result shows that the limiting probability of satisfying the true constraints only depends on the correlation structure between active constraints and the uncertainty set radii for the active constraints.

Proof of Theorem G.2. At the first stage, we identify $J \subset [K]$ as active set with probability p_J. Here the randomness is introduced by the data samples $Z_1, \ldots, Z_n$. Let $2^{[K]}$ be the power set of $[K]$. We have

$$\sum_{J \in 2^{[K]}} p_J = 1.$$

By Proposition G.1, as $n \to \infty$ we have

$$p_{J_+} \to 1$$

and

$$p_J \to 0 \text{ for any } J \in 2^{[K]} \text{ and } J \neq J_+.$$

By Corollary 3.2, we have

$$\lim_{n\to\infty} \mathbb{P}\left\{ \widehat{\theta}_n \text{ is feasible} \mid \text{identify } J_+ \text{ as active set} \right\} = \mathbb{P}\left\{ \mathcal{N}\left(\mathbf{0}_m, \mathrm{Corr}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_m(\theta^\star; Z) \end{bmatrix} \right) \le \begin{bmatrix} \sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_m} \end{bmatrix} \right\}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible}\right\} \\
&= \sum_{J\in 2^{[K]}} \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible} \mid \text{identify } J \text{ as active set}\right\} \mathbb{P}\left\{\text{identify } J \text{ as active set}\right\} \\
&= \underbrace{p_{J_+}}_{\to 1} \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible} \mid \text{identify } J_+ \text{ as active set}\right\} + \sum_{J\neq J_+} \underbrace{p_J \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible} \mid \text{identify } J \text{ as active set}\right\}}_{\to 0} \\
&\to \mathbb{P}\left\{\mathcal{N}\left(\mathbf{0}_m, \mathrm{Corr}_{P_0}\begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_m(\theta^\star; Z)\end{bmatrix}\right) \leq \begin{bmatrix}\sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_m}\end{bmatrix}\right\}.
\end{aligned}$$

Hence we complete the proof of Theorem G.2. □

H Proxy dual ascent algorithm for non-differentiable constraints

We summarize in Algorithm 3 the proxy dual ascent algorithm for solving a stochastic optimization problem with single non-differentiable constraint. The difference between Algorithm 3 and 1 is in the step of evaluating (proxy) dual function: Algorithm 3 uses a differentiable surrogate $\tilde{g}$ (highlighted in orange) instead of the non-differentiable g in Algorithm 1 in this step.

Algorithm 3 Proxy dual ascent algorithm for single non-differentiable constraint

1: **Input:** starting dual iterate $\lambda_0 \geq 0$
2: **repeat**
3: Evaluate proxy dual function:

$$(\theta_t, \mu_t, \nu_t) \leftarrow \arg\min_{\theta,\mu\geq 0,\nu} \frac{1}{n}\sum_{i=1}^n f(\theta; Z_i) + \lambda_t \left\{\frac{1}{n}\sum_{i=1}^n \mu\varphi^*\left(\frac{\tilde{g}(\theta; Z_i)-\nu}{\mu}\right) + \mu\frac{\rho_\alpha}{n} + \nu\right\}$$

4: Dual ascent update: $\lambda_{t+1} \leftarrow \left[\lambda_t + \eta_t \left\{\frac{1}{n}\sum_{i=1}^n \mu_t\varphi^*\left(\frac{g(\theta_t; Z_i)-\nu_t}{\mu_t}\right) + \mu_t\frac{\rho_\alpha}{n} + \nu_t\right\}\right]_+$
5: **until** converged

Algorithm 4 summarizes the proxy dual ascent algorithm for solving a stochastic optimization problem with multiple non-differentiable constraints. In contrast to Algorithm 2, Algorithm 4 replaces the non-differentiable functions g_k's by their differentiable surrogates $\tilde{g}_k$'s (highlighted in orange) in the (proxy) dual function evaluation step.

Algorithm 4 Proxy dual ascent algorithm for multiple non-differentiable constraints

1: **Input:** starting dual iterate $\boldsymbol{\lambda}_0 = (\lambda_{01}, \ldots, \lambda_{0K})^\top \in \mathbb{R}_+^K$
2: **repeat**
3: Evaluate proxy dual function:

$$(\theta_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t) \leftarrow \arg\min_{\theta,\boldsymbol{\mu}\in\mathbb{R}_+^K,\boldsymbol{\nu}} \frac{1}{n}\sum_{i=1}^n f(\theta; Z_i) + \sum_{k=1}^K \lambda_{tk}\left\{\frac{1}{n}\sum_{i=1}^n \mu_{tk}\varphi^*\left(\frac{\tilde{g}_k(\theta; Z_i)-\nu_{tk}}{\mu_{tk}}\right) + \mu_{tk}\frac{\rho_k}{n} + \nu_{tk}\right\}$$

4: Dual ascent update:

$$\lambda_{t+1,k} \leftarrow \left[\lambda_{tk} + \eta_t\left\{\frac{1}{n}\sum_{i=1}^n \mu_{tk}\varphi^*\left(\frac{g_k(\theta_t; Z_i)-\nu_{tk}}{\mu_{tk}}\right) + \mu_{tk}\frac{\rho_\alpha}{n} + \nu_{tk}\right\}\right]_+, k \in [K]$$

5: **until** converged

I Real data experiments: details and more

In this section, we provide details for experiments on Adult dataset with ε-demographic parity as the fairness goal (which we present in the main text), and conduct additional experiments on Adult dataset with ε-false positive rate parity as the fairness goal.

I.1 Adult with ε-demographic parity

Recall that in Section 5.3 the fairness goal is ε-demographic parity (ε-DP):

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for male is the advantaged group and $A = 0$ for female is the disadvantaged group. We use a logistic regression model for classification by predicting $\widehat{Y} = \mathbf{1}\{\theta^\top X \geq 0\}$ and training such model parameterized in θ by logistic loss. We implement the two-stage method and proxy dual ascent algorithm by replacing indicator function by the sigmoidal function $h_1(t) = (1 + e^{-at})^{-1}$ for $a = 2$. We do bootstrap evaluation by treating the full dataset as the true probability measure and sample such probability measure with replacement as the training distribution (with sampling rate 50%).

I.2 Adult with ε-false positive rate parity

Similar to the preceding subsection, now we consider the fairness goal to be ε-false positive rate parity:

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid Y = 0, A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid Y = 0, A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for male is the advantaged group and $A = 0$ for female is the disadvantaged group.

In Figure 5, we have line plots for frequency of constraint satisfaction and box plots for classification error rate, all of which are summarized over 100 replicates. The patterns shown in the left and right panels are similar to that in Figure 2. We observe more variation in frequencies of constraint satisfaction across different false positive rate tolerance designs due to the label and demographic attribute imbalance of Adult dataset.

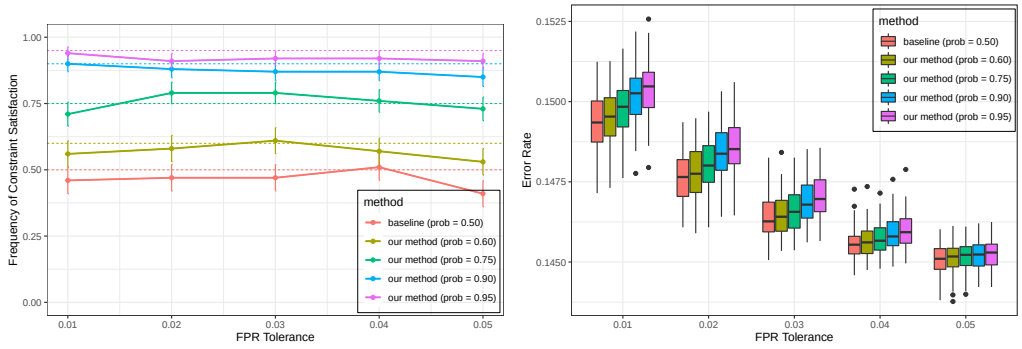

Figure 5: Frequency of constraint satisfaction (left) and classification error rate (right) for different false positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA) and our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$) are compared.

J Dual form of the robust constraint function (2.5)

In this section, we provide a standard derivation for the dual form of the robust constraint function (2.5).

We introduce the likelihood ratio $L(Z) = dP(Z)/d\widehat{P}_n(Z)$. By change of variable, we can rewrite the robust constraint function (2.5) as

$$\sup_{P: D_\varphi(P\|\widehat{P}_n) \leq \rho} \mathbb{E}_P\big[g(\theta; Z)\big] = \sup_{L \geq 0}\{\mathbb{E}_{\widehat{P}_n}\big[L(Z)g(\theta; Z)\big] \mid \mathbb{E}_{\widehat{P}_n}\big[\varphi(L(Z))\big] \leq \rho, \mathbb{E}_{\widehat{P}_n}\big[L(Z)\big] = 1\},$$

where the supremum takes over measurable functions. This gives us a constrained optimization problem. Let $\mu \geq 0$ be the Lagrange multiplier for $\mathbb{E}_{\widehat{P}_n}\big[\varphi(L(Z))\big] \leq \rho$ and $\nu \in \mathbb{R}$ be the Lagrange multiplier for $\mathbb{E}_{\widehat{P}_n}\big[L(Z)\big] = 1$. The corresponding Lagrangian is

$$\mathcal{L}(L, \mu, \nu) = \mathbb{E}_{\widehat{P}_n}\big[(g(\theta; Z) - \nu)L(Z) - \mu\varphi(L(Z))\big] + \mu\rho + \nu.$$

For regular φ-divergence, we have

$$
\begin{aligned}
&\sup_{P:D_\varphi(P\|\widehat{P}_n)\le\rho}\mathbb{E}_P\big[g(\theta;Z)\big]\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\sup_{L\ge0}\mathcal{L}(L,\mu,\nu)\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\sup_{L\ge0}\left\{\sum_{i=1}^n\big[(g(\theta;Z_i)-\nu)L(Z_i)-\mu\varphi(L(Z_i))\big]+\mu\rho+\nu\right\}\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\sup_{L\ge0}\left\{\sum_{i=1}^n\mu\left[\frac{g(\theta;Z_i)-\nu}{\mu}L(Z_i)-\varphi(L(Z_i))\right]+\mu\rho+\nu\right\}\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\left\{\sum_{i=1}^n\mu\sup_{t_i\ge0}\left\{\frac{g(\theta;Z_i)-\nu}{\mu}t_i-\varphi(t_i)\right\}+\mu\rho+\nu\right\}\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\left\{\sum_{i=1}^n\mu\varphi^*\left(\frac{g(\theta;Z_i)-\nu}{\mu}\right)+\mu\rho+\nu\right\}.
\end{aligned}
$$

Here the last equality holds according to the definition of the convex conjugate $\varphi^*(\cdot)$.

K Experiments on additional baseline and dataset

In this section, we conduct more experiments using two-dataset approach of [8] as an additional baseline and default of credit card clients dataset from UCI [13] as an additional dataset.

K.1 Adult with ε-demographic parity (continued)

We continue the experiments on Adult dataset with ε-demographic parity in Section 5.3 and I.1 by adding two-dataset approach of [8] as an additional baseline.

The two dataset approach of [8] splits the training set into two parts, one for updating the model parameters and the other for updating the Lagrangian multipliers, with the goal to improve the generalization of fairness constraints. Figure 6 demonstrates that the two-dataset approach marginally improves the frequency of constraint satisfaction over the SAA. The cost of such an improvement is increased variation in classification error rate. Due to the fact that the two-dataset approach does not use the entire training set to update the model parameters, statistical efficiency is sacrificed. Although the two-dataset approach outperforms the SAA in terms of constraint satisfaction frequency, both are inferior to our methods, which achieve the user-prescribed level of frequency.

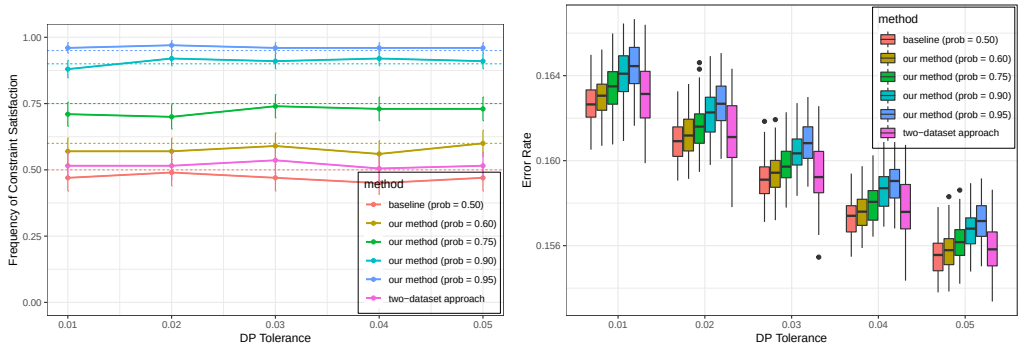

Figure 6: Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon\in\{0.01,0.02,0.03,0.04,0.05\}$. Baseline (sample average approximation, SAA), our methods (with nominal probability $0.60,0.75,0.90,0.95$), and two-dataset approach [8] are compared.

K.2 Adult with ε-false positive rate parity (continued)

We continue the experiments on Adult dataset with ε-false positive rate parity in Section I.2 by adding two-dataset approach of [8] as an additional baseline.

Figure 7 demonstrates similar patterns as Figure 6. The two-dataset approach improves constraint satisfaction frequency over the SAA, but has a worse classification error rate or one comparable to it. Our methods achieve desired frequency of constraint satisfaction at the user-prescribed level.

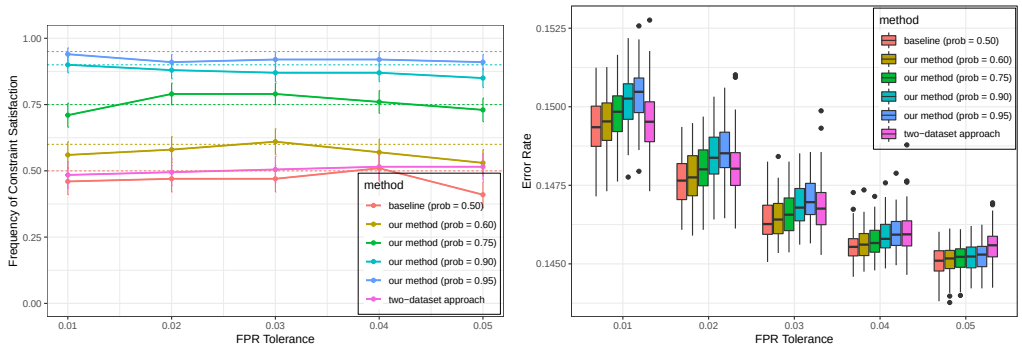

Figure 7: Frequency of constraint satisfaction (left) and classification error rate (right) for different false positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA), our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$), and two-dataset approach [8] are compared.

K.3 Credit with ε-demographic parity

Predicting whether or not an individual defaulted on a loan is the classification task of the UCI default of credit card clients (Credit) dataset. Membership in the demographic group is determined by an individual's level of education: $A = 1$ if a person has earned a graduate degree; otherwise, $A = 0$.

We consider the fairness goal to be ε-demographic parity:

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for individuals with a graduate degree is the advantaged group.

As depicted in Figure 8, our methods achieve the level of constraint satisfaction frequency specified by the user at the expense of a slight increase in classification error.

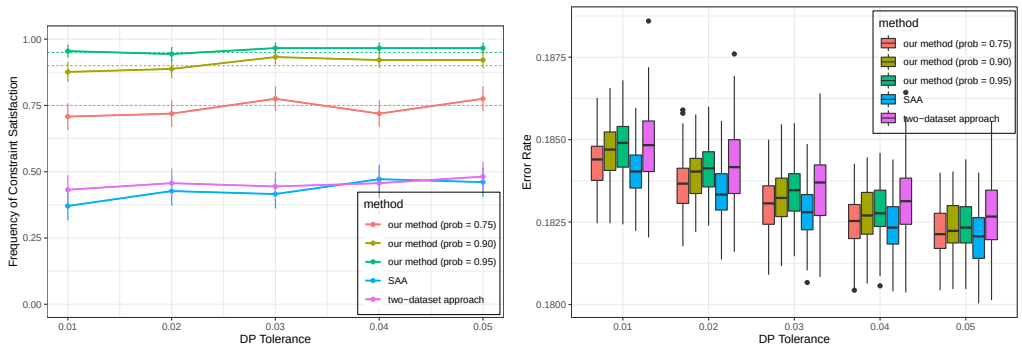

Figure 8: Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Our methods (with nominal probability $0.75, 0.90, 0.95$), sample average approximation (SAA) and two-dataset approach [8] are compared.

K.4 Credit with ε-true positive rate parity

Similar to the preceding subsection, now we consider the fairness goal to be ε-true positive rate parity:

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid Y = 1, A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid Y = 1, A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for individuals with a graduate degree is the advantaged group.

The patterns shown in Figure 9 are similar to that in Figure 8. The SAA has the lowest error rate but the worst generalization of constraint satisfaction. The two-dataset approach increases the SAA's frequency of constraint satisfaction at the cost of an increase in classification error rates. Our methods can achieve the desired high probability of constraint satisfaction, whereas neither of the two baselines can.

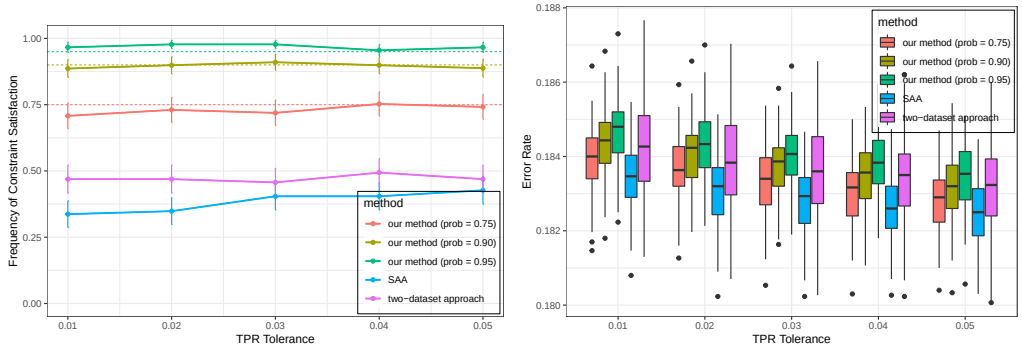

Figure 9: Frequency of constraint satisfaction (left) and classification error rate (right) for different true positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Our methods (with nominal probability $0.75, 0.90, 0.95$), sample average approximation (SAA) and two-dataset approach [8] are compared.