Non-Convex Robust Hypothesis Testing using Sinkhorn Uncertainty Sets

Jie Wang, Rui Gao, Yao Xie

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619088

***Abstract*—We present a new framework to address the non-convex robust hypothesis testing problem, wherein the goal is to seek the optimal detector that minimizes the maximum of worst-case type-I and type-II risk functions. The distributional uncertainty sets are constructed to center around the empirical distribution derived from samples based on Sinkhorn discrepancy. Given that the objective involves non-convex, non-smooth probabilistic functions that are often intractable to optimize, existing methods resort to approximations rather than exact solutions. To tackle the challenge, we introduce an exact mixed-integer exponential conic reformulation of the problem, which can be solved into a global optimum with a moderate amount of input data. Subsequently, we propose a convex approximation, demonstrating its superiority over current state-of-the-art methodologies in literature. Furthermore, we establish connections between robust hypothesis testing and regularized formulations of non-robust risk functions, offering insightful interpretations.**

I. Introduction

Hypothesis testing is a fundamental problem in statistics, whose primary goal is to decide the true hypothesis while minimizing the risk of wrong decisions. Hypothesis testing is a building block for various statistical problems such as change-point detection [1]–[5], model criticism [6]–[8], and it has applications in broad domains including healthcare [9]. Practically, the underlying true distributions corresponding to each hypothesis are unknown, and we only have access to a small amount of data collected for each hypothesis.

Distributionally robust hypothesis testing has emerged as a popular approach to tackle the challenge of establishing an optimal decision in the presence of limited sample size, model misspecification, and adversarial data perturbation. It formulates the problem as seeking the optimal decision over uncertainty sets that contain candidate distributions for each hypothesis. The construction of such distributional uncertainty sets plays a key role in both computational tractability and testing performance.

The earlier work of finding a robust detector dates back to Huber's seminar work [10], which constructs the uncertainty sets as total-variation probability balls centered around the reference distributions. Unfortunately, the computational complexity of seeking the optimal detector within this framework, particularly for multivariate distributions, hinders its practical applications. Two primary approaches have emerged for constructing uncertainty sets in robust hypothesis testing. The first involves defining uncertainty sets using descriptive statistics such as moment conditions [11]. The second approach considers all possible distributions within a pre-specified statistical divergence from a reference distribution. Commonly adopted statistical divergences include the KL-divergence [12], [13], Wasserstein distance [4], [14], [15], entropic regularized Wasserstein distance (i.e., Sinkhorn discrepancy) [16], and maximum mean discrepancy [17].

It is noteworthy that distributionally robust optimization (DRO) with Sinkhorn discrepancy-based uncertainty set has recently received great attention in the literature [16], [18]–[23], mainly due to its data-driven nature, computational tractability, and flexibility to obtain worst-case distributions yielding satisfactory performance. Considering its empirical success, we propose a new framework for robust hypothesis testing, whose distributional uncertainty sets are constructed using the Sinkhorn discrepancy. Our goal is to seek the optimal detector to minimize the *maximum of worst-case type-I and type-II error*. In contrast to the recent works [14], [16] considering a special *smooth* and *convex* relaxation of the objective function, we aim to solve the non-convex problem by (i) either *directly* optimizing the probabilistic objective, or (ii) providing a *tighter* convex relaxation.

Our proposed framework balances the trade-off between computational efficiency and statistical testing performance. The contributions are summarized as follows. Proofs and numerical study for our framework can be found in [24].

1) Under the random feature model, we obtain a finite-dimensional optimization reformulation for this robust hypothesis testing problem (Section II). Besides, we provide a closed-form expression for the worst-case distributions (Remark 1).
2) We develop novel optimization algorithms for the non-convex robust testing problem. First, we provide an exact mixed-integer conic reformulation of the problem (Section III-A), enabling the attainment of global optimum even with a moderate data size. Subsequently, we introduce a convex approximation and illustrate its superiority as a tighter relaxation compared to the state-of-the-art (Section III-B).
3) We connect robust hypothesis testing and regularized formulations of non-robust risk functions under two hyper-parameter scaling regimes, offering insightful in-

J. Wang and Y. Xie are with H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. R. Gao is with Department of Information, Risk, and Operations Management, University of Texas at Austin. This work is partially supported by an NSF CAREER CCF-1650913, NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, DMS-1830210, and the Coca-Cola Foundation.

terpretations of Sinkhorn robust testing (Section IV).

Notations. The base of the logarithm function $\log$ is e. For any positive integer N, define $[N] = \{1, \ldots, N\}$. For scalar $x \in \mathbb{R}$, define $(x)_+ = \max\{x, 0\}$. Let $\mathcal{K}_{\exp}$ denote the exponential cone:

$$\mathcal{K}_{\exp} = \Big\{(\nu, \mu, \delta) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} : \ e^{\delta/\nu} \leq \mu/\nu\Big\}.$$

For a given event E, define the indicator function $\mathbf{1}_E(\cdot)$ such that $\mathbf{1}_E(z) = 1$ if $z \in E$ and otherwise $\mathbf{1}_E(z) = 0$. Given a function $\phi : \ \Omega \to \mathbb{R}$ and scalar $r \in [1, \infty]$, define the norm $\|\phi\|_{L^r} = \big(\int_\Omega \phi(x)^r \mathrm{d}x\big)^{1/r}$. For a non-negative measure ν, define the norm $\|\phi\|_{L^r(\nu)} = \big(\int_\Omega \phi(x)^r \mathrm{d}\nu(x)\big)^{1/r}$.

II. Problem Setup

Let $\Omega \subseteq \mathbb{R}^d$ be the sample space where the observed samples take their values, and $\mathcal{P}(\Omega)$ be the set of all distributions supported on Ω. Denote by $\mathcal{P}_1, \mathcal{P}_2 \subseteq \mathcal{P}(\Omega)$ the uncertainty sets under hypotheses H_1 and H_2, respectively. Given two sets of training samples $\{x_1^k, \ldots, x_{n_k}^k\}$ generated from $\mathbb{P}_k \in \mathcal{P}_k$ for $k = 1, 2$, denote the corresponding empirical distributions as $\widehat{\mathbb{P}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \delta_{x_i^k}$. For notation simplicity, assume that $n_0 = n_1 = n$, but our formulation can be naturally extended for unequal sample sizes. Given a new testing sample ω, the goal of *composite hypothesis testing* is to distinguish between the null hypothesis $H_1 : \ \omega \sim \mathbb{P}_1$ and the alternative hypothesis $H_2 : \ \omega \sim \mathbb{P}_2$, where $\mathbb{P}_k \in \mathcal{P}_k$ for $k = 1, 2$. For a detector $T : \ \Omega \to \mathbb{R}$, it accepts the null hypothesis H_1 when $T(\omega) \geq 0$; otherwise, it accepts the alternative hypothesis H_2. Under the Bayesian setting, we quantify the risk of this detector as the *maximum of the worst-case type-I and type-II errors*:

$$\mathcal{R}(T; \mathcal{P}_1, \mathcal{P}_2) = \max_{k=1,2} \sup_{\mathbb{P}_k \sim \mathcal{P}_k} \mathbb{P}_k\{\omega : (-1)^{k+1} T(\omega) < 0\}.$$

In this paper, we aim to find the detector T such that its risk is minimized:

$$\inf_{T:\ \Omega \to \mathbb{R}} \ \mathcal{R}(T; \mathcal{P}_1, \mathcal{P}_2). \tag{1}$$

It is worth noting that there are three major challenges when solving such a formulation: (i) First, seeking the optimal detector among all measurable functions is an infinite dimensional optimization problem, which is intractable; (ii) Second, finding the worst-case distributions over ambiguity sets $\mathcal{P}_1, \mathcal{P}_2$ is also an infinite dimensional optimization, which is not always tractable; (iii) Finally, the objective involves probability functions, which are non-smooth and non-convex. In the following, we provide methodologies to tackle these difficulties.

A. Random Feature Model

We use the random feature model to address challenges (i). Consider the following assumption on the space of detectors.

Assumption 1 (Optimal Detector in RKHS)**.** The underlying true detector $T^* : \ \Omega \to \mathbb{R}$ belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{F}_K$ equipped with a kernel function $K(x, y) = \mathbb{E}_{\omega \sim \pi_0}[\phi(x; \omega)\phi(y; \omega)]$ for some feature map ϕ and feature distribution π_0. Besides, there exists a constant $M > 0$ such that for π_0-almost ω, it holds that $\|\phi(\cdot; \omega)\|_{L^2} \leq M$. ♣

We highlight that such a restriction does not limit the generality. Instead, since the RKHS (with universal kernel choice, such as Gaussian kernel) is dense in the continuous function space, i.e., it approximates any continuous function within arbitrarily small error. For commonly used kernels, the feature map expressions are also easily satisfied. For example, when considering the kernel function to be continuous, real-valued, and shift-invariant, by Bochner's Theorem [25], it holds that

$$K(x, y) = \mathbb{E}_{(z,b)}[\cos(z^{\mathrm{T}} x + b)\cos(z^{\mathrm{T}} y + b)],$$

where the vector z follows the distribution from the density function $p(\omega) = \frac{1}{2\pi}\int e^{-\mathbf{i}\langle \omega, \delta\rangle} K(\delta)\mathrm{d}\delta$, $\mathbf{i} = \sqrt{-1}$, and scalar b follows the uniform distribution supported on $[0, 2\pi]$. In such case, Assumption 1 holds by taking $\phi(x; \omega) := \cos(z^{\mathrm{T}} x + b)$ with $\omega := (z, b)$.

For any detector $T \in \mathcal{F}_K$, there exists a function $\theta(\cdot) \in L^2(\pi_0)$ such that

$$T(x) = \mathbb{E}_{\omega \sim \pi_0}[\theta(\omega)\phi(x; \omega)].$$

Denote the feature vector

$$\Phi(x) = \left(\frac{1}{D}\phi(x; \omega_1), \ldots, \frac{1}{D}\phi(x; \omega_D)\right) \in \mathbb{R}^D,$$

with $\{\omega_i\}_{i \in [D]}$ being i.i.d. samples generated from π_0, and the vector $\bar{\theta} = (\theta(\omega_1), \ldots, \theta(\omega_D))$. Then, the random feature model

$$\hat{T}(x) = \langle \bar{\theta}, \Phi(x)\rangle = \frac{1}{D}\sum_{i \in [D]} \theta(\omega_i)\phi(x; \omega_i)$$

is an unbiased estimator of $T(x)$ with respect to $\{\omega_i\}$. This motivates us to propose

$$\mathcal{F}_D = \big\{T : \ x \mapsto \langle \theta, \Phi(x)\rangle, \ \exists \theta \in \mathbb{R}^D\big\}$$

as an approximation of $\mathcal{F}_K$. The following presents the approximation theoretical guarantees for $\mathcal{F}_D$. Similar results have also been explored in [26, Theorem 5].

Proposition 1 (Direct Approximation Theorem)**.** *Fix the error probability $\alpha \in (0, 1)$. Suppose Assumption 1 holds and define*

$$\|T^*\|_\infty := \inf_{\theta(\cdot)} \left\{\|\theta(\cdot)\|_{L^\infty(\pi_0)} : \ T^*(x) = \mathbb{E}_{\pi_0}[\theta(\omega)\phi(x; \omega)]\right\}.$$

Then there exists a function T in $\mathcal{F}_D$ such that with probability at least $1 - \alpha$, it holds that

$$\|T^* - T\|_{L^2} \leq \frac{M}{\sqrt{D}}\left(\|T^*\|_{\mathcal{F}_K} + \sqrt{2\|T\|_\infty^2 \log\frac{1}{\alpha}}\right). \qquad \diamondsuit$$

By Proposition 1, the number of samples $D = \Omega(\frac{1}{\epsilon^2}\log\frac{1}{\alpha})$ is enough to control the approximation error within ϵ with probability at least $1-\alpha$, which is *data dimension independent*. This justifies that the random feature model is a suitable choice for approximate detectors, especially for high-dimensional

scenarios. As a consequence, we obtain the finite-dimensional reformulation of Problem (1):

$$\begin{aligned} \min_{\theta\in\mathbb{R}^D, s\geq 0} \quad & s \\ \text{s.t.} \quad & \sup_{\mathbb{P}_k\in\mathcal{P}_k} \mathbb{P}_k\{\omega:\ (-1)^{k+1}\langle\theta,\Phi(\omega)\rangle<0\}\leq s, \quad k=1,2. \end{aligned} \tag{2}$$

From the above, we find θ is optimal implies $\alpha\theta$ is also optimal for any $\alpha\in\mathbb{R}_+$. To make the solution θ well-conditioned, we additionally add the constraint $\|\theta\|_2\leq 1$ when solving (2).

B. Preliminaries on Sinkhorn DRO

In the following, we specify uncertainty sets $\mathcal{P}_k, k=1,2$ using Sinkhorn discrepancy and discuss the corresponding tractable reformulation.

Assumption 2 (Sinkhorn Uncertainty Sets)**.** For $k=1,2$, we specify the uncertainty set

$$\mathcal{P}_k=\Big\{\mathbb{P}:\ \mathcal{W}_{\varepsilon_k}(\mathbb{P},\widehat{\mathbb{P}}_k)\leq\rho_k\Big\}, \tag{3}$$

where the Sinkhorn discrepancy $\mathcal{W}_\varepsilon(\cdot,\cdot)$ is defined as

$$\mathcal{W}_\varepsilon(\mathbb{P},\mathbb{Q})\ =\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}\ \left\{\mathbb{E}_{(x,y)\sim\gamma}[c(x,y)]+\varepsilon H(\gamma)\right\}.$$

Here $\Gamma(\mathbb{P},\mathbb{Q})$ denotes the set of joint distributions whose first and second marginal distributions are $\mathbb{P}$ and $\mathbb{Q}$ respectively, $c(x,y)$ denotes the transport cost, and $H(\gamma)$ denotes the relative entropy of γ with respect to product measure $\mathbb{P}\otimes\Lambda$, where $\Lambda(\cdot)$ denotes the Lebesgue measure on Ω:

$$H(\gamma)\ =\mathbb{E}_{(x,y)\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mathbb{P}(x)\mathrm{d}y}\right)\right]. \qquad \clubsuit$$

Subsequently from Assumption 2, we define the radii

$$\overline{\rho}_k\triangleq\rho_k+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\varepsilon_k\log\int e^{-c(x,z)/\varepsilon_k}\mathrm{d}z\Big], \quad k=1,2. \tag{4}$$

With a measurable variable $f:\ \Omega\to\mathbb{R}$, we associate value

$$V=\sup_{\mathbb{P}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{P}}[f], \tag{5}$$

where the ambiguity set $\mathcal{P}$ is in the form of (3). Define the dual problem of (5) as

$$V_D=\inf_{\lambda\geq 0}\ \Big\{\lambda\overline{\rho}+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\lambda\varepsilon\log\mathbb{E}_{z\sim\mathbb{Q}_{x,\varepsilon}}\left[e^{f(z)/(\lambda\varepsilon)}\right]\Big]\Big\}, \tag{6}$$

where we define the constant

$$\overline{\rho}=\rho+\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\varepsilon\log\int e^{-c(x,z)/\varepsilon}\mathrm{d}z\Big] \tag{7}$$

and $\mathbb{Q}_{x,\varepsilon}$ as the kernel probability distribution with density

$$\frac{\mathrm{d}\mathbb{Q}_{x,\varepsilon}(z)}{\mathrm{d}z}\propto e^{-c(x,z)/\varepsilon}. \tag{8}$$

For example, when considering the optimal transport cost function $c(x,z)=\frac{1}{2}\|x-z\|_2^2$, $\mathbb{Q}_{x,\varepsilon}$ reduces to the Gaussian distribution $\mathcal{N}(x,\varepsilon\mathbf{I}_D)$. By [18, Theorem 1], V_D defined in (6) is the dual reformulation of Problem (5). This observation indicates the computational tractability when using Sinkhorn uncertainty sets: solving the worst-case expectation problem in (5) is *always* tractable when solving its one-dimensional dual problem in (6) using the random sampling approach developed in [18, Section 4]. Besides, we usually tune the radius $\overline{\rho}$ that appeared in dual formulation instead of the original radius ρ.

Proposition 2 (Reformulation of Sinkhorn DRO)**.** *Suppose that $\int e^{-c(x,z)/\varepsilon}\mathrm{d}z<\infty$ for $\widehat{\mathbb{P}}$-almost every x and $\overline{\rho}\geq 0$, then it holds that $V=V_D$.* $\diamondsuit$

Remark 1 (Recovery of Worst-case Distributions)**.** *After solving Problem* (2) *to obtain (near-)optimal solution (θ^*,s^*), one can recover the worst-case distributions corresponding to hypothesis H_1 and H_2 based on [18, Remark 4], denoted as $\mathbb{P}_k^*, k\in\{1,2\}$. Assume the optimal Lagrangian multipliers $(\lambda_k^*)_{k=1,2}$ to Problem* (2) *is positive, then the density of the worst-case distribution, denoted as $\mathrm{d}\mathbb{P}_k^*(z)/\mathrm{d}z$, becomes*

$$\mathbb{E}_{x\sim\widehat{\mathbb{P}}_k}\left[\alpha_x\cdot\exp\left(\frac{\mathbf{1}\{(-1)^{k+1}\langle\theta^*,\Phi(z)\rangle<0\}-\lambda_k^*c(x,z)}{\lambda_k^*\varepsilon_k}\right)\right],$$

where α_x is a normalizing constant. $\square$

III. Optimization Methodology

In this section, we first discuss how to solve the formulation (2) directly based on mixed-integer conic programming and then talk about how to solve its convex relaxation using the CVaR approximation approach.

A. A Mixed-Integer Conic Formulation

According to the definition of ambiguity sets $\mathcal{P}_k, k=1,2$ and Proposition 2, as probabilistic constraints can always be written as expectations of indicator functions, Problem (2) can be reformulated as

$$\vartheta^*=\min_{\substack{\|\theta\|_2\leq 1, s\geq 0,\\ \lambda_1,\lambda_2\geq 0}}\ \Big\{s:\ F_k(\theta,\lambda_k)\leq s, \quad k=1,2\Big\}, \tag{9}$$

where the function F_k is defined as

$$\begin{aligned} F_k(\theta,\lambda_k)\triangleq\lambda_k\overline{\rho}_k+\mathbb{E}_{x\sim\widehat{\mathbb{P}}_k}\Big[\lambda_k\varepsilon_k\cdot \\ \log\mathbb{E}_{y\sim\mathbb{Q}_{x,\varepsilon_k}}\Big[\exp\Big\{\frac{\mathbf{1}\{(-1)^{k+1}\langle\theta,\Phi(y)\rangle<0\}}{\lambda_k\varepsilon_k}\Big\}\Big]\Big]. \end{aligned}$$

Here, the radii $\overline{\rho}_k$ and distribution $\mathbb{Q}_{x,\varepsilon_k}$ are defined in (4) and (8), respectively. Next, we adopt the idea of sample average approximation (SAA) to approximate those two constraints in (9). Recall that $\widehat{\mathbb{P}}_k=\frac{1}{n}\sum_{i=1}^n\delta_{x_i^k}, k\in\{1,2\}$. For each sample x_i^k, we generate m i.i.d. sample points $y_{i,j}^k$ following distribution $\mathbb{Q}_{x_i^k}$ for $j\in[m]$. Hence, we obtain the sample estimates of functions F_k for $k=1,2$:

$$\begin{aligned} \widetilde{F}_k(\theta,\lambda_k)=\lambda_k\overline{\rho}_k+ \\ \frac{\lambda_k\overline{\rho}_k}{n}\sum_{i\in[n]}\log\left[\frac{1}{m}\sum_{j\in[m]}e^{\frac{\mathbf{1}\{(-1)^{k+1}\langle\theta,\Phi(y_{i,j}^k)\rangle<0\}}{\lambda_k\varepsilon_k}}\right]. \end{aligned}$$

Consequently, the sample estimate of the optimal value ϑ^* defined in (9) is given by

$$\widehat{\vartheta}^*=\min_{\substack{\|\theta\|_2\leq 1, s\geq 0,\\ \lambda_1,\lambda_2\geq 0}}\ \Big\{s:\ \widetilde{F}_k(\theta,\lambda_k)\leq s, \quad k=1,2\Big\}. \tag{10}$$

We present a consistency result between $\widehat{\vartheta}^*$ and ϑ^* below.

Proposition 3 (Consistency of $\widehat{\vartheta}^*$)**.** *Assume the radii $\overline{\rho}_k > 0$ for $k = 1, 2$, and there exists an optimal solution $(\theta^*, s^*, \lambda_1^*, \lambda_2^*)$ to* (9) *such that for any $\delta > 0$, there exists $(\theta, s, \lambda_1, \lambda_2)$ with $\|\theta\|_2 \leq 1, s \geq 0, \lambda_1 \geq 0, \lambda_2 \geq 0$, $\|(\theta, s, \lambda_1, \lambda_2) - (\theta^*, s^*, \lambda_1^*, \lambda_2^*)\| \leq \delta$ and $F_k(\theta, \lambda_k) < s^*, k = 1, 2$. As a consequence, $\widehat{\vartheta}^* \to \vartheta^*$.* $\diamondsuit$

The assumptions in Proposition 3 are essential. The first is to ensure the optimal multipliers λ_1, λ_2 exist and are bounded. For the second, assume on the contrary that there exists a case where $F_k(\theta, \lambda_k) \leq s^*$ only defines one feasible point $(\bar{\theta}, \bar{\lambda}_k)$ such that $F_k(\bar{\theta}, \bar{\lambda}_k) = s^*$. Then arbitrarily small perturbations regarding the constraint $\widetilde{F}_k(\theta, \lambda_k) \leq s^*$ may cause the SAA problem (10) infeasible to solve.

Besides, the SAA problem (10) admits a finite-dimensional mixed-integer exponential conic program (MIECP) reformulation. Consequently, the moderate-sized instances of such a formulation could be handled by state-of-the-art solvers [27]–[29] in a reasonable amount of time.

Theorem 1 (MIECP Reformulation of (10))**.** *Assume there exist constants for $i \in [n], j \in [m], k \in \{1, 2\}$:*

$$M_{i,j}^k = \max_{\|\theta\|_2 \leq 1} \ (-1)^{k+1} \langle \theta, \Phi(y_{i,j}^k) \rangle.$$

Then, Problem (10) *is equivalent to*

$$\begin{array}{ll} \text{Minimize} & s \\ \textit{s.t.} & \begin{cases} \|\theta\|_2 \leq 1 \\ (-1)^{k+1} \langle \theta, \Phi(y_{i,j}^k) \rangle \leq M_{i,j}^k (1 - z_{i,j}^k) \end{cases} \\ & \begin{cases} \lambda_k \overline{\rho}_k + \dfrac{1}{n} \displaystyle\sum_{i \in [n]} t_i^k \leq s \\ \lambda_k \varepsilon_k \geq \dfrac{1}{m} \displaystyle\sum_{j \in [m]} a_{i,j}^k \\ (\lambda_k \varepsilon_k, a_{i,j}^k, z_{i,j}^k - t_i^k) \in \mathcal{K}_{\exp}, \\ i \in [n], j \in [m], k \in \{1, 2\} \end{cases} \end{array} \tag{11}$$

subject to the following decision variables

$$\begin{array}{l} s \in [0, 1], \theta \in \mathbb{R}^D, \lambda_1, \lambda_2 \in \mathbb{R}_+, \{t_i^k\}_{i,k} \in \mathbb{R}^{n \times 2}, \\ \{z_{i,j}^k\}_{i,j,k} \in \{0, 1\}^{n \times m \times 2}, \{a_{i,j}^k\}_{i,j,k} \in \mathbb{R}^{n \times m \times 2}. \end{array} \quad \diamondsuit$$

Although Problem (11) can be directly handled by off-the-shelf Mosek solver [30], we do not implement in this way because it involves $2nm$ binary variables and $2nm$ exponential conic constraints, which incurs heavy computational cost. Instead, we solve it using the outer approximation algorithm developed in [28], which iteratively solves the subproblem of (11) for fixed values of binary variables $\{z_{i,j}^k\}_{i,j,k}$ and then update them using the cutting plane algorithm.

B. Convex Approximation

Since the probabilistic constraints in Problem (2) make it intractable to solve, an alternative approach to solving this problem is to construct convex approximations of those constraints. The most popular approach is to replace the probabilistic constraints with the conditional value-at-risk (CVaR) approximation [31], since the following relation holds for any random variable Z and probability level ϵ:

$$\inf_{\beta \leq 0} \ \Big\{ \epsilon \beta + \mathbb{E}[Z - \beta]_+ \Big\} \leq 0 \implies \mathbb{P}\{Z > 0\} \leq \epsilon.$$

Inspired by this approach, we replace two constraints in Problem (2) using the CVaR approximation:

$$\min_{\|\theta\|_2 \leq 1, s \geq 0} \quad s \tag{12a}$$

$$\text{s.t.} \ \sup_{\mathbb{P}_k \in \mathcal{P}_k} \inf_{\beta_k \leq 0} \Big\{ s \beta_k + \mathbb{E}_{\mathbb{P}_k} [(-1)^k \langle \theta, \Phi(\omega) \rangle - \beta_k]_+ \Big\} \leq 0, \forall k. \tag{12b}$$

Remark 2 (Superior Performance of CVaR Approximation)**.** *Recall references [14], [16] used the generating function approach for convex approximation, which can be viewed as a special case of our formulation by specifying $\beta_k = -1, \forall k$ in* (12b)*. In such cases, this constraint becomes*

$$\sup_{\mathbb{P}_k \in \mathcal{P}_k} \ \mathbb{E}_{\mathbb{P}_k} \Big[\ell \circ \Big((-1)^k \langle \theta, \Phi(\omega) \rangle \Big) \Big] \leq s, \quad k = 1, 2,$$

with the generating function $\ell(x) = (x + 1)_+$ that leads to the tightest theoretical approximation ratio proposed in [14, Theorem 1]. Their approaches can be strengthened by taking the optimization over β_k into account. $\square$

Assume the sample space Ω is compact. By Prohorov's Theorem (see, e.g., [32, Theorem 2.4]), the ambiguity sets $\mathcal{P}_k, k = 1, 2$ are compact as well, which ensures that one can apply Sion's minimax Theorem [33] to exchange the sup and inf operators in those two constraints of the problem above. Next, one can leverage the strong duality result in Proposition 2 to obtain its equivalent formulation:

$$\min_{\substack{\|\theta\|_2 \leq 1, s \geq 0, \\ \beta_k \leq 0, \lambda_k \geq 0, k = 1, 2}} \ \Big\{ s : \ G_k(s, \beta_k, \lambda_k) \leq 0, \quad k = 1, 2 \Big\}, \tag{13}$$

where the function G_k is defined as

$$\begin{array}{l} G_k(s, \beta_k, \lambda_k) = s \beta_k + \Big\{ \lambda_k \overline{\rho}_k + \\ \quad \mathbb{E}_{x \sim \widehat{\mathbb{P}}_k} \Big[\lambda_k \varepsilon_k \log \mathbb{E}_{y \sim \mathbb{Q}_{x, \varepsilon_k}} \Big[e^{[(-1)^k \langle \theta, \Phi(y) \rangle - \beta_k]_+ / (\lambda_k \varepsilon_k)} \Big] \Big] \Big\}. \end{array}$$

Here, the radii $\overline{\rho}_k$ and distributions $\mathbb{Q}_{x, \varepsilon_k}$ are defined in (4) and (8), respectively. It is worth noting that Problem (13) does not preserve convexity due to the bilinear structure of (s, β_k) for $k = 1, 2$ in two constraints. Fortunately, we can apply the bisection search method outlined in Algorithm 1 that finds the global optimum solution efficiently.

Algorithm 1 Bisection Search for Solving Problem (13)

Require: Interval $[s^{\text{lb}}, s^{\text{ub}}]$, precision level Υ.

1: **while** $s^{\text{ub}} - s^{\text{lb}} < \Upsilon$ **do**

2: $s \leftarrow \frac{1}{2}(s^{\text{lb}} + s^{\text{ub}})$.

3: Compute

$$T(s) = \min_{\substack{\theta\in\mathbb{R}^D, \|\theta\|_2\le 1, \beta_k\le 0, \\ \lambda_k\ge 0, k=1,2}} \left\{ \max_k \; G_k(s, \beta_k, \lambda_k) \right\}. \tag{14}$$

4: Update $s^{\text{ub}} \leftarrow s$ **if** $T(s) \le 0$ and **otherwise** $s^{\text{lb}} \leftarrow s$.

5: **end while**

Return s

The most computationally expansive step in Algorithm 1 is to solve the subproblem (14). One can apply the projected stochastic subgradient method [34] to obtain its optimal solution with a negligible optimality gap. The main difficulty is obtaining unbiased gradient estimates of G_k since the objective function involves nonlinear operators of expectations. Instead, one can follow the approach outlined in [35]–[37] to efficiently generate biased gradient estimates with controlled gradient bias and variance. Consequently, one can still obtain the optimal solution with convergence guarantees. We leave the complexity analysis of this method for future study.

It is also noteworthy that CVaR approximation has been used to solve the Sinkhorn robust chance-constrained program in literature [21]. Unlike their algorithm idea that solves a large-scale convex program using interior-point methods, we provide a first-order method that enables us to solve such problem more efficiently.

IV. Regularization Effects of Robust Testing

Recall we have used Sinkhorn ambiguity sets to robustify the probabilistic constraints in (1). In this section, we provide interpretations of such robustness by showing that the robust risk of a detector can be approximated by the non-robust risk with certain regularizations, called the *regularization effects*.

To begin with, we study the worst-case 0-1 loss function for a generic event E and a general nominal distribution $\widehat{\mathbb{P}}$:

$$\sup_{\mathbb{P}:\ \mathcal{W}_\varepsilon(\mathbb{P},\widehat{\mathbb{P}})\le\rho} \mathbb{P}(E). \tag{15}$$

We assume hyper-parameters $\overline{\rho}$ defined in (7) and regularization parameter ε both converges to 0, and we consider two scaling regimes between $\overline{\rho}$ and ε: either $\overline{\rho}/\varepsilon \to 0$ or $\varepsilon/\overline{\rho} \to 0$.

Case 1: $\overline{\rho}/\varepsilon \to 0$**.** In this case, the decaying rate of the radius $\overline{\rho}$ is faster than that of the regularization parameter ε. Define the variance regularizer $\sigma^2(E;\widehat{\mathbb{P}},\varepsilon) = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\Big[\mathbb{V}\mathrm{ar}_{z\sim\mathbb{Q}_{x,\varepsilon}}[\mathbf{1}_E(z)]\Big]$. The following proposition shows Problem (15) is asymptotically equivalent to variance regularized 0-1 loss, whose proof follows a similar argument from [38].

Proposition 4. *For any $b_0 > 0$, the following holds for all $\varepsilon > 0$ and Borel probability measures $\mathbb{P}$ satisfying $\inf_{\varepsilon>0}\sigma^2(E;\widehat{\mathbb{P}},\varepsilon) \ge b_0$:*

$$\sup_{\mathbb{P}:\ \mathcal{W}_\varepsilon(\mathbb{P},\widehat{\mathbb{P}})\le\rho} \mathbb{P}(E) - \Big(\mathbb{E}_{x\sim\widehat{\mathbb{P}}}[\mathbb{Q}_{x,\varepsilon}(E)] + (2\overline{\rho}/\varepsilon)^{1/2}\sigma(E;\widehat{\mathbb{P}},\varepsilon)\Big) = o((\overline{\rho}/\varepsilon)^{1/2}). \qquad \diamondsuit$$

Based on Proposition 4, the objective in Problem (1) can be viewed as the variance-regularized non-robust testing problem with residual error $O(\max_{k=1,2}\overline{\rho}_k/\varepsilon_k)$:

$$\max_{k=1,2}\left(\widehat{\mathbb{P}}_k(E_k) + (2\overline{\rho}_k/\varepsilon_k)^{1/2}\sigma(E_k;\widehat{\mathbb{P}}_k,\varepsilon_k)\right),$$

where $E_1 = \{\omega :\ T(\omega) < 0\}, E_2 = E_1^c$.

Case 2: $\varepsilon/\overline{\rho} \to 0$**.** Next, we consider the case where the convergence rate of the entropic regularization ε is faster than that of $\overline{\rho}$. To simplify the analysis, we consider the quadratic transport cost function $c(x,z) = \frac{1}{2}\|x-z\|_2^2$ for Sinkhorn discrepancy defined in Assumption 2. In such a case, we show Problem 15 is well approximated by the Wasserstein robust loss, whose proof is mainly based on Laplace's method [39].

Proposition 5. *For any measurable subset $E \subseteq \Omega$,*

$$\sup_{\mathbb{P}:\ \mathcal{W}_\varepsilon(\mathbb{P},\widehat{\mathbb{P}})\le\rho} \mathbb{P}(E) = \sup_{\mathbb{P}:\ \mathcal{W}_0(\mathbb{P},\widehat{\mathbb{P}})\le\overline{\rho}} \mathbb{P}(E) + O(\varepsilon/\overline{\rho}), \tag{16}$$

where $\mathcal{W}_0(\cdot,\cdot)$ denotes the standard optimal transport distance with quadratic transport cost function. $\diamondsuit$

We additionally assume that $\widehat{\mathbb{P}}$ is an empirical distribution constructed from n i.i.d. samples from the underlying true distribution $\mathbb{P}_*$, and specify the radius $\overline{\rho} = O(n^{-b})$ for some $b \in (0,1]$. Based on Proposition 5 and a recent study [40] that provides the regularization effect analysis on Wasserstein DRO with 0-1 loss, we further expand Problem (15) as

$$\sup_{\mathbb{P}:\ \mathcal{W}_\varepsilon(\mathbb{P},\widehat{\mathbb{P}})\le\rho} \mathbb{P}(E) = \widehat{\mathbb{P}}(E) + O(1)\cdot\mathfrak{g}(0)^{2/3}\overline{\rho}^{2/3} + O(\varepsilon/\overline{\rho}),$$

where the density $\mathfrak{g}(0) := \lim_{s\downarrow 0}\frac{1}{s}\mathbb{P}_*\{\omega : \mathsf{d}_{E^c}(\omega) \in (0,s)\}$.

Based on the argument above, the objective in Problem (1) can be viewed as the following density-regularized non-robust testing problem with residual error $O(\varepsilon/\overline{\rho})$:

$$\max_{k=1,2}\left(\widehat{\mathbb{P}}_k(E_k) + O(1)\cdot\mathfrak{g}_k(0)^{2/3}\overline{\rho}_k^{2/3}\right),$$

where events $E_k = \{\omega :\ (-1)^{k+1}\langle\theta,\Phi(\omega)\rangle < 0\}$, and density

$$\mathfrak{g}_k(0) = \lim_{s\downarrow 0}\frac{1}{s}\mathbb{P}_k\{\omega\in E_k : \mathsf{d}_{E_k^c}(\omega)\in(0,s)\}, \quad k=1,2.$$

A large value of $\mathfrak{g}_k(0)$ means the detector has a small empirical margin around the decision boundary. Hence, the robust testing in the regime $\varepsilon/\overline{\rho} \to 0$ tends to penalize this density to penalize detectors with a small margin.

V. Concluding Remarks

Our proposed framework opens avenues for further study. First, it is of research interest to develop more scalable optimization algorithms for solving the MIECP formulation (11) and consider enhanced convex approximations. Second, there is potential for relaxing the assumptions when showing the regularization effects for robust hypothesis testing. Finally, we are curious to explore guarantees for choosing the hyper-parameters of the robust testing model, including the radii and regularization parameters.

REFERENCES

[1] H. Poor and O. Hadjiliadis, *Quickest detection*. Cambridge University Press, Jan. 2008.

[2] L. Xie and Y. Xie, "Sequential change detection by optimal weighted ℓ_2 divergence," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 747–761, Apr. 2021.

[3] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 494–514, Apr. 2021.

[4] L. Xie, "Minimax robust quickest change detection using wasserstein ambiguity sets," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1909–1914.

[5] L. Xie, Y. Liang, and V. V. Veeravalli, "Distributionally robust quickest change detection using wasserstein uncertainty sets," *arXiv preprint arXiv:2309.16171*, 2023.

[6] J. R. Lloyd and Z. Ghahramani, "Statistical model criticism using kernel two sample tests," *Advances in neural information processing systems*, vol. 28, 2015.

[7] K. Chwialkowski, H. Strathmann, and A. Gretton, "A kernel test of goodness of fit," in *International conference on machine learning*. PMLR, 2016, pp. 2606–2615.

[8] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[9] P. Schober and T. R. Vetter, "Two-sample unpaired t tests in medical research," *Anesthesia & Analgesia*, vol. 129, no. 4, p. 911, 2019.

[10] P. J. Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753 – 1758, Dec. 1965.

[11] A. Magesh, Z. Sun, V. V. Veeravalli, and S. Zou, "Robust hypothesis testing with moment constrained uncertainty sets," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, Jan. 2009.

[13] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5572–5587, Apr. 2017.

[14] R. Gao, L. Xie, Y. Xie, and H. Xu, "Robust hypothesis testing using wasserstein uncertainty sets," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Dec. 2018, p. 7913–7923.

[15] L. Xie, R. Gao, and Y. Xie, "Robust hypothesis testing with wasserstein uncertainty sets," *arXiv preprint arXiv:2105.14348*, May 2021.

[16] J. Wang and Y. Xie, "A data-driven approach to robust hypothesis testing using sinkhorn uncertainty sets," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 3315–3320.

[17] Z. Sun and S. Zou, "Kernel robust hypothesis testing," *IEEE Transactions on Information Theory*, 2023.

[18] J. Wang, R. Gao, and Y. Xie, "Sinkhorn distributionally robust optimization," *arXiv preprint arXiv:2109.11926*, 2023.

[19] W. Azizian, F. Iutzeler, and J. Malick, "Regularization for wasserstein distributionally robust optimization," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 29, p. 33, 2023.

[25] W. Rudin, *Fourier analysis on groups*. Courier Dover Publications, 2017.

[20] J. Wang, R. Moore, Y. Xie, and R. Kamaleswaran, "Improving sepsis prediction model generalization with optimal transport," in *Machine Learning for Health*. PMLR, 2022, pp. 474–488.

[21] S.-B. Yang and Z. Li, "Distributionally robust chance-constrained optimization with sinkhorn ambiguity set," *AIChE Journal*, vol. 69, no. 10, p. e18177, 2023.

[22] C. Dapogny, F. Iutzeler, A. Meda, and B. Thibert, "Entropy-regularized wasserstein distributionally robust shape and topology optimization," *Structural and Multidisciplinary Optimization*, vol. 66, no. 3, p. 42, 2023.

[23] J. Song, N. He, L. Ding, and C. Zhao, "Provably convergent policy optimization via metric-aware trust region methods," *arXiv preprint arXiv:2306.14133*, 2023.

[24] J. Wang, R. Gao, and Y. Xie, "Non-convex robust hypothesis testing using sinkhorn uncertainty sets," *arXiv preprint arXiv:2403.14822*, 2024.

[26] C. Ma, S. Wojtowytsch, L. Wu *et al.*, "Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't," *arXiv preprint arXiv:2009.10713*, 2020.

[27] J. Dahl and E. D. Andersen, "A primal-dual interior-point algorithm for nonsymmetric exponential-cone optimization," *Mathematical Programming*, vol. 194, no. 1-2, pp. 341–370, 2022.

[28] C. Coey, M. Lubin, and J. P. Vielma, "Outer approximation with conic certificates for mixed-integer convex problems," *Mathematical Programming Computation*, vol. 12, no. 2, pp. 249–293, 2020.

[29] Q. Ye and W. Xie, "Second-order conic and polyhedral approximations of the exponential cone: application to mixed-integer exponential conic programs," *arXiv preprint arXiv:2106.09123*, 2021.

[30] M. ApS, "Mosek optimization suite," 2019.

[31] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, 2007.

[32] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[33] M. Sion, "On general minimax theorems." *Pacific J. Math.*, vol. 8, no. 4, pp. 171–176, 1958.

[34] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[35] Y. Hu, X. Chen, and N. He, "On the bias-variance-cost tradeoff of stochastic optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 119–22 131, 2021.

[36] Y. Hu, S. Zhang, X. Chen, and N. He, "Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2759–2770, 2020.

[37] Y. Hu, J. Wang, Y. Xie, A. Krause, and D. Kuhn, "Contextual stochastic bilevel optimization," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[38] J. Blanchet and A. Shapiro, "Statistical limit theorems in distributionally robust optimization," *arXiv preprint arXiv:2303.14867*, 2023.

[39] A. Erdélyi, *Asymptotic expansions*. Courier Corporation, 1956, no. 3.

[40] Z. Yang and R. Gao, "Wasserstein regularization for 0-1 loss," *Optimization Online Preprint*, 2022.