Parametric Scenario Optimization under Limited Data: A Distributionally Robust Optimization View

HENRY LAM*, Department of Industrial Engineering and Operations Research, Columbia University
FENGPEI LI*, Department of Industrial Engineering and Operations Research, Columbia University

We consider optimization problems with uncertain constraints that need to be satisfied probabilistically. When data are available, a common method to obtain feasible solutions for such problems is to impose sampled constraints, following the so-called scenario optimization approach. However, when the data size is small, the sampled constraints may not support a guarantee on the feasibility of the obtained solution. This paper studies how to leverage parametric information and the power of Monte Carlo simulation to obtain feasible solutions for small-data situations. Our approach makes use of a distributionally robust optimization (DRO) formulation that translates the data size requirement into a Monte Carlo sample size requirement drawn from what we call a generating distribution. We show that, while the optimal choice of this generating distribution is the one eliciting the data or the baseline distribution in a nonparametric divergence-based DRO, it is not necessarily so in the parametric case. Correspondingly, we develop procedures to obtain generating distributions that improve upon these basic choices. We support our findings with several numerical examples.

CCS Concepts: • **Mathematics of computing → Probability and statistics**; • **Computing methodologies → Modeling and simulation**; • **Simulation types and techniques** → *Uncertainty quantification*; • **Mathematical analysis** → *Mathematical optimization.*

Additional Key Words and Phrases: chance constraint, distributionally robust optimization, scenario optimization, parametric uncertainty

ACM Reference Format:
Henry Lam and Fengpei Li. 2021. Parametric Scenario Optimization under Limited Data: A Distributionally Robust Optimization View. 1, 1 (May 2021), 41 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

We consider optimization problems in the form

$$\begin{aligned} \min_{x\in\mathcal{X}\subseteq\mathbb{R}^d} \quad & c^T x, \\ \text{s.t.} \quad & \mathbb{P}(x\in\mathcal{X}_\xi)\geq 1-\epsilon, \end{aligned} \tag{1.1}$$

where $\mathbb{P}$ is a probability measure governing the random variable ξ on some space $\mathcal{Y}$ and $\mathcal{X}_\xi \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ is a set depending on ξ. Problem (1.1) enforces a solution x to satisfy $x \in \mathcal{X}_\xi$ with high probability, namely at least $1-\epsilon$. This problem is often known as a probabilistically constrained or chance-constrained program (CCP) [61]. It provides a natural framework for decision-making under stochastic resource capacity or risk tolerance, and has been applied in various domains such as production planning [55], inventory management [50], reservoir design [62, 63], communications [65], and ranking and selection [35].

We focus on the situations where $\mathbb{P}$ is unknown, but some i.i.d. data, say $\xi_1, \ldots, \xi_n$, are available. One common approach to handle (1.1) in these situations is to use the so-called scenario optimization

Authors' addresses: Henry Lam, khl2114@columbia.edu, Department of Industrial Engineering and Operations Research, Columbia University, 500 W. 120th Street, New York, NY, 10027; Fengpei Li, fl2412@columbia.edu, Department of Industrial Engineering and Operations Research, Columbia University, 500 W. 120th Street, New York, NY, 10027.

XXXX-XXXX/2021/5-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(SO) or constraint sampling [11, 56]. This replaces the unknown constraint in (1.1) with $x \in \mathcal{X}_{\xi_i}, i = 1, \ldots, n$, namely, by considering

$$\begin{array}{ll} \min\limits_{x \in \mathcal{X} \subseteq \mathbb{R}^d} & c^T x, \\ \text{s.t.} & x \in \mathcal{X}_{\xi_i},\ i = 1, \ldots, n. \end{array} \tag{1.2}$$

Note that CCP (1.1) is generally difficult to solve even when the set $\mathcal{X}_\xi$ is convex for any given ξ and the distribution $\mathbb{P}$ is known [61]. Thus, the sampled problem (1.2) offers a tractable approximation for the difficult CCP even in non-data-driven situations, assuming the capability to generate these samples.

Our goal is to find a good feasible solution for (1.1) by solving (1.2) under the availability of i.i.d. data described above. Intuitively, as the sample size n increases, the number of constraints in (1.2) increases and one expects them to sufficiently populate the safety set $\{\xi : x \in \mathcal{X}_\xi\}$, thus ultimately give rise to a feasible solution for (1.1). To make this more precise, we first mention that because of the statistical noise from the data, one must settle for finding a solution that is feasible with a high confidence. More specifically, define, for any given solution x,

$$V(x, \mathbb{P}) = \mathbb{P}(x \notin \mathcal{X}_\xi)$$

to be the violation probability of x under probability measure $\mathbb{P}$ that generates ξ. Obviously, x is feasible for (1.1) if and only if

$$V(x, \mathbb{P}) \leq \epsilon. \tag{1.3}$$

We would like to obtain a solution, say $\hat{x}$, from the data such that

$$\mathbb{P}_{data}(V(\hat{x}, \mathbb{P}) \leq \epsilon) \geq 1 - \alpha, \tag{1.4}$$

where $\mathbb{P}_{data}$ is the distribution that generates the i.i.d. data $\xi_i, i = 1, \ldots, n$ (each sampled from $\mathbb{P}$), and $1 - \alpha$ is a given confidence level (e.g., $\alpha = 5\%$). In other words, we want $\hat{x}$ to satisfy the chance constraint in (1.1) with the prescribed confidence.

Under the convexity of $\mathcal{X}_\xi$ and mild additional assumptions (namely, that every instance of (1.2) has a feasible region with nonempty interior and a unique optimal solution), the seminal work [10] provides a tight estimate on the required data size n to guarantee (1.4). They show that a solution $\hat{x}$ obtained by solving (1.2) satisfies

$$\mathbb{P}_{data}(V(\hat{x}, \mathbb{P}) > \epsilon) \leq \sum_{i=0}^{d-1} \binom{n}{i} \epsilon^i (1 - \epsilon)^{n-i}, \tag{1.5}$$

with equality held for the class of "fully-supported" optimization problems [10]. Thus, suppose we have a sample size n large enough such that

$$B(\epsilon, d, n) = \sum_{i=0}^{d-1} \binom{n}{i} \epsilon^i (1 - \epsilon)^{n-i} \leq \alpha, \tag{1.6}$$

then from (1.5) we have $\mathbb{P}_{data}(V(\hat{x}, \mathbb{P}) > \epsilon) \leq \alpha$ or (1.4).

However, in small-sample situations in which the data size n is not large enough to support (1.6), the feasibility guarantee described above may not hold. It can be shown [10] that the minimum n that achieves (1.4) is linear in d and reciprocal in ϵ, thus may impose challenges especially in high-dimensional and low-tolerance problems. Similar dependence on the key problem parameters also appears in other related methods such as [20], which uses the Vapnik-Chervonenkis dimension to infer required sample sizes, the sampling-and-discarding approach in [11], and the closely related approach using sample average approximation in [53]. Several recent lines of techniques have been suggested to overcome these challenges and reduce sample size requirements, including the

use of support rank and solution-dependent support constraints [12, 64], regularization [9], and sequential approaches [7, 8, 13, 14].

In this paper, we offer a different path to alleviate the data size requirement than the above methods, when $\mathbb{P}$ possesses known parametric structures. Namely, we assume $\mathbb{P} \in \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ for some parametric family of distribution, where $\mathbb{P}_\theta$ satisfies two basic requirements: It is estimatable, i.e., the unknown quantity or parameter θ can be estimated from data, and simulatable, i.e., given θ, samples from $\mathbb{P}_\theta$ can be drawn using Monte Carlo methods. Under these presumptions, our approach turns the CCP (1.1), with an unknown parameter, into a CCP that has a definite parameter and a suitably re-adjusted tolerance level, which then allows us to generate enough Monte Carlo samples and consequently utilize the guarantee provided from (1.5). On a high level, this approach replaces the data size requirement in using (1.2) (or, in fact, any of its variant methods) with a Monte Carlo size requirement, the latter potentially more available given cheap modern computational power. Our methodological contributions consist of the development of procedures, related statistical results on their sample size requirement translations, and also showing some key differences between parametric and nonparametric regimes.

Our approach starts with a distributionally robust optimization (DRO) to incorporate the data-driven parametric uncertainty. The latter is a framework for decision-making under modeling uncertainty on the underlying probability distributions in stochastic problems. It advocates the search for decisions over the worst case, among all distributions contained in a so-called uncertainty set or ambiguity set (e.g., [21, 30, 67]). In CCP, this entails a worst-case chance constraint over this set (e.g., [15–17, 32, 33, 37, 39, 40, 51, 69–71]). When the uncertainty set covers the true distribution with a high confidence (i.e., the set is a confidence region), then feasibility for the distributionally robust CCP converts into a confidence guarantee on the feasibility for the original CCP. We follow this viewpoint and utilize uncertainty sets in the form of a neighborhood ball surrounding a baseline distribution, where the ball size is measured by a statistical distance (e.g., [1, 5, 6, 22, 23, 25, 27, 31, 34, 37, 44, 46, 52, 60]). In the parametric case, a suitable choice of this distance (such as the ϕ-divergence that we focus on) allows easy and meaningful calibration of the ball size from the data, so that the resulting DRO provides a provable feasibility conversion to the CCP.

Our next step is to combine this DRO with Monte Carlo sampling and scenario approximation. The definition of DRO means that there are many possible candidate distributions that can govern the truth, whereas the statistical guarantee for SO assumes a specific distribution that generates the data or Monte Carlo samples. To resolve this discrepancy, we select a *generating distribution* that draws the Monte Carlo samples, and develop a translation of the guarantee from a fixed distribution into one on the DRO. We highlight the benefits in using SO to handle this DRO, as opposed to other potential methods. While there exist many good results on tractable reformulations of DRO for chance constraints (e.g., [32, 33, 39, 51, 69]), the reformulation tightness typically relies on using moment-based uncertainty sets and particular forms of the safety condition. Compared to moments, divergence-based uncertainty sets can be calibrated with data to consistently shrink to the true distribution. Importantly, in the parametric case, the calibration of divergence-based sets is especially convenient, and achieves a tight convergence rate by using maximum likelihood theory that efficiently captures parametric information. Our condition for applying SO to this DRO is at the same level of generality as applying SO to an unambiguous CCP, which, as mentioned before, only requires the convexity of $\mathcal{X}_\xi$ and mild conditions.

To exploit the full capability of our approach, we investigate the optimal choice of the generating distribution in relation to the target DRO, in the sense of requiring the least Monte Carlo size. We show that, if there is no ambiguity on the distribution (i.e., a standard CCP), or when the uncertainty set of a DRO is constructed via a divergence ball in the nonparametric space, the best

generating distribution is, in a certain sense, the true or the baseline distribution at the center of the ball. However, if there is parametric information, the optimal choice of the generating distribution can deviate from the baseline distribution in a divergence-based DRO. We derive these results by casting the problem of selecting a generating distribution into a hypothesis testing problem, which connects the sampling efficiency of the generating distribution with the power of the test and the Neyman-Pearson lemma [48]. The results on DRO in particular combine this Neyman-Pearson machinery with the established DRO reformulation of chance constraints in [37, 40], with the discrepancy between the best generating distribution and the baseline distribution in the parametric case stemming from the removal of the extremal distributions in the corresponding nonparametric uncertainty set. These connections among hypothesis testing, SO and DRO are, to our best knowledge, the first of its kind in the literature.

Finally, given the non-optimality of the baseline distribution of a divergence-based DRO in generating Monte Carlo samples, we further develop procedures to search over generating distributions that improve upon this baseline. On a high level, this can be achieved by increasing the sampling variability to incorporate the uncertainty of the distributional parameters (one may intuit this from the perspective of a posterior distribution in a Bayesian framework), which is implemented by utilizing suitable mixture distributions. We provide several classes of mixture distributions to attain such a variability enlargement, and study descent-type algorithms to search for good distributions in these classes. In the experiments, we show our methods can be combined with SO or other SO-based methods including FAST [13] to solve a variety of optimization problems and data distributions, some of which are not amenable to RO, especially when the objective function is non-linear or the feasible sets are jointly chance-constrained. Furthermore, we also demonstrate how to search for more judicious choices of generating distributions that can significantly reduce the required number of Monte Carlo samples.

We conclude this introduction by briefly discussing a few other lines of related literature. The first is the so-called robust Monte Carlo or robust simulation that, like us, also considers using Monte Carlo sampling together with DRO [28, 29, 36, 38, 42, 43]. However, this literature focuses on approximating DRO with stochasticity in the objective function, and does not study the chance constraint feasibility and SO that constitute our main focus. We also contrast our work with [44] that also considers likelihood theory and utilizes simulation in tackling uncertain constraints. [44] focuses on the nonparametric regime and uses the empirical likelihood to construct uncertainty sets. Unlike our work, there is no parametric information there that can be leveraged to overcome sample size requirements in SO. Moreover, the simulation used in [44] is for calibrating the uncertainty set, instead of drawing sample constraints. Next, [24] considers a scenario approach to distributionally robust CCP with an uncertainty set based on the Prohorov distance. Like [20], [24] utilizes the Vapnik-Chervonenkis dimension in studying feasibility, in contrast to the convexity-based argument in [10] that we utilize. More importantly, we aim to optimize the efficiency of Monte Carlo sampling in handling limited-data CCP, thus motivating us to study the choice of distance, calibration schemes, and selection of generating distributions that are different from [24]. Finally, a preliminary conference version of this work has appeared in [45], which contains a basic introduction of our framework, without detailed investigation of the optimality of generating distributions, improvement strategies, and extensive numerical demonstrations.

To summarize, our main contributions of this paper are:

(1) We propose a framework to obtain good feasible solutions in data-driven CCPs in small-sample situations, where the data size is insufficient to support the use of SO with valid statistical guarantees. Focusing on the parametric regime, our framework operates by setting up a DRO, with an uncertainty set constructed from parameter estimates using the data, that

can in turn be tackled by using SO with Monte Carlo samples. In doing so, our framework effectively leverages the parametric information to convert the SO requirement on the data size into a requirement on the Monte Carlo size, the latter can be much more abundant given cheap modern computational power. The overview of this framework and the DRO construction are in Sections 2.1 and 2.3.

(2) We investigate and present the Monte Carlo size requirements needed to give statistically feasible solutions to the divergence-based DRO used in our framework. This relies on developing an implementable mechanism to connect the sample size requirement for SO, which attempts to solve a CCP with a fixed underlying distribution, to the sample size requirement needed to solve a DRO, by selecting a suitable generating distribution to draw the Monte Carlo samples. This contribution is presented in Section 2.2.

(3) We study the optimality of generating distributions, in a sense of minimizing the Monte Carlo effort that we will describe precisely. In particular, we show that the optimal generating distributions for an unambiguous CCP, and for a distributionally robust CCP with nonparametric divergence-based uncertainty sets, are simply their respective natural choices, namely the original underlying distribution and the baseline distribution (i.e., center of the divergence ball). In contrast, the optimal generating distribution for a distributionally robust CCP in the parametric case is more delicate, and the baseline distribution there can be readily dominated by other generating distributions. These results are derived by bridging the Neyman-Pearson lemma in statistical hypothesis testing with SO and DRO, which appears to be the first of its kind in the literature as far as we know. This contribution is presented in Section 3.

(4) Motivated by the non-optimality of the baseline distribution, we propose several approaches to construct generating distributions that dominate the baseline distributions for parametric DRO, by using mixture schemes that, on a high level, enlarge the variability of the generating distributions. We show how to use descent-type search procedures to construct these distributions. This contribution is presented in Section 4.

Lastly, we also present in full detail our implementation algorithms in Section 5, numerically demonstrate our approach and compare with other methods in Section 6, and conclude in Section 7.

2 FROM DATA-DRIVEN DRO TO SCENARIO OPTIMIZATION

This section introduces our overall framework. Recall our goal as to find a good feasible solution $\hat{x}$ for (1.1), and suppose that we have an i.i.d. data size n possibly less than the requirement shown in (1.6). As discussed in the introduction, we first formulate a DRO that incorporates the parametric estimation noise and subsequently allows us to resort to Monte Carlo sampling to obtain a feasible solution for (1.1). In the following, Section 2.1 first describes the basic guarantees from DRO. Section 2.2 investigates Monte Carlo sampling that provides guarantees on DRO. Section 2.3 discusses the choice of the uncertainty set.

2.1 Overview of Data-Driven DRO

For concreteness, suppose the unknown true distribution $\mathbb{P} \in \mathcal{P}$, the class of possible probability distributions for ξ (to be specified later). Given the observed data $\xi_1, ..., \xi_n$, the basic steps in our data-driven DRO are:

- Step 1: Find a data-driven uncertainty set $\mathcal{U}_{data} = \mathcal{U}_{data}(\xi_1, \ldots, \xi_n) \subseteq \mathcal{P}$ such that

$$\mathbb{P}_{data}(\mathbb{P} \in \mathcal{U}_{data}) \geq 1 - \alpha, \tag{2.1}$$

  where $\mathbb{P}_{data}$ denotes the measure generating the data $\xi_i, i = 1, \ldots, n$.

- Step 2: Given $\mathcal{U}_{data}$, set up the distributionally robust CCP:

$$\begin{array}{ll} \min\limits_{x\in\mathcal{X}\subseteq\mathbb{R}^d} & c^T x, \\ \text{s.t.} & \min\limits_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(x \in \mathcal{X}_\xi) \geq 1-\epsilon, \end{array} \tag{2.2}$$

  where the probability measure $\mathbb{Q}$ is the decision variable in the minimization in the constraint.
- Step 3: Find a solution $\hat{x}$ feasible for (2.2).

It is straightforward to see that $\hat{x}$ obtained from the above procedure is feasible for (1.1) with confidence at least $1-\alpha$: If $\mathbb{P} \in \mathcal{U}_{data}$, then any $\hat{x}$ feasible for (2.2) satisfies

$$\mathbb{P}(\hat{x} \in \mathcal{X}_\xi) \geq \min_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(\hat{x} \in \mathcal{X}_\xi) \geq 1-\epsilon$$

Thus

$$\mathbb{P}_{data}(\mathbb{P}(\hat{x} \in \mathcal{X}_\xi) \geq 1-\epsilon) \geq \mathbb{P}_{data}(\mathbb{P} \in \mathcal{U}_{data}) \geq 1-\alpha, \tag{2.3}$$

which gives our conclusion.

2.2 Monte Carlo Sampling for DRO

To use the above procedure, we need to provide a way to construct the depicted $\mathcal{U}_{data}$ and to find a (confidently) feasible solution for (2.2). We postpone the set construction to the next subsection and focus on finding a feasible solution here. We resort to SO, via Monte Carlo sampling, to handle (2.2). Note that, unlike in the standard SO discussed in the introduction, the distribution $\mathbb{Q}$ here can be any candidate within the set $\mathcal{U}_{data}$. Thus, let us select a generating distribution, called $\mathbb{P}_0$ (which can depend on the data), to generate Monte Carlo samples ξ_i^{MC}, $i = 1, \ldots, N$, and solve

$$\begin{array}{ll} \min\limits_{x\in\mathcal{X}\subseteq\mathbb{R}^d} & c^T x, \\ \text{s.t.} & x \in \mathcal{X}_{\xi_i^{MC}},\ i = 1, \ldots, N. \end{array} \tag{2.4}$$

For convenience, denote, for any $\epsilon, \beta > 0$,

$$N_{exact}(\epsilon, \beta, d) = \min\left\{ n : \sum_{i=0}^{d-1} \binom{n}{i} \epsilon^i (1-\epsilon)^{n-i} \leq \beta \right\}. \tag{2.5}$$

From the result of [10] discussed in the introduction, using $N_{exact}(\epsilon, \beta, d)$ or more Monte Carlo samples from $\mathbb{P}_0$ in (2.4) would give a solution $\hat{x}^{MC}$ that satisfies $V(\hat{x}^{MC}, \mathbb{P}_0) \leq \epsilon$ with confidence level $1-\beta$. This is not exactly the distributionally robust feasibility statement for problem (2.2). To address this discrepancy, we consider, conditional on the data $\xi_1, \ldots, \xi_n$,

$$\begin{array}{ll} \max\limits_{\mathbb{Q}\in\mathcal{U}_{data}} & V(\hat{x}^{MC}, \mathbb{Q}) \\ \text{s.t.} & V(\hat{x}^{MC}, \mathbb{P}_0) \leq \delta. \end{array} \tag{2.6}$$

This optimization problem serves to translate a guarantee on the violation probability under $\mathbb{P}_0$ to any $\mathbb{Q}$ in $\mathcal{U}_{data}$. If we can bound the optimal value in (2.6), then we can trace back the level of δ that is required to ensure a chance constraint validity of tolerance level ϵ. However, the event involved in defining $V(\hat{x}^{MC}, \mathbb{P}_0)$ and $V(\hat{x}^{MC}, \mathbb{Q})$, namely $\{\xi : \hat{x}^{MC} \notin \mathcal{X}_\xi\}$, can be challenging to handle in general. Thus, we relax (2.6) to

$$\begin{array}{ll} \max\limits_{\mathbb{Q}\in\mathcal{U}_{data}, A\subset\mathcal{Y}} & \mathbb{Q}(A) \\ \text{s.t.} & \mathbb{P}_0(A) \leq \delta. \end{array} \tag{2.7}$$

where the decision variables now include the set A in addition to the probability measure $\mathbb{Q}$. Conditional on the data $\xi_1, \ldots, \xi_n$, the optimal value of optimization problem (2.7), which we denote

$M(\mathbb{P}_0, \boldsymbol{\mathcal{U}}_{data}, \delta)$, is clearly an upper bound for that of (2.6). In fact, it is also clear from (2.7) that $M(\mathbb{P}_0, \boldsymbol{\mathcal{U}}_{data}, \delta)$ is non-decreasing in $\delta > 0$ and

$$\max_{\mathbb{Q}\in\boldsymbol{\mathcal{U}}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) \leq M(\mathbb{P}_0, \boldsymbol{\mathcal{U}}_{data}, V(\hat{x}^{MC}, \mathbb{P}_0)), \tag{2.8}$$

by simply taking $A = \{\xi : \hat{x}^{MC} \notin \mathcal{X}_{\xi}\}$ and $\delta = V(\hat{x}^{MC}, \mathbb{P}_0)$ in (2.7). We have the following guarantee:

Theorem 2.2.1. *Given $\mathbb{P}_0$, $\boldsymbol{\mathcal{U}}_{data}$ and $\epsilon > 0$, suppose there exists $\delta_\epsilon > 0$ small enough such that*

$$M(\mathbb{P}_0, \boldsymbol{\mathcal{U}}_{data}, \delta_\epsilon) \leq \epsilon. \tag{2.9}$$

If we solve (2.4) *with $N_{exact}(\delta_\epsilon, \beta, d)$ number of samples drawn from $\mathbb{P}_0$, then the obtained solution $\hat{x}^{MC}$ would be feasible for* (2.2) *with confidence at least $1-\beta$. Furthermore, if*

$$\mathbb{P}_{data}(\mathbb{P} \in \boldsymbol{\mathcal{U}}_{data}) \geq 1-\alpha, \tag{2.10}$$

where $\mathbb{P}_{data}$ is the measure governing the real-data generation under the true distribution $\mathbb{P}$, then the obtained solution $\hat{x}^{MC}$ would be feasible for (1.1) *with confidence at least $1-\alpha-\beta$.*

Proof. By results in [10], we know that by solving (2.4) with $N_{exact}(\delta_\epsilon, \beta, d)$ number of samples from $\mathbb{P}_0$, the obtained solution $\hat{x}^{MC}$ would satisfy

$$\mathbb{P}_{MC,0}(V(\hat{x}^{MC}, \mathbb{P}_0) > \delta_\epsilon) \leq \beta \tag{2.11}$$

where $\mathbb{P}_{MC,0}$ is the measure with respect to the Monte Carlo samples drawn from $\mathbb{P}_0$. Moreover, based on the monotonicity property of $M(\cdot)$ and (2.8), we have

$$V(\hat{x}^{MC}, \mathbb{P}_0) \leq \delta_\epsilon \implies \max_{\mathbb{Q}\in\boldsymbol{\mathcal{U}}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) \leq M(\mathbb{P}_0, \boldsymbol{\mathcal{U}}_{data}, \delta_\epsilon). \tag{2.12}$$

Thus (2.9) implies that

$$\mathbb{P}_{data}\left(\max_{\mathbb{Q}\in\boldsymbol{\mathcal{U}}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) > \epsilon\right) \leq \mathbb{P}_{data}(V(\hat{x}^{MC}, \mathbb{P}_0) > \delta_\epsilon) \leq \beta$$

and hence $\hat{x}^{MC}$ is feasible for (2.2) with confidence at least $1-\beta$. Furthermore, if $\mathbb{P} \in \boldsymbol{\mathcal{U}}_{data}$, then a $\hat{x}^{MC}$ feasible for (2.2) is also feasible for (1.1) since $\max_{\mathbb{Q}\in\boldsymbol{\mathcal{U}}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) \geq V(\hat{x}^{MC}, \mathbb{P})$ and hence

$$\max_{\mathbb{Q}\in\boldsymbol{\mathcal{U}}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) \leq \epsilon \implies V(\hat{x}^{MC}, \mathbb{P}) \leq \epsilon. \tag{2.13}$$

Thus, if we denote $\Xi = \{\xi_1, ..., \xi_n, \xi_1^{MC}, ..., \xi_N^{MC}\}$ to be entire sequence consisting of real data and the generated Monte Carlo samples, it then follows that

$$\{\Xi : V(\hat{x}^{MC}, \mathbb{P}) > \epsilon\} \subseteq \{\Xi : \mathbb{P} \notin \boldsymbol{\mathcal{U}}_{data}\} \cup \{\Xi : V(\hat{x}^{MC}, \mathbb{P}_0) > \delta_\epsilon\}. \tag{2.14}$$

It now follows by (2.10) and (2.11) that $\hat{x}^{MC}$ is feasible for (1.1) with probability at least $1-\alpha-\beta$. □

Theorem 2.2.1 can be cast in terms of asymptotic instead of finite-sample guarantees by following the same line of arguments. We summarize it as the following corollary.

Corollary 2.2.2. *In Theorem 2.2.1, if the condition $\mathbb{P}_{data}(\mathbb{P} \in \boldsymbol{\mathcal{U}}_{data}) \geq 1-\alpha$ is substituted by the asymptotic condition*

$$\liminf_{n\to\infty} \mathbb{P}_{data}(\mathbb{P} \in \boldsymbol{\mathcal{U}}_{data}) \geq 1-\alpha, \tag{2.15}$$

then the feasibility of $\hat{x}^{MC}$ in the last conclusion of Theorem 2.2.1 holds with confidence asymptotically tending to at least $1-\alpha-\beta$.

To summarize, in the presence of data insufficiency, if we choose $\mathcal{U}_{data}$ to satisfy the confidence property (2.1), and are able to evaluate the bounding function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ that translates the violation probability under $\mathbb{P}_0$ to a worst-case violation probability over $\mathcal{U}_{data}$, then we can run SO with $N_{exact}(\delta_\epsilon, \beta, d)$ Monte Carlo samples from $\mathbb{P}_0$ to obtain a solution for (1.1) with confidence $1 - \alpha - \beta$.

We also note that the above scheme still holds if the $N_{exact}(\epsilon, \beta, d)$ in (2.5) is replaced by the sample size requirements of other variants of SO (e.g., FAST [13]) that are potentially smaller. This works as long as we stay with the same SO-based procedure in using the Monte Carlo samples. For clarity, throughout most of our exposition we will focus on the sample size requirement depicted in (2.5), but we will discuss other variants in our implementation and numerical sections.

Finally, let us take a step back and justify why we use SO to tackle (2.2), as opposed to other potential means. Indeed, as pointed out in the introduction, there exist many good results on tractable reformulations of DRO. As will be discussed in detail in the next subsection, in the present context we will choose an uncertainty set that can leverage parametric information efficiently. Sets based on the neighborhoods of distributions measured by ϕ-divergences are particularly attractive choices, as they can be calibrated easily (both the ball center and the size) in a way that efficiently uses parametric information. The dependence on the parameter dimension in particular is reflected in the degree of freedom in the χ^2-distribution used in the calibrating the ball size, which shrinks to zero at a canonical rate as the data size increases. Other sets, such as moment-based ones, though possibly amenable to tight tractable reformulations, do not enjoy these statistical properties in the parametric context. Thus, in view of tackling ϕ-divergence-based DRO, SO appears to be a natural choice, and we have set up a framework to utilize it under conditions at the same level of generality as required for the unambiguous counterpart. Sections 3 and 4 will study this framework in further depth and enhance its efficiency. We caution, however, that the conservativeness in our proposed uncertainty set (which affects the optimality of the obtained solution) relies on the dimensionality of the distributional parameters. Our approach is expected to work well when this dimension is moderate, but not in high-dimensional problems where other approaches could be better choices.

2.3 Constructing Uncertainty Sets

In this section we discuss the construction of the uncertainty set $\mathcal{U}_{data}$, using the ϕ-divergence approach [1]. We assume the true distribution $\mathbb{P}$ of ξ lies in a parametric family. We denote the true parameter as θ_{true}. To highlight the parametric dependence, we call the true distribution $\mathbb{P}_{\theta_{true}} \in \mathcal{P}_{para} = \{\mathbb{P}_\theta\}_{\theta \in \Theta \subset \mathbb{R}^D}$ indexed by θ, where D is the dimension of parameter space. Given data $\xi_1, \xi_2, ..., \xi_n$, we want to construct an uncertainty set $\mathcal{U}_{data}$ satisfying

$$\lim_{n\to\infty} \mathbb{P}_{data}(\mathbb{P}_{\theta_{true}} \in \mathcal{U}_{data}) = 1 - \alpha \tag{2.16}$$

so that Corollary 2.2.2 applies. To do so, we first estimate θ_{true} from the data. There are various approaches to do so; here we apply the common maximum likelihood estimator (MLE) $\hat{\theta}_n$, and set $\mathcal{U}_{data}$ to be

$$\mathcal{U}_{data} = \left\{ \mathbb{Q} \in \mathcal{P}_{para} : d_\phi(\mathbb{P}_{\hat{\theta}_n}, \mathbb{Q}) \le \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n} \right\}, \tag{2.17}$$

where $\chi^2_{1-\alpha,D}$ is the $1-\alpha$ quantile of χ^2_D, the χ^2-distribution with degree of freedom D, and $d_\phi(\cdot,\cdot)$ is the ϕ-divergence between two probability measures, i.e., given a convex function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, with $\phi(1) = 0$, a distance between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ defined as

$$d_\phi(\mathbb{P}_1, \mathbb{P}_2) = \int_{\mathcal{Y}} \phi\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}\right) \mathbb{P}_1(dy), \tag{2.18}$$

assuming $\mathbb{P}_2$ is absolutely continuous with respect to $\mathbb{P}_1$ with Radon-Nikodym derivative $\frac{d\mathbb{P}_2}{d\mathbb{P}_1}$ on $\mathcal{Y}$. Moreover, we assume that ϕ is twice continuously differentiable with $\phi''(1) \neq 0$, and if necessary set the continuation of ϕ to $\mathbb{R}_-$ as $\phi(x) = +\infty$ for $x < 0$. In (2.17), we call the center of the divergence ball, $\mathbb{P}_{\hat{\theta}_n}$, the baseline distribution.

To guarantee desirable asymptotic properties of our uncertainty set, we make the following assumption:

A1. Let $\theta_{true} \in \Theta$ be the true parameter and let $\hat{\theta}_n$ be the MLE of θ_{true} estimated from n i.i.d. data points. Then, as $n \to \infty$, $\hat{\theta}_n$ satisfies consistency and asymptotic normality condition:

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_{true} \quad \text{and} \quad \sqrt{n}(\hat{\theta}_n - \theta_{true}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}^{-1}(\theta_{true})), \tag{2.19}$$

where $\mathcal{I}(\theta)$ is the Fisher information for the parametric family $\mathcal{P}_{para}$ with well-defined inverse that is continuous in the domain $\theta \in \Theta$.

Assumption A1 of MLE estimator is known to hold under various regularity conditions [47, 66]. We list a set of such conditions in Appendix A.

Under Assumption A1, it can be shown [59, 66] that $\mathcal{U}_{data}$ in (2.17) satisfies the confidence guarantee (2.16). Furthermore, since we can identify each $\mathbb{P}_\theta$ in $\mathcal{P}_{data}$ with θ, we can equivalently view $\mathcal{U}_{data}$ as a subset of $\theta \in \Theta$, and write it as

$$\mathcal{U}_{data} \triangleq \left\{ \theta \in \Theta : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_\theta) \leq \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n} \right\}. \tag{2.20}$$

For convenience, we shall use the two definitions of $\mathcal{U}_{data}$ interchangeably depending on the context. It is also known that the asymptotic confidence properties of (2.17) or (2.20) are the same among different choices within the ϕ-divergence class. These can be seen via a second order expansion of the ϕ-divergences. Moreover, they are asymptotically equivalent to

$$\left\{ \theta \in \Theta : (\theta - \hat{\theta}_n)^T \mathcal{I}(\hat{\theta}_n)(\theta - \hat{\theta}_n) \leq \frac{\chi^2_{1-\alpha,D}}{n} \right\}, \tag{2.21}$$

where $\mathcal{I}(\hat{\theta}_n)$ is the estimated Fisher information, under the regularity conditions above [58, 59, 66]. In other words, under Assumption A1, both (2.20) and (2.21) satisfy

$$\lim_{n\to\infty} \mathbb{P}_{data}(\theta_{true} \in \mathcal{U}_{data}) = 1 - \alpha. \tag{2.22}$$

Note that the convergence rate of (2.16) or (2.22) depends on the higher-order properties of the parametric model, which in turn can depend on the parameter dimension. Different from the sample size requirements in SO, this convergence rate is a consequence of MLE properties. Some details on finite-sample behaviors of MLE can be found in [41].

The $\mathcal{U}_{data}$ discussed above is a set over the parametric class of distributions (or parameter values). Considering tractability, DRO over nonparametric space could be easier to handle than parametric, which suggests a relaxation of the parametric constraint to estimate the bounding function M. This also raises the question of whether one can possibly contain $\mathcal{U}_{data}$ in a nonparametric ball with a shrunk radius and subsequently obtain a better M. These would be the main topics of Sections 3 and 4.

3 BOUNDING FUNCTIONS AND GENERATING DISTRIBUTIONS

Given the uncertainty set $\mathcal{U}_{data}$ in (2.20), we turn to the choice of the generating measure $\mathbb{P}_0$ and the bounding function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ which, as we recall, is the optimal value of optimization

problem (2.7). In the discussed parametric setup, the latter becomes

$$\begin{aligned} \max_{\theta\in\mathcal{U}_{data}, A\subset\mathcal{Y}} \quad & \mathbb{P}_\theta(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \le \delta. \end{aligned} \tag{3.1}$$

From Theorem 2.2.1 and the fact that $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ is non-decreasing in δ, we want to choose $\mathbb{P}_0$ that minimizes $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ so that we can take the maximum δ_ϵ and subsequently achieve overall confident feasibility with the least Monte Carlo sample size. Note that $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ is a multi-input function depending on both $\mathbb{P}_0$ and δ, and so a priori it is not clear that a uniform minimizer $\mathbb{P}_0$ can exist across all values of δ so that the described task is well-defined. It turns out that this is possible in some cases, which we shall investigate in detail. In the following, we discuss results along this line at three levels: The unambiguous case, namely when $\mathcal{U}_{data}$ in (3.1) is a singleton (Section 3.1), the case where $\mathcal{U}_{data}$ is nonparametric (Section 3.2), and the case where $\mathcal{U}_{data}$ is parametric (Section 3.3). The first two cases pave the way to the last one, which is most important to our development and also motivates Section 4. With these results in hand, we also discuss the possibility of using other statistical distances in our framework in Section 3.4.

3.1 Neyman-Pearson Connections and A Least Powerful Null Hypothesis

We first consider, for a given $\theta_1 \in \mathcal{U}_{data}$, the optimization problem

$$\begin{aligned} \max_{A\subset\mathcal{Y}} \quad & \mathbb{P}_{\theta_1}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \le \delta. \end{aligned} \tag{3.2}$$

This problem can be viewed as choosing a most powerful decision rule in a statistical hypothesis test. More precisely, one can think of A as a rejection region for a simple test with null hypothesis $\mathbb{P}_0$ and alternate hypothesis $\mathbb{P}_{\theta_1}$. Subject to a tolerance of δ Type-I error, optimization problem (3.2) looks for a decision rule that maximizes the power of the test. By the Neyman-Pearson lemma [48], under mild regularity conditions on the parametric family, the optimal set $A^\star_{0,\theta_1,\delta}$ of (3.2) takes the form

$$A^\star_{0,\theta_1,\delta} = \{\xi \in \mathcal{Y} : \frac{d\mathbb{P}_{\theta_1}}{d\mathbb{P}_0}(\xi) > K^\star_{0,\theta_1,\delta}\}, \tag{3.3}$$

with $K^\star_{0,\theta_1,\delta}$ chosen so that $\mathbb{P}_0(A^\star_{0,\theta_1,\delta}) = \delta$. Also, then, the optimal value of (3.2) is $\mathbb{P}_{\theta_1}(A^\star_{0,\theta_1,\delta})$. Generalizing the above analysis to all $\theta \in \mathcal{U}_{data}$, we conclude that

$$M(\mathbb{P}_0, \mathcal{U}_{data}, \delta) = \sup_{\theta\in\mathcal{U}_{data}} \mathbb{P}_\theta(A^\star_{0,\theta,\delta}), \tag{3.4}$$

is the optimal value of (3.1). These observations will be useful for deriving our subsequent results.

Our goal is to choose $\mathbb{P}_0$ to minimize (3.4). To start our analysis, let us first consider the extreme case where the uncertainty set $\mathcal{U}_{data}$ consists of only one point $\mathbb{Q}$. In this case, we look for $\mathbb{P}_0$ that minimizes $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta)$, the optimal value of

$$\begin{aligned} \max_{A\subset\mathcal{Y}} \quad & \mathbb{Q}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \le \delta. \end{aligned} \tag{3.5}$$

That is, for a given measure $\mathbb{Q}$, we seek for the maximum discrepancy between $\mathbb{Q}$ and $\mathbb{P}_0$ over all $\mathbb{P}_0$-measure sets that have δ or less content. This is similar to minimizing the total variation distance between $\mathbb{Q}$ and $\mathbb{P}_0$, and hints that the optimal choice of $\mathbb{P}_0$ is $\mathbb{Q}$. The following theorem, utilizing the Neyman-Pearson lemma depicted above, confirms this intuition. We remark that the assumptions of the theorem can be relaxed by using more general versions of the lemma, but the presented version suffices for most purposes and also the subsequent examples we will give.

Theorem 3.1.1. *Given a measure $\mathbb{Q}$ with continuous density on $\mathcal{X}$, among all $\mathbb{P}_0$ such that $\frac{d\mathbb{Q}}{d\mathbb{P}_0}$ exists and is continuous and positive almost surely, the minimum $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta)$ is obtained by choosing $\mathbb{P}_0 = \mathbb{Q}$, giving $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta) = \delta$.*

Proof. Under the assumptions, by the Neyman-Pearson lemma, for a fixed measure $\mathbb{P}_0$, the set achieving the optimal value of (3.5) takes the form $A^\star = \{\xi \in \mathcal{Y} : \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K^\star\}$ for some $K^\star \geq 0$ with $\mathbb{P}_0(A^\star) = \delta$. It then follows that

$$M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta) - \delta = \mathbb{Q}(A^\star) - \mathbb{P}_0(A^\star) = \int_{\frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K^\star} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi).$$

Under the absolute continuity assumption, we define

$$g(K) = \int_{\frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi),$$

which can be seen to be a non-increasing function for $K \geq 1$ and a non-decreasing function for $K \leq 1$. To see this, take $K_1 \geq K_2$, and we have

$$g(K_2) = g(K_1) + \int_{K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi).$$

Thus, when $K_1 \geq K_2 \geq 1$, we have $g(K_2) \geq g(K_1)$ because

$$\int_{K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi) \geq (K_2 - 1)\mathbb{P}_0(K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2) \geq 0,$$

while when $1 \geq K_1 \geq K_2$, we have $g(K_2) \leq g(K_1)$ because

$$\int_{K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi) \leq (K_1 - 1)\mathbb{P}_0(K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2) \leq 0.$$

Then, to identify the minimum of $g(K)$, we either decrease K from 1 to 0 which gives

$$\liminf_{K \to 0} g(K) = \int (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi) = 0, \tag{3.6}$$

by using the dominated convergence theorem (e.g., by considering the set $\{1 > d\mathbb{Q}/d\mathbb{P}_0(\xi) > K\}$) or we increase K from 1 to ∞ which gives

$$\liminf_{K \to \infty} g(K) \geq 0. \tag{3.7}$$

by Fatou's lemma. Observations (3.6) and (3.7) suggest that $g(K) \geq 0$ for all $K \geq 0$ and imply that $g(K^\star) \geq 0$. Thus, we must have $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta) \geq \delta$. Note that this holds for any $\mathbb{P}_0$. Now, since choosing $\mathbb{P}_0 = \mathbb{Q}$ gives $M(\mathbb{Q}, \{\mathbb{Q}\}, \delta) = \delta$, an optimal choice of $\mathbb{P}_0$ is $\mathbb{Q}$. □

Theorem 3.1.1 shows that under mild regularity conditions, in terms of choosing the generating distribution $\mathbb{P}_0$ and minimizing $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta)$, we cannot do better than simply choosing $\mathbb{Q}$ itself. This means that if we had known the true distribution was $\mathbb{Q}$, and without additional knowledge of the event of interest, the safest choice (in the minimax sense) for sampling would be $\mathbb{Q}$, a quite intuitive result. In the language of hypothesis testing, given the simple alternate hypothesis $\mathbb{Q}$, the null hypothesis $\mathbb{P}_0$ that provides the least power for the test, i.e., makes it most difficult to distinguish between the two hypotheses, is $\mathbb{Q}$.

3.2 Nonparametric DRO

Building on the discussion in Section 3.1, we now consider the choice of generating distribution $\mathbb{P}_0$ to minimize the bounding function obtained from (3.1). Before so, we first discuss the nonparametric case, where the analog of (3.1) is in the form:

$$\begin{array}{ll} \max\limits_{d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda, A\subset\mathcal{Y}} & \mathbb{Q}(A) \\ \text{s.t.} & \mathbb{P}_0(A)\le\delta. \end{array} \tag{3.8}$$

for some ball radius $\lambda > 0$, where the decision variables are $\mathbb{Q}$ in the space of all distributions absolutely continuous with respect to $\mathbb{P}_{\hat{\theta}}$, and A.

We show that the above setting can be effectively reduced to the unambiguous case, i.e., when $\mathbb{Q}$ lies in a singleton discussed in Section 3.1. This comes from an established equivalence between a distributionally robust chance constraint and an unambiguous chance constraint evaluated by the center of the divergence ball, when the event A is fixed [37, 40]. In particular, suppose the stochasticity space is $\mathcal{Y} = \mathbb{R}^k$, and $\mathbb{P}_{\hat{\theta}}$ admits a density $p_{\hat{\theta}}$. Theorem 1 in [40] shows that for any A,

$$\max_{d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda} \mathbb{Q}(A)\le\epsilon \iff \mathbb{P}_{\hat{\theta}}(A)\le\epsilon', \tag{3.9}$$

where $\epsilon' = \epsilon'(\epsilon,\lambda,\phi) > 0$ can be explicitly determined by ϵ, λ and ϕ as

$$\epsilon'(\epsilon,\lambda,\phi) = \max\left\{1 - \inf_{\substack{z>0, z+\pi z\le\ell_\phi \\ \underline{m}(\phi^*)\le z_0+z\le\overline{m}(\phi^*)}} \left\{\frac{\phi^*(z_0+z)-z_0-\epsilon z+\lambda}{\phi^*(z_0+z)-\phi^*(z_0)}\right\}, 0\right\} \tag{3.10}$$

with $\phi^*(t) = \sup_x\{tx - g(x)\}$ being the conjugate function of ϕ and $\underline{m}(\phi^*) = \sup\{m\in\mathbb{R} : \phi^* \text{ is a finite constant on } (-\infty, m]\}$, $\overline{m}(\phi^*) = \inf\{m\in\mathbb{R} : \phi^*(m) = +\infty\}$, $\ell_\phi = \lim_{x\to+\infty}\phi(x)/x$, and $\pi = -\infty$ if $Leb\{[p_{\hat{\theta}} = 0]\} = 0$, 0 if $Leb\{[p_{\hat{\theta}} = 0]\} > 0$ and $Leb\{[p_{\hat{\theta}} = 0]\setminus A\} = 0$, and 1 otherwise, where $Leb\{\cdot\}$ is the Lebesgue measure on $\mathbb{R}^k$.

The above equivalence can be used to obtain the following result.

Theorem 3.2.1. *Suppose $\mathcal{Y} = \mathbb{R}^k$ and $\mathbb{P}_{\hat{\theta}}$ admits a density. Among all $\mathbb{P}_0$ such that $\frac{d\mathbb{P}_{\hat{\theta}}}{d\mathbb{P}_0}$ exists and is continuous, positive almost surely, an optimal choice of $\mathbb{P}_0$ that minimizes $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda\},\delta)$, namely the optimal value of* (3.8)*, is the center of the ϕ-divergence ball $\mathbb{P}_{\hat{\theta}}$. Moreover, this gives $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda\},\delta) = \epsilon'^{-1}(\delta,\lambda,\phi)$, where $\epsilon'^{-1}(\cdot,\lambda,\phi)$ is the inverse of the function $\epsilon' = \epsilon'(\epsilon,\lambda,\phi)$ defined in* (3.10) *with respect to ϵ, given by*

$$\epsilon'^{-1}(x,\lambda,\phi) \triangleq \min\{\epsilon\ge 0 : \epsilon'(\epsilon,\lambda,\phi)\ge x\} \tag{3.11}$$

Proof. From Theorem 1 in [40], we know that, for any $A\subset\mathcal{Y}$ and $0\le\epsilon\le 1$, (3.9) holds. We can rewrite the optimal value of problem (3.8) in the form:

$$\begin{array}{ll} \min\limits_{\epsilon\ge 0} & \epsilon \\ \text{s.t.} & \max\limits_{d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda} \mathbb{Q}(A)\le\epsilon \text{ for all } A\subset\mathcal{Y} \text{ such that } \mathbb{P}_0(A)\le\delta, \end{array} \tag{3.12}$$

which, according to (3.9), has the same optimal value as

$$\begin{array}{ll} \min\limits_{\epsilon\ge 0} & \epsilon \\ \text{s.t.} & \mathbb{P}_{\hat{\theta}}(A)\le\epsilon' \text{ for all } A\subset\mathcal{Y} \text{ such that } \mathbb{P}_0(A)\le\delta. \end{array} \tag{3.13}$$

Since, fixing ϕ and λ, ϵ' is a non-decreasing function of ϵ, we see that minimizing ϵ is equivalent to minimizing ϵ'. Denoting $v^\star$ as the optimal value of the optimization problem

$$\begin{array}{ll} \max\limits_{A\subset\mathcal{Y}} & \mathbb{P}_{\hat{\theta}}(A) \\ \text{s.t.} & \mathbb{P}_0(A) \le \delta, \end{array} \tag{3.14}$$

then the optimal value of (3.13) is $\epsilon'^{-1}(v^*, \lambda, \phi)$. Moreover, this is achievable by setting $\mathbb{P}_0 = \mathbb{P}_{\hat{\theta}}$ that gives the optimal value $v^* = \delta$ to (3.14) by Theorem 3.1.1.

□

An implication of Theorem 3.2.1 is that, by noting that a parametric divergence ball lies inside a corresponding nonparametric ball, we can compute a bound for M to obtain a required Monte Carlo size, drawn from the baseline $\mathbb{P}_{\hat{\theta}}$, to get a feasible solution for the distributionally robust CCP (2.2) and subsequently the CCP (1.1). More precisely, recall the bounding function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta) = M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ with $\lambda = \phi''(1)\chi^2_{1-\alpha,D}/(2n)$, given by (3.1), as the optimal value of

$$\begin{array}{ll} \max\limits_{d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda, \mathbb{Q}\in\mathcal{P}_{para}, A\subset\mathcal{Y}} & \mathbb{Q}(A) \\ \text{s.t.} & \mathbb{P}_0(A) \le \delta. \end{array} \tag{3.15}$$

We have:

Corollary 3.2.2. *Given a data size n, suppose $\mathcal{Y} = \mathbb{R}^k$ and $\mathbb{P}_{\hat{\theta}}$ admits a density, where $\hat{\theta}$ is the MLE under Assumption A1. If we choose $\delta_\epsilon = \epsilon'(\epsilon, \phi''(1)\chi^2_{1-\alpha,D}/(2n), \phi)$ and draw $N_{exact}(\delta_\epsilon, \beta, d)$ Monte Carlo samples from the generating distribution $\mathbb{P}_{\hat{\theta}}$ to construct the sampled problem* (2.4)*, then the obtained solution will be feasible for* (1.1) *with asymptotic confidence level at least $1 - \alpha - \beta$.*

Proof. Note that a parametric divergence ball lies inside a corresponding nonparametric ball in the sense that

$$\{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}$$

Thus, by the definition of M, we have

$$M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) \le M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta)$$

In particular,

$$M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) \le M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta) = \epsilon'^{-1}(\delta, \lambda, \phi)$$

where the equality follows from Theorem 3.2.1. Thus, if we choose δ_ϵ such that $\epsilon'^{-1}(\delta_\epsilon, \lambda, \phi) \le \epsilon$, or $\delta_\epsilon = \epsilon'(\epsilon, \lambda, \phi)$, where $\lambda = \phi''(1)\chi^2_{1-\alpha,D}/(2n)$ as presented in (2.17), and the generating distribution as $\mathbb{P}_{\hat{\theta}}$, then Corollary 2.2.2 guarantees that running SO on $N_{exact}(\delta_\epsilon, \beta, d)$ Monte Carlo samples gives a feasible solution for (1.1) with confidence asymptotically at least $1 - \alpha - \beta$. □

Corollary 3.2.2 thus provides an implementable procedure to handle (1.1) through (2.2).

3.3 Parametric DRO

Next we discuss further the choice of generating distributions in parametric DRO beyond $\mathbb{P}_{\hat{\theta}}$. While the ball center $\mathbb{P}_{\hat{\theta}}$ is a valid choice, the equivalence relation (3.9) does not apply when the divergence ball is in a parametric class, and the optimal choice of the generating distribution may no longer be $\mathbb{P}_{\hat{\theta}}$, as shown in the next result.

Theorem 3.3.1. *In terms of selecting a generating distribution $\mathbb{P}_0$ to minimize $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$, the optimal value of* (3.15)*, the choice $\mathbb{P}_{\hat{\theta}}$ can be strictly dominated by other distributions.*

Intuitively, Theorem 3.3.1 arises because the extreme distribution that achieves the equivalence relation (3.9) may not be in the considered parametric family. It implies more flexibility in choosing the generating measure $\mathbb{P}_0$, in the sense of requiring less Monte Carlo samples than using $\mathbb{P}_{\hat{\theta}}$.

From the standpoint of hypothesis testing in Section 3.1, the imposed minimax problem (3.15) in searching for the best $\mathbb{P}_0$ can be viewed as finding a simple null hypothesis that is uniformly least powerful across the uncertainty set. This question is related and appears more general than finding the least favorable or powerful prior in testing against composite null hypothesis [48]. In the latter context, given a set Θ_1, one aims to find a distribution $\mu^\star(d\theta_0)$ such that $\Gamma(\mu^\star) \leq \Gamma(\mu)$ for all distributions $\mu(d\theta_0)$ on Θ_0, where $\Gamma(\mu)$ is the optimal value of

$$\begin{aligned} \max_{\theta_1 \in \Theta_1} \quad & \mathbb{P}_{\theta_1}(A) \\ \text{s.t.} \quad & \int_{\Theta_0} \mathbb{P}_{\theta_0}(A)\mu(d\theta_0) \leq \delta. \end{aligned} \tag{3.16}$$

The distribution $\mu(d\theta_0)$ is interpreted as a prior on a composite null hypothesis parametrized by θ_0, and $\mu^*(d\theta_0)$ is the least favorable prior. The difference between (3.16) and our formulation (3.15) lies in the restriction to measures of the form $\mathbb{P}_0 = \int_{\Theta_0} \mathbb{P}_{\theta_0}\mu(d\theta_0)$ for the former, leading to a smaller search space than ours. This mixture-type $\mathbb{P}_0$ and the Bayesian connection will partly motivate our investigation in Section 4.

To prove Theorem 3.3.1, we present a counter example and also some related discussion.

Example 3.3.2. *Consider the uncertainty set $\mathcal{U}_{data} = \{\mathbb{P}_\theta, : -1 \leq \theta \leq 1\}$ within Gaussian location family on $\mathbb{R}$ with $\mathbb{P}_\theta(dy) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\theta)^2}{2}}$. This can be thought of, e.g., as an uncertainty set based on the χ^2-distance, the latter defined between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ as*

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) = \int_{\mathcal{Y}} (\frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1)^2 \mathbb{P}_1(dy). \tag{3.17}$$

Note that the χ^2-distance is in the family of ϕ-divergences, by choosing $\phi = (x-1)^2$. We aim to find a generating distribution $\mathbb{P}_0$ to minimize $M(\mathbb{P}, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, \delta)$, the optimal value of

$$\begin{aligned} \max_{\theta \in \mathcal{U}_{data}, A \subset \mathbb{R}} \quad & \mathbb{P}_\theta(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{3.18}$$

We consider several symmetric distributions as $\mathbb{P}_0$ (symmetry is reasonably conjectured as a good property since an imbalanced shift might increase the power for the alternative hypothesis on one side and the worst case overall). We list these symmetric distributions in increasing variability:

$$\begin{aligned} \mathbb{P}_0^1(dy) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\ \mathbb{P}_0^2(dy) &= \frac{1}{\sqrt{2\pi \cdot 2}} e^{-\frac{y^2}{2 \cdot 2}} \\ \mathbb{P}_0^3(dy) &= \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(y-1)^2}{2}} + e^{-\frac{(y+1)^2}{2}} \right). \end{aligned} \tag{3.19}$$

Given $0 \leq \theta \leq 1$, it can be shown by the Neyman-Pearson lemma that the rejection region $A^\star$ (i.e. the set giving the optimal value of (3.18) for a given θ) for $\mathbb{P}_0^1$ has the form $\{y : y > c_1\}$, for $\mathbb{P}_0^2$ the form $\{y : |y - 2\theta| \leq c_2\}$ and for $\mathbb{P}_0^3$ the form $\{y : \frac{e^{\theta y}}{e^y + e^{-y}} > c_3\}$, for some c_1, c_2 and c_3. Let $\delta = 0.05$ be the

tolerance level, it can be shown through numerical verification that

$$\begin{aligned} M(\mathbb{P}_0^1, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, 0.05) &= 0.2595 \\ M(\mathbb{P}_0^2, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, 0.05) &= 0.1160 \\ M(\mathbb{P}_0^3, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, 0.05) &= 0.0995. \end{aligned} \tag{3.20}$$

Thus, the natural choice $\mathbb{P}_{\hat{\theta}} = \mathbb{P}_0^1$ based on relaxing to nonparametric DRO yields a bounding function $M(\cdot)$ that is outperformed by $\mathbb{P}_0^2$ or $\mathbb{P}_0^3$. Later in Section 4 we will see numerically how $\mathbb{P}_0^2$ and $\mathbb{P}_0^3$ can lead to a smaller sample size requirements.

Although Theorem 3.3.1 reveals room to search for the best generating distribution, the involved optimization, or even just finding an improved distribution over $\mathbb{P}_{\hat{\theta}}$, appears to be nontrivial. In particular, the maximization problem in (3.15) depends on the computation of $A^\star$ for each alternative of $\theta \in \mathcal{U}_{data}$. Section 4 discusses some approaches to search for improvements. We conclude the current section with some discussion on the choice of statistical distances used in the uncertainty set.

3.4 Choice of Statistical Distance

We have chosen to use ϕ-divergence to construct our uncertainty set $\mathcal{U}_{data}$, and we have seen how this allows us to effectively translate sample size requirements from the data to Monte Carlo. Note that another common type of distance is the Wasserstein distance (e.g., [6, 25, 27]). If one can translate the violation probability under a generating distribution into the worst-case violation probability over a Wasserstein ball, then the same line of arguments in Section 2 applies to using SO on this DRO. Presuming that the size of a parametric Wasserstein-based confidence region can be properly calibrated from data, it is conceivable that the above can give rise to an alternate solution route. It is known (Theorem 3 in [6]), under suitable regularity conditions, that one can equate a Wasserstein-ambiguous probability $\sup_{d_W(\mathbb{Q},\mathbb{P}_{\hat{\theta}}) \le \lambda} \mathbb{Q}(\xi \in A)$, where d_W denotes a Wasserstein distance of order 1 and cost function c, and A is an event, to $\mathbb{P}_{\hat{\theta}}(c(\xi, A) \le 1/\nu^*)$ where $\nu^* \ge 0$ is a dual multiplier for the associated optimization problem, and $c(\xi, A)$ denotes the cost-induced distance between a point ξ and a set A. Thus, $M(\mathbb{P}_0, \{\mathbb{Q} : d_W(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta)$ can be written as

$$\begin{aligned} \max_{A \subset \mathcal{Y}} \quad & \mathbb{P}_{\hat{\theta}}(c(\xi, A) \le 1/\nu^*) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \le \delta. \end{aligned} \tag{3.21}$$

Compared to the evaluation of $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta)$ in Theorem 3.2.1, the tightening of the tolerance level from ϵ to ϵ' is now replaced by the set inflation from A to the $(1/\nu^*)$-neighborhood of A given by $\{\xi : c(\xi, A) \le 1/\nu^*\}$. Note that, regardless of the distance used, one could reduce the conservativeness of our analysis by focusing on A in the form $\{x \notin \mathcal{X}_\xi\}$, but this would require looking at the specific form of the safety set $\mathcal{X}_\xi$.

4 IMPROVING GENERATING DISTRIBUTIONS

This section discusses some approaches to search for better generating distributions beyond the baseline distribution in a divergence ball of DRO. Section 4.1 first states a general result to create better generating distributions. Section 4.2 then specializes to using a mixture distribution on θ to exploit this result. Sections 4.3 and 4.4 then provide two specific ways to construct these mixtures. Finally, Section 4.5 demonstrates some numerical comparisons in using these new mixing generating distributions and also simply using the baseline.

4.1 A Framework to Reduce Divergence Ball Size by Incorporating Parametric Information

The reason why the best choice of generating distribution $\mathbb{P}_0$ is not the baseline of the divergence ball, $\mathbb{P}_{\hat{\theta}}$, in minimizing $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ is that the equivalence relation (3.9) does not hold when $\mathbb{Q}$ is restricted to a parametric class. In some sense the reduction to the unambiguous chance constraint in the right hand side of (3.9) is over-conservative as it does not account for parametric information. Suppose we would still like to use the analytically tractable relation (3.9), but at the same time be less conservative. Then, one approach is to find a new baseline distribution, say $\tilde{\mathbb{P}}$, such that the parametrically restricted divergence ball $\{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}$ lies inside a new nonparametric divergence ball at the center $\tilde{\mathbb{P}}$, namely $\{\mathbb{Q} : d_\phi(\tilde{\mathbb{P}}, \mathbb{Q}) \le \tilde{\lambda}\}$. If we can obtain a nonparametric ball size $\tilde{\lambda}$ such that $\tilde{\lambda} < \lambda$ and the set inclusion holds, then this new ball is also a valid uncertainty set, and, when simply setting the generating distribution as $\mathbb{P}_0 = \tilde{\mathbb{P}}$ and applying Theorem 3.2.1, we have a smaller upper bound for $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ than $\epsilon'^{-1}(\delta, \lambda, \phi)$ obtained from using Theorem 3.2.1 directly with the parametric constraint relaxed.

To above mechanism can be executed as follows. Let $\mathcal{U}_{data} = \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}$. For any $\mathbb{P}_0$, let

$$\mathcal{D}_{data}(\mathbb{P}_0, \phi) \triangleq \sup_{\mathbb{Q} \in \mathcal{U}_{data}} d_\phi(\mathbb{P}_0, \mathbb{Q}). \tag{4.1}$$

Then we clearly have

$$\mathcal{U}_{data} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_0, \mathbb{Q}) \le \mathcal{D}_{data}(\mathbb{P}_0, \phi)\}, \tag{4.2}$$

since the right-hand-side set includes distributions outside of the parametric family as well.

Our goal is to find $\mathbb{P}_0$ to minimize $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ or any upper bound of $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ so that it is smaller than the ball size λ appearing in the original parametric divergence ball $\mathcal{U}_{data}$. We state the implication of this as follows:

Theorem 4.1.1. *Suppose $\mathcal{Y} = \mathbb{R}^k$ and $\mathbb{P}_{\hat{\theta}}$ admits a density. Consider the parametric divergence ball $\mathcal{U}_{data} = \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}$. Suppose we can find $\mathbb{P}_0$ such that $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ defined in (4.1) satisfies $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \lambda$. Then we have*

$$\begin{aligned}\min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) &\le \min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \mathcal{D}_{data}(\mathbb{P}_0, \phi)\}, \delta) \\ &\le \min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta) \end{aligned} \tag{4.3}$$

and

$$\min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) \le \epsilon'^{-1}(\delta, \mathcal{D}_{data}(\mathbb{P}_0, \phi), \phi) \le \epsilon'^{-1}(\delta, \lambda, \phi) \tag{4.4}$$

where $\epsilon'^{-1}(\epsilon, \lambda, \phi)$ is defined in (3.11).

Proof. By the definition of $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$, (4.2) holds. Together with the condition $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \lambda$, we have the set inclusions

$$\{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_0, \mathbb{Q}) \le \mathcal{D}_{data}(\mathbb{P}_0, \phi)\} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\} \tag{4.5}$$

The inequalities (4.3) then follow from the definition of M. The inequalities (4.4) in turn follow immediately from Theorem 3.2.1. □

Theorem 4.1.1 stipulates that choosing $\mathbb{P}_0$ depicted in the theorem as the generating distribution, and setting $\epsilon'^{-1}(\delta, \mathcal{D}_{data}(\mathbb{P}_0, \phi), \phi)$ as an upper bound for $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ to obtain the required Monte Carlo size $N_{exact}(\delta_\epsilon, \beta, d)$ implied by Corollary 2.2.2, will give a

lighter Monte Carlo requirement than using the bound $\epsilon'^{-1}(\delta, \lambda, \phi)$ directly obtained by relaxing the parametric constraint and using $\mathbb{P}_{\hat{\theta}}$ as the generating distribution as in Corollary 3.2.2.

4.2 Mixture as Generating Distribution

Since optimization (4.1) can be difficult to solve generally, we focus on finding improved generating distribution $\mathbb{P}_0$ so that the implication of Theorem 4.1.1 holds, instead of fully optimizing (4.1). In this and the next subsections, we design a search space $\mathcal{P}_0$ for $\mathbb{P}_0$ that allows the construction of tractable procedures to achieve such improvements, while at the same time ensures the obtained $\mathbb{P}_0$ are amenable to Monte Carlo simulation.

From now on we will focus on χ^2-distance as our choice of ϕ for convenience (as will be seen). Suppose that $\mathbb{P}_\theta$ has density $p(y;\theta)$. We then set $\mathcal{P}_0$ to be the collection of distributions with densities in the form

$$p_0(y) = \int_\Theta p(y;\theta)\mu(d\theta), \tag{4.6}$$

for some probability measure μ on Θ. This class of distributions is easy to sample assuming $p(y;\theta)$ and μ are, as one can first sample $\theta \sim \mu(d\theta)$ and then $\xi \sim \mathbb{P}_\theta$ given θ.

Searching for the best $p_0(y)$ requires minimizing $\mathcal{D}_{data}(\mathbb{P}_0)$ over $\mathbb{P}_0 \in \mathcal{P}_0$ (where for convenience we denote $\mathcal{D}_{data}(\mathbb{P}_0)$ as $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ with ϕ representing the χ^2-distance). We first use (3.17) to write

$$\begin{aligned}
\mathcal{D}_{data}(\mathbb{P}_0) &= \sup_{\theta\in\mathcal{U}_{data}} \int_{\mathcal{Y}} \left(\frac{p(y;\theta)}{p_0(y)} - 1\right)^2 p_0(y)dy \\
&= \sup_{\theta\in\mathcal{U}_{data}} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{p_0(y)}dy - 1 \\
&= \sup_{\theta\in\mathcal{U}_{data}} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu(d\theta')}dy - 1.
\end{aligned} \tag{4.7}$$

Denoting $\mathcal{P}(\Theta)$ as the space of probability measures on Θ, we define the function $L : \mathcal{P}(\Theta)\times\Theta \to \mathbb{R}$ to be

$$L(\mu,\theta) \triangleq \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu(d\theta')}dy, \tag{4.8}$$

assuming the integral is well-defined for $\mathcal{P}(\Theta)\times\Theta$ and further define

$$l(\mu) \triangleq \sup_{\theta\in\mathcal{U}_{data}} L(\mu,\theta). \tag{4.9}$$

Thus (4.7) can be written as $\mathcal{D}_{data}(\mathbb{P}_0) = l(\mu) - 1$, and minimizing $\mathcal{D}_{data}(\mathbb{P}_0)$ is equivalent to solving

$$\min_{\mu\in\mathcal{P}(\Theta)} l(\mu) = \min_{\mu\in\mathcal{P}(\Theta)} \max_{\theta\in\mathcal{U}_{data}} L(\mu,\theta). \tag{4.10}$$

Optimization (4.10) has the following convexity property:

Lemma 4.2.1. *The outer minimization in problem* (4.10) *is convex.*

Lemma 4.2.1 can be proved by direct verification, which is shown in Appendix C. Note also that, if μ is the point mass δ_θ for $\theta \in \Theta$, then the mixture distribution would recover the parametric distribution $\mathbb{P}_\theta$. Hence the proposed family $\mathcal{P}_0$ includes $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$, and in particular the original baseline distribution $\mathbb{P}_{\hat{\theta}}$. Although the outer minimization of (4.10) is a convex problem, computing $l(\mu)$ involves a non-convex optimization and is difficult in general. Our approach is to search for a descent direction for the convex function $l(\cdot)$ from $\delta_{\hat{\theta}}$. In the following, we will study two types of

search directions, each using its own version of Danskin's Theorem [3, 4]. To proceed, we introduce the following definition:

Definition 4.2.2. Define $\Theta^*(\mu)$ to be the set of optimal points for the maximization problem in $l(\mu) = \sup_{\theta \in \mathcal{U}_{data}} L(\mu, \theta)$ given $\mu \in \mathcal{P}(\Theta)$:

$$\Theta^*(\mu) = \text{argmax}_{\theta \in \mathcal{U}_{data}} L(\mu, \theta) \tag{4.11}$$

It can be shown that $\Theta^*(\mu)$ is non-empty and $\Theta^*(\mu) \subseteq \mathcal{U}_{data}$ because $\mathcal{U}_{data}$ is compact and $L(\mu, \theta)$ is continuous in θ.

4.3 Mixing with a Proposed Distribution

We consider mixing distributions in the form $(1-t)\delta_{\hat{\theta}} + t\mu_{prop}$ for some proposed distribution μ_{prop}, and look for a descent direction by varying t from 0 to 1. We have the following result that is a consequence of Danskin's Theorem that involves a one-sided derivative. We provide proofs both for this theorem and our following result in Appendix C.

Theorem 4.3.1. *Fix any $\mu_1, \mu_2 \in \mathcal{P}(\Theta)$ and $\theta \in \Theta$. Under the assumptions that $\psi(t) = L((1-t)\mu_1 + t\mu_2, \theta)$ is well defined for $0 \le t \le 1$, we know that the function $g(y,t)$*

$$g(y,t) : \mathcal{Y} \times [0,1] \triangleq \frac{(p(y;\theta))^2}{(1-t)\int_\Theta p(y;\theta')\mu_1(d\theta') + t\int_\Theta p(y;\theta')\mu_2(d\theta')}$$

is integrable for $t \in [0,1]$. If we further assume that there exists a integrable function $g_0(y)$ such that

$$\left| \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\mu_1 - \mu_2)(d\theta')}{(\int_\Theta p(y;\theta')((1-t)\mu_1 + t\mu_2)(d\theta'))^2} \right| \le g_0(y),$$

then we have the right derivative of $\psi(t)$ at $t = 0$ given by

$$\begin{aligned} \psi^+(0) &= \sup_{\theta \in \Theta^*(\mu_1)} \lim_{t \downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t} \\ &= \sup_{\theta \in \Theta^*(\mu_1)} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\mu_1 - \mu_2)(d\theta')}{(\int_\Theta p(y;\theta')\mu_1(d\theta'))^2} dy. \end{aligned} \tag{4.12}$$

The quantity $\psi^+(0)$ is the directional derivative of $L(\mu_1)$ in the direction $\mu_2 - \mu_1$. Thus, to improve on $\mathcal{D}_{data}(\mathbb{P}_{\hat{\theta}})$, we can propose a mixing distribution $\mu_{prop}(d\theta')$, and substitute $\mu_1 = \delta_{\hat{\theta}}$ and $\mu_2 = \mu_{prop}$ in (4.12) to check if

$$\sup_{\theta \in \Theta^*(\delta_{\hat{\theta}})} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\delta_{\hat{\theta}} - \mu_{prop})(d\theta')}{p(y;\hat{\theta})^2} dy < 0, \tag{4.13}$$

which indicates a strict descent for $l(\cdot)$ from $\delta_{\hat{\theta}}$ to μ_{prop}. In this case, it follows from the convexity of $l(\cdot)$ that we can find some $0 < t \le 1$ such that $l((1-t)\delta_{\hat{\theta}} + t\mu_{prop}) < l(\delta_{\hat{\theta}})$, so that

$$p_t(y) = \int_\Theta p(y;\theta')((1-t)\delta_{\hat{\theta}} + t\mu_{prop})(d\theta'), \tag{4.14}$$

gives rise to $\mathcal{D}_{data}(\mathbb{P}_0) < \mathcal{D}_{data}(\mathbb{P}_{\hat{\theta}})$. Finding such a t can be done by a bisection search or enumerating $\mathcal{D}_{data}(\mathbb{P}_0)$ on p_t over a grid of t. Note that the above can be implemented only if (4.13) can be verified and also if $\mathcal{D}_{data}(\mathbb{P}_0)$ is computable. We will show that both properties are satisfied for the case of multivariate Gaussian when μ_{prop} is properly chosen. In particular, we will identify general sufficient conditions for μ_{prop} to guarantee (4.13), and also find μ_{prop} such that

the maximization involved in computing $\mathcal{D}_{data}(\mathbb{P}_0)$ in (4.7) can be reduced to a one-dimensional problem.

Consider a multivariate Gaussian distribution with unknown mean $\Theta \subset \mathbb{R}^D$ in an open convex set with density

$$p(y;\theta) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \cdot e^{-\frac{1}{2}(y-\theta)^\intercal \Sigma^{-1}(y-\theta)}, \tag{4.15}$$

where Σ is a fixed positive semi-definite covariance matrix. Direct verification (in Appendix C) shows that

$$\begin{aligned}\mathcal{U}_{data} \triangleq \left\{\theta \in \Theta : \chi^2(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_\theta) \le \frac{\chi^2_{1-\alpha,D}}{n}\right\} &= \left\{\theta \in \Theta : e^{(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta-\hat{\theta})} - 1 \le \frac{\chi^2_{1-\alpha,D}}{n}\right\} \\ &= \left\{\theta : \hat{\theta} + \Sigma^{\frac{1}{2}} v, \quad \text{for } \|v\|_2^2 \le \log(1 + \frac{\chi^2_{1-\alpha,D}}{n})\right\}, \end{aligned} \tag{4.16}$$

and thus

$$\Theta^*(\delta_{\hat{\theta}}) = \text{argmax}_{\theta \in \mathcal{U}_{data}} e^{(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta-\hat{\theta})} = \left\{\theta : \hat{\theta} + \Sigma^{\frac{1}{2}} v, \quad \text{for } \|v\|_2^2 = \log(1 + \frac{\chi^2_{1-\alpha,D}}{n})\right\}. \tag{4.17}$$

We propose the following μ_{prop}. First, we call a distribution on Θ symmetrical around $\theta \in \Theta$ if its probability density or mass function has the same value for any $\theta_1, \theta_2 \in \Theta$ such that $\theta = \frac{\theta_1+\theta_2}{2}$.

Proposition 4.3.2. *Let $\mu_{prop}(d\theta')$ be any symmetrical distribution around $\hat{\theta}$. Given $\theta \in \Theta^*(\delta_{\hat{\theta}})$, we define $Y_\theta = (\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})$ with $\theta' \sim \mu_{prop}(d\theta')$. Suppose there exists an integrable random variable Y under the measure μ_{prop} such that $e^{2Y_\theta} \le Y$ for all $\theta \in \Theta^*(\delta_{\hat{\theta}})$. If, for each $\theta \in \Theta^*(\delta_{\hat{\theta}})$, Y_θ does not equal to 0 with probability 1, then* (4.13) *holds and the mixture distribution produced by $\mu_{prop}(d\theta)$ would result in a descent direction on $\mathcal{D}_{data}(\mathbb{P}_{\hat{\theta}})$.*

One can check that any Gaussian distribution with mean $\hat{\theta}$ satisfies the conditions of Proposition 4.3.2, and so does any $\mu_{prop}(d\theta')$ that is discrete, symmetrical around $\hat{\theta}$, whose outcome directions $\theta' - \theta$ constitute a basis of $\mathbb{R}^D$. Alternately, we also consider the following continuous μ_{prop}. We set $\theta' \sim \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta$ where η is a random vector uniformly distributed on the surface of the D-dimension unit ball. Note that this θ' can be efficiently simulated by sampling D independent standard Gaussian random variables and scaling their norm to unit length to obtain η. While this μ_{prop} can be readily checked to satisfy the conditions in Proposition 4.3.2, we also provide an alternate proof on the validity of this μ_{prop} in achieving a descent direction in Lemma C.0.5 in the Appendix, as results proven therein provide important reference to calculations in numerical experiments regarding μ_{prop}.

Next, we discuss the computation of $\mathcal{D}_{data}(\mathbb{P}_0)$ for a given $\mathbb{P}_0$. First, we call a random variable Y on $\mathcal{Y} \subset \mathbb{R}^k$ rotationally invariant if $Y \overset{\mathcal{D}}{=} Q^\intercal Y$ for any rotational matrix $Q \in \mathbb{R}^{k\times k}$. Using this notion, the following shows how one can reduce the D-dimensional maximization problem in the definition of $\mathcal{D}_{data}(\mathbb{P}_0)$ into a one-dimensional problem.

Proposition 4.3.3. *Given a nominal distribution $Y \sim \mathbb{P}_0$ and a multivariate Gaussian family with known covariance Σ denoted $\mathbb{P}_\theta = \mathcal{N}(\theta, \Sigma)$. If the nominal distribution $Y \sim \mathbb{P}_0$ satisfies the condition that the random variable $Z = \Sigma^{-1/2}(Y - \hat{\theta})$ is rotationally invariant, then for any θ_1, θ_2 satisfying $(\theta_1 - \hat{\theta})^\intercal \Sigma^{-1}(\theta_1 - \hat{\theta}) = (\theta_2 - \hat{\theta})^\intercal \Sigma^{-1}(\theta_2 - \hat{\theta})$, we have*

$$\chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_1}) = \chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_2}). \tag{4.18}$$

Thus, for $\mathcal{D}_{data}(\mathbb{P}_0) = \max_{\theta\in\mathcal{U}_{data}} \chi^2(\mathbb{P}_0, \mathbb{P}_\theta)$ *with* $\mathcal{U}_{data} = \{\theta \in \Theta : (\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta - \hat{\theta}) \le \lambda\}$ *as in* (4.16)*, we have*

$$\mathcal{D}_{data}(\mathbb{P}_0) = \max_{0\le t\le 1} \chi^2(\mathbb{P}_0, \mathbb{P}_{(1-t)\hat{\theta}+t\theta^\star}), \tag{4.19}$$

given any $\theta^\star$ *satisfying* $(\theta^\star - \hat{\theta})^\intercal \Sigma^{-1}(\theta^\star - \hat{\theta}) = \lambda$.

Proposition 4.3.4. *Given* $0 \le t \le 1$ *and* $\mu_{prop}(d\theta) = \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta$*, where* η *is a random vector uniformly distributed on the surface of the* D*-dimension unit ball, the nominal measure* $\mathbb{P}_t$ *with density*

$$p_t(y) = \int_\Theta p(y;\theta')((1-t)\delta_{\hat{\theta}} + t\mu_{prop})(d\theta') = (1-t)\mathbb{P}_{\hat{\theta}} + t\int_\Theta p(y;\theta')\mu_{prop}(d\theta'),$$

satisfies the conditions in Proposition 4.3.3.

Therefore, in computing $\mathcal{D}_{data}(\mathbb{P}_0)$ derived from the proposed distribution $\mu_{prop}(d\theta) = \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta$, using Propositions 4.3.3 and 4.3.4 we can change the domain of the involved maximization from $\Theta \subset \mathbb{R}^D$ into $\mathbb{R}$, leading to a substantial reduction in the search space and a tractable problem.

4.4 Enlarging Mixture Variability

Our next proposal is to consider a continuous mixing distribution $\mu_r(d\theta')$ on Θ where $r \ge 0$ controls the variability of the distribution, so that $r = 0$ corresponds to $\delta_{\hat{\theta}}$. Here, we can parametrize the density of the generating distribution as

$$p_r(y) = \int_\Theta p(y;\theta')\mu_r(d\theta'), \tag{4.20}$$

and our search direction is along r starting from $r = 0$. We propose two possible ways to define $\mu_r(d\theta')$. First is to let $\mu_r^1(d\theta')$ follow the distribution of $\theta' \sim \hat{\theta} + \Sigma^{\frac{1}{2}} \cdot \eta_{\sqrt{r}}$ where $\eta_{\sqrt{r}}$ is the uniform distribution inside the D-dimensional unit ball with radius $\sqrt{r}$. Second is to let $\mu_r^2(d\theta')$ follow $\mathcal{N}(\hat{\theta}, r\Sigma)$. The second approach in particular can be intuited as the posterior distribution of the parameter from a Bayesian perspective. In both cases, we notice that letting $r = 0$ would recover the original baseline distribution $p(y;\hat{\theta})$.

To analyze these schemes, we abuse notation slightly and now define $L : \mathbb{R}^+ \times \Theta \to \mathbb{R}$ to be

$$L(r,\theta) \triangleq \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{p_r(y)} dy, \tag{4.21}$$

and

$$l(r) \triangleq \sup_{\theta\in\mathcal{U}_{data}} L(r,\theta). \tag{4.22}$$

We show that increasing r to positive values would produce a descent direction for $l(r)$ at $r = 0$, when the underlying distribution is Gaussian. Recall that in this case $\Theta^*(\delta_{\hat{\theta}})$ can be expressed by (4.17). As $l(r)$ is not necessarily convex in this situation, we use a generalized version of Danskin's Theorem [18] for non-convex problems to get the following result:

Theorem 4.4.1. *With* $l(r)$ *and* $L(r,\theta)$ *defined in* (4.21) *and* (4.22)*, and* $p(y;\theta)$ *multivariate Gaussian with mean* θ *and known positive definite covariance* Σ*, we have*

$$l^+(0) = \lim_{r\downarrow 0} \frac{l(r) - l(0)}{r} = (1 + \frac{\chi^2_{1-\alpha,D}}{n}) \cdot \lim_{r\downarrow 0} \frac{1}{r}\Big(1 - \inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_r}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}]\Big) \tag{4.23}$$

The proof is in Appendix C. With Theorem 4.4.1, we can show that both μ_r^1 and μ_r^2 proposed above are valid choices to produce descent directions. Moreover, we can also show that they allow tractable computation of $\mathcal{D}_{data}(\mathbb{P}_0)$. These are depicted as follows.

Corollary 4.4.2. *Under the assumptions in Theorem 4.4.1, $l^+(0) < 0$ for both μ_r^1 and μ_r^2.*

Corollary 4.4.3. *Given $r \geq 0$ and μ_{prop} being $\mu_r^1(d\theta)$ or $\mu_r^2(d\theta)$, the nominal measure $\mathbb{P}_r$ with density given by* (4.20) *satisfies the conditions in Proposition 4.3.3.*

The proofs of Corollaries 4.4.2 and 4.3.4 are in Appendix C.

4.5 Numerical Demonstrations

To confirm our findings in Section 4.3 and 4.4, we perform several numerical experiments. Consider $\mathbb{P}_\theta$ to be multivariate Gaussian $\mathcal{N}(\theta, I_D)$ with $k = D = 10$. We set $\epsilon = \alpha = 0.05$ while $\beta = 0.01$ and data size $n = 10$ or 5. Notice in this case, the dimension D is high but the available sample n is low and we would actually need $N_{exact} = 371$ data points to perform standard SO. Based on our discussion, we compare three choices of μ_{prop}:

- $\mu_1 = \delta_{\hat{\theta}}$, the point mass at $\hat{\theta}$.
- $\mu_2 \sim \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \eta$, where η is the uniform random vector on the surface of a D-dimension unit ball, discussed in Section 4.3.
- $\mu_3 \sim \mathcal{N}(\hat{\theta}, I_D/n)$, the Gaussian distribution with mean $\hat{\theta}$ and covariance matrix I_D/n, discussed in Section 4.4.

For μ_1, μ_2 and μ_3, the calculation of $\mathcal{D}_{data}(\mathbb{P}_0)$ is tractable. We leave the details in the Appendix as remarks following Lemma C.0.5 and summarize the results in Table 1 and 2. We use N to denote the number of Monte Carlo samples needed. Moreover, we use both algorithms Extended SO and Extended FAST discussed in Section 5 for demonstration. As we can see, the decrease in N under a better sampling distribution can be considerable, down to less than a third compared to using the baseline in some cases. Mixing with a proposed uniform distribution (μ_2) appears to reduce N more than applying a Gaussian mixture (μ_3). As a side note, we also observe Extended FAST requires significantly less sample size than Extended SO in this example.

Table 1. Comparisons among choices of $\mathbb{P}_0$ for 10 dimensional multivariate Gaussian when $n = 5$.

	$\mathcal{D}_{data}(\mathbb{P}_0)$	δ_ϵ	N for Extended SO	N for Extended FAST
$\mu_1(\delta_{\hat{\theta}})$	37.9161	6.5766×10^{-5}	285601	70221
μ_2	11.0368	2.2454×10^{-4}	83649	20707
μ_3	14.7391	1.6850×10^{-4}	111465	27528

5 PROCEDURAL DESCRIPTION

This section presents our procedures to find solutions for CCP (1.1) using SO-based methods, when the direct use of data $\xi_1, ..., \xi_n$ from $\mathbb{P}$ is possibly insufficient to achieve feasibility with a given confidence. Algorithm 1, which we call "Extended SO", first presents the basic and most easily applicable procedure arising from Corollary 3.2.2. Notice that, given an overall target confidence level, say c, we have flexibility in choosing α and β such that $\alpha + \beta = c$. In our experiments, we simply choose $\alpha = \beta = c/2$. However, if the requirement for confidence level is high, it is a more

Table 2. Comparisons among choices of $\mathbb{P}_0$ for 10 dimensional multivariate Gaussian when $n = 10$.

	$\mathcal{D}_{data}(\mathbb{P}_0)$	δ_ϵ	N for Extended SO	N for Extended FAST
$\mu_1(\delta_{\hat{\theta}})$	5.2383	4.6857×10^{-4}	40081	10026
μ_2	3.3139	7.3298×10^{-4}	25621	6481
μ_3	3.7926	6.4275×10^{-4}	29219	7363

beneficial to choose a relatively small β, since the required Monte Carlo sample size depends only logarithmically on β (i.e., required sample size is of $O(\log \frac{1}{\beta})$) [10]. However, as the confidence level $1-\alpha$ grows larger, the size of uncertainty $\mathcal{U}_{data}$ would grow and cause the tolerance level ϵ for the SO (under the baseline $\mathbb{P}_0$) to decrease. On the other hand, the dependence of Monte Carlo sample size on ϵ is less favorable, typically of $O(\frac{1}{\epsilon})$ [10].

Algorithm 1 *Extended SO* to obtain a feasible solution $\hat{x}$ for (1.1) with asymptotic confidence $1-\alpha-\beta$

1: **Inputs:** data points $\xi_1, \ldots, \xi_n$, a ϕ-divergence, parametric information $\mathcal{P}_{para} = \{\mathbb{P}_\theta\}_{\theta \in \Theta \subset \mathbb{R}^D}$.
2: Find the MLE $\hat{\theta}$ from the data $\xi_1, \ldots, \xi_n$ for parameter θ.
3: Set $\lambda \leftarrow \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n}$ where $\chi^2_{1-\alpha,D}$ is the $1-\alpha$ quantile of a χ^2_D distribution.
4: Set $\delta_\epsilon \leftarrow \epsilon'(\epsilon, \lambda, \phi)$ where ϵ' is defined in (3.10).
5: Set $N \leftarrow N_{exact}(\delta_\epsilon, \beta, d)$ where N_{exact} is defined in (2.5).
6: Generate $\xi_1^{MC}, ..., \xi_N^{MC}$ from $\mathbb{P}_{\hat{\theta}}$ to construct (2.4) and obtain a solution $\hat{x}$.

There are several variants of Algorithm 1. First, we have discussed the use of plain SO and that the required sample size is (2.5), while on the other hand, as mentioned at the end of Section 2.2, we can use other variants of SO such as FAST that requires a smaller sample size for either the data or the Monte Carlo samples we generate. In the case of FAST, we would have $N_{exact}(\epsilon, \beta, d) = 20d + \frac{1}{\epsilon}\log\frac{1}{\beta}$, as suggested by [13]. Thus, a variant of Algorithm 1 is to replace N_{exact} with this latter quantity, and replace (2.4) with the FAST procedure in [13] for the last step of Algorithm 1 (we call this algorithm "Extended FAST" which will also be used in the next section).

The explicit expression for $\epsilon'(\epsilon, \lambda, \phi)$ for different ϕ, ϵ and λ can be found in [40]. For example, if we choose $\phi = (x-1)^2$ which corresponds to the χ^2-distance, then for $\epsilon < 1/2$, we have $\epsilon' = \max\{0, \epsilon - \frac{\sqrt{\lambda^2+4\lambda(\epsilon-\epsilon^2)}-(1-2\epsilon)\lambda}{2\lambda+2}\}$. We can also replace $\epsilon'(\epsilon, \lambda, \phi)$ by any δ_ϵ that achieves $M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta_\epsilon) \le \epsilon$. In Appendix B, we derive a self-contained easy upper bound for $M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta)$ in the case of χ^2-distance and use it to find such a δ_ϵ. This easy computation of δ_ϵ will also be used in our numerics in the next section.

Section 4.1 has investigated some proposals to improve the generating distributions. Algorithm 2 depicts these proposals in a general form. The main difference of Algorithm 2 compared to Algorithm 1 is the introduction of $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ that one can attempt to minimize over a class of generating distribution $\mathbb{P}_0$ or evaluate for trial-and-error choices of $\mathbb{P}_0$, so that at the end we have $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \phi''(1)\chi^2_{1-\alpha,D}/(2n)$. As discussed in Section 4.1, using this $\mathbb{P}_0$ allows us to obtain a smaller Monte Carlo size requirement than simple relaxation of the parametric constraint. Sections 4.3 and 4.4 describe the possibilities of achieving such a reduction, in the case of Gaussian underlying distributions and using χ^2-distance. Note that, just like in Algorithm 1, we can consider

other variants such as incorporating FAST and using alternate bounds for M instead of ϵ', by undertaking the same modifications as in Algorithm 1.

Algorithm 2 *Extended SO with improved generating distribution* to obtain a feasible solution $\hat{x}$ for (1.1) with asymptotic confidence $1-\alpha-\beta$

1: **Inputs:** data points $\xi_1, \ldots, \xi_n$, a ϕ-divergence, parametric information $\mathcal{P}_{para} = \{\mathbb{P}_\theta\}_{\theta\in\Theta\subset\mathbb{R}^D}$.
2: Find the MLE $\hat{\theta}$ from the data $\xi_1, \ldots, \xi_n$ for parameter θ.
3: Set $\lambda \leftarrow \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n}$ where $\chi^2_{1-\alpha,D}$ is the $1-\alpha$ quantile of a χ^2_D distribution.
4: Obtain $\mathbb{P}_0$ by minimizing $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ defined in (4.1) over a class of distributions or simple trial-and-error search so that $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \lambda$.
5: Set $\delta_\epsilon \leftarrow \epsilon'(\epsilon, \mathcal{D}_{data}(\mathbb{P}_0, \phi), \phi)$ where ϵ' is defined in (3.10).
6: Set $N \leftarrow N_{exact}(\delta_\epsilon, \beta, d)$ where N_{exact} is defined in (2.5).
7: Generate $\xi_1^{MC}, ..., \xi_N^{MC}$ from $\mathbb{P}_0$ to construct (2.4) and obtain a solution $\hat{x}$.

6 NUMERICAL EXPERIMENTS

This section presents some numerical examples to support our theoretical findings and illustrate the performance of our proposed procedures for data-driven CCPs. We focus on Algorithm 1 (Extended SO) and its FAST variant discussed in Section 5 (Extended FAST). We consider both single and joint CCPs (i.e., one and multiple inequalities respectively in the safety condition of the probability) as well as quadratic optimization problems. Moreover, we compare numerically with methods of robust optimization (RO) in [2, 54]. The experimental outputs that we report include:

- Under each setting, we repeat the experiment 1000 times with new data generated each time. For the solution $\hat{x}$ obtained in each trial from a given algorithm, we evaluate the violation probability $V(\hat{x}, \mathbb{P})$ under the true probability measure $\mathbb{P}$ (under θ_{true}) either through exact calculation or Monte Carlo simulation with sample size 10000. Moreover, using the empirical distribution for the violation probabilities, we report $\hat{\epsilon}$ as the average violation probability $V(\hat{x}, \mathbb{P})$ as well as Q_{95}, the 95-percentile. Finally, we report and compare "f_{val}", the average objective value for the optimization problem across all 1000 runs.
- We fix $\alpha = 0.05$ and $\beta = 0.01$ across different values of ϵ and d. However, when we compare our methods with robust optimization approaches, we set $\alpha = 0.05$ and $\beta = 0.001$, since RO approaches essentially guarantee $\beta = 0$. On the other hand, the sample size chosen for FAST is taken with default values $N_1 = 20d$ in stage 1 and $N_2 = \frac{\log\beta - \log(B_\epsilon^{N_1,d})}{\log(1-\epsilon)}$ in stage 2 as discussed in [13].
- For given ϵ and d, we denote N_{exact} as the required sample size if we can directly sample from $\mathbb{P}$ and use standard SO. We denote n as the available data size ($n < N_{exact}$) and N as the Monte Carlo size needed for the our DRO-based methods. In DRO-based methods, we fix our generating distribution $\mathbb{P}_0$ as $\mathbb{P}_{\hat{\theta}}$ and use the χ^2-distance across the experiments.

6.1 Single Linear Chance Constraint Problem

We first consider a single linear CCP

$$\begin{aligned} \min_{x\in\mathcal{X}\subseteq\mathbb{R}^d} \quad & c^T x \\ \text{s.t.} \quad & \mathbb{P}((a+\xi)^T x \le b) \ge 1-\epsilon, x \ge 0 \end{aligned} \tag{6.1}$$

where $x \in \mathbb{R}^d$ is the decision variable, $a, c \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are fixed and $\xi \in \mathbb{R}^d$ is a random vector following some parametric distribution. We fix $a = [5, 5, ..., 5] \in \mathbb{R}^d$, $b = 5$ and $c = [-1, -1, ..., -1] \in$

$\mathbb{R}^d$ and the problem would have a non-empty feasible region with high probability for ξ considered here. Moreover, a robustly feasible point for FAST [13] is chosen to be $\bar{x} = \mathbf{0} \in \mathbb{R}^d$ and an explicit $\mathcal{U}_{data}$ is constructed as (2.21) for our DRO.

6.1.1 Multivariate Gaussian . We conduct experiments when $\xi \sim \mathcal{N}(\theta, \Sigma)$ with fixed but a priori randomly generated positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and unknown $\theta \in \mathbb{R}^d$. Due to the normality of ξ, for any given θ, we can reformulate the chance constraint exactly as a second-order cone constraint, which can be robustified straightforwardly in the ambiguous chance constraint case. The underlying true parameter is taken to be $\theta_{true} = \mathbf{0} \in \mathbb{R}^d$ and the results are summarized in Table 3 and 4.

Table 3. Single linear CCP under Gaussian with unknown mean for different ϵ and d.

	$\epsilon = 0.1$ $d = 5$	$\epsilon = 0.1$ $d = 10$	$\epsilon = 0.1$ $d = 20$	$\epsilon = 0.05$ $d = 5$	$\epsilon = 0.05$ $d = 10$	$\epsilon = 0.05$ $d = 20$
n	50	80	200	50	80	200
N_{exact}	113	183	312	229	371	631
N	449	743	1016	1443	2349	3118
$\hat{\epsilon}$	0.0050	0.0041	0.0041	0.0015	0.0015	0.0014
Q_{95}	0.0136	0.0103	0.0088	0.0045	0.0037	0.0031
f_{val}	-0.7577	-0.7447	-0.7360	-0.7353	-0.7243	-0.7128

Table 4. Comparisons for single linear CCP under Gaussian: $\epsilon = 0.05$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	80	80	80
N_{exact}	NA	447	447
N	NA	2887	1079
$\hat{\epsilon}$	0.0180	0.0011	0.00069
Q_{95}	0.0272	0.0029	0.0019
f_{val}	-0.8008	-0.7212	-0.7093

6.1.2 Exponential Distribution. We conduct experiment when each coordinate ξ_i of $\xi \in \mathbb{R}^d$ independently follows exponential distribution with rate λ_i. Since ξ is no longer Gaussian and the domain of the moment generating moment function for exponential distribution depends on $\lambda = (\lambda_1, \ldots, \lambda_d)$, for convenience we use RO constructed from a convex approximation using Chebyshev's inequality:

$$\mathbb{P}_\lambda\Big(\xi^T x - \sum_{i=1}^d \frac{x_i}{\lambda_i} > \epsilon^{-1/2}\sqrt{Var(\xi^T x)}\Big) \le \epsilon$$

which, combined with $\mathcal{U}_{data}$ as in (2.21), reduces the ambiguous chance constraint into a robust conic quadratic constraint

$$\epsilon^{-1/2}\sqrt{\sum_{i=1}^d (\frac{x_i}{\lambda_i})^2} + a^T x + \epsilon^{-1/2}\sum_{i=1}^d \frac{x_i}{\lambda_i} - b \le 0, \quad \forall \lambda : \sum_{i=1}^d (1 - \frac{\lambda_i}{\hat{\lambda}_i})^2 \le \frac{\chi^2_{1-\alpha,d}}{n},$$

The above can be tractably reformulated as in Section 5 of [54] on problems in the form of 5(b), with $\Omega = (\min_i(\hat{\lambda}_i)(1 - \frac{\chi^2_{1-\alpha,d}}{n}))^{-1}$ where $\hat{\lambda}_i$ represents the MLE estimate. Finally, the underlying true parameters are taken as $\lambda_i = 1, \forall i$, and results are summarized in Table 5.

Table 5. Comparisons for single linear CCP under Exponential: $\epsilon = 0.05$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	80	80	80
N_{exact}	NA	447	447
N	NA	2887	1079
$\hat{\epsilon}$	0.0045	0.0047	0.0016
Q_{95}	0.0094	0.0100	0.0050
f_{val}	-0.6978	-0.6981	-0.6701

From the results of the experiments, we can see the vast majority of solutions produced by three methods satisfy statistical feasibility. In fact, all methods are conservative with respect to the violation probability ϵ, although some are more conservative than the other. In particular, when ξ is Gaussian, RO takes advantage of an exact formulation to produce less conservative solution with lower objective value (closer to the optimal value). This can be seen in Table 4, where $\hat{\epsilon} = 0.018$ $f_{val} = -0.80$ for RO and $\hat{\epsilon} = 0.0011$ $f_{val} = -0.72$ only for Extended SO. When ξ is no longer Gaussian, RO appears to produce similar-quality solutions as Extended SO in terms of feasibility or optimality. For example in Table 5, we have $\hat{\epsilon} = 0.0045$ $f_{val} = -0.6978$ for RO and $\hat{\epsilon} = 0.0047$ $f_{val} = -0.6981$ for Extended SO. Note that while the validity of RO depends crucially on the applicability and accuracy of convex approximation, the validity of Extended SO or Extended FAST is not restricted by the distributions of ξ, and they also do not require intensive, case-specific analysis as RO. In general, we observe consistent performances of our methods in both experiments.

6.2 Joint Linear Chance Constraint Problem

Next, we consider a joint chance-constrained linear problem:

$$\begin{aligned} \min_{x \in X \subseteq \mathbb{R}^d} \quad & c^T x \\ \text{s.t.} \quad & \mathbb{P}((A + \Xi)x \leq b) \geq 1 - \epsilon, x \geq 0 \end{aligned} \tag{6.2}$$

where $x \in \mathbb{R}^d$ is the decision variable, $A \in \mathbb{R}^{m \times d}$, $c \in \mathbb{R}^d$ and $b \in \mathbb{R}^m$ are fixed and $\Xi \in \mathbb{R}^{m \times d}$ is a random matrix following some parametric distribution. We set c, each row of A and b to be the same as in the single linear CCP. We treat $\Xi \in \mathbb{R}^{m \times d}$ as a matrix concatenated from a random vector $\xi \in \mathbb{R}^{md} \sim \mathcal{N}(\theta, \Sigma)$ with fixed but a priori randomly generated positive definite covariance matrix $\Sigma \in \mathbb{R}^{md \times md}$ and unknown $\theta \in \mathbb{R}^{md}$. To solve RO, we use Bonferroni's inequality as in [57] to first divide the violation probability ϵ uniformly across m individual chance constraints and then follow the procedure as in single linear CCP. The results are summarized in Table 6.

In this joint linear example, Extended FAST provides the least conservative solution in terms of the achieved tolerance level ($\hat{\epsilon} = 0.0226$, which is closer to 0.05, compared to 0.0003 in RO and 0.0012 in Extended SO), and Extended SO is the least conservative in terms of the objective value ($f_{val} = -0.6626$ compared to -0.6448 in RO and -0.6466 in Extended FAST). RO appears to be the most conservative in terms of both the achieved tolerance level and objective value. Note that this occurs even though the underlying randomness is Gaussian, which allows exact reformulation in

Table 6. Comparisons for Joint linear CCP under Gaussian: $\epsilon = 0.05$, $m = 3$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	80	80	80
N_{exact}	NA	291	291
N	NA	2388	1214
$\hat{\epsilon}$	0.0003	0.0012	0.0226
Q_{95}	0.0007	0.0033	0.0564
f_{val}	-0.6448	-0.6626	-0.6466

the single chance constraint case. The conservative performance here is likely (and unsurprisingly) due to the crude Bonferroni's correction. Note that other alternatives to using Bonferroni, if one considers tractable reformulation, would be to use moment-based DRO where tractability can be achieved (e.g., [69]). However, it is unclear if using moment-based DRO would be more or less conservative than using Bonferroni correction along with exact reformulation for the individualized constraints, which could comprise an interesting comparison for a future study. Nonetheless, our Extended SO/FAST, being purely sampled-based, avoids the additional conservativeness coming from the Bonferroni correction. However, we note that a large number of Monte Carlo samples are required due to the large size of $\mathcal{U}_{data}$ in this high-dimensional problem.

6.3 Non-Linear Chance Constrained Problems

In this section, we conduct numerical experiments for non-linear CCP. We consider two examples. First is a quadratic objective with joint linear chance constraints, and second is a linear objective with a quadratic chance constraint, similar as the QM problem considered in [49].

6.3.1 Quadratic Objective with Joint Linear Chance Constraint. We adopt the same setup (thus the robust feasibility condition remains the same) as in (6.2) except we modify the objective with a quadratic term

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & \frac{1}{2} x^T H x + c^T x \\ \text{s.t.} \quad & \mathbb{P}((A + \Xi)x \leq b) \geq 1 - \epsilon, x \geq 0 \end{aligned} \tag{6.3}$$

for a fixed but a priori randomly generated positive definite matrix H. We use $\epsilon = 0.05$. Results are summarized in Table 7. As we can see, feasibility in terms of violation probability is satisfied by all methods, though RO suffers from higher conservativeness compared to Extended SO/FAST in terms of the objective value ($f_{val} = -0.48$ compared to -0.5547 and -0.5476 for Extended SO and FAST respectively). Like the previous example, this could be attributed to the Bonferroni correction used in the RO. Extended FAST gives the least conservative solution in terms of the tolerance level ($\hat{\epsilon} = 0.0096$), using only one third of the samples compared to Extended SO (3888 vs 1384). On the other hand, Extended SO gives the least conservative solution in terms of the objective value ($f_{val} = -0.5547$).

6.3.2 Linear Objective with Quadratic Chance Constraint. We consider the following setup:

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & c^T x \\ \text{s.t.} \quad & \mathbb{P}(x^T \Xi x + a^T x \leq b) \geq 1 - \epsilon, x \geq 0 \end{aligned} \tag{6.4}$$

Table 7. Comparisons for quadratic objective with joint linear chance constraint under Gaussian: $\epsilon = 0.05$, $m = 5$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	200	200	200
N_{exact}	NA	447	447
N	NA	3888	1384
$\hat{\epsilon}$	0	0.0006	0.0096
Q_{95}	0	0.0017	0.0253
f_{val}	-0.4800	-0.5547	-0.5476

We set $\Xi = \frac{1}{m}\sum_{i=1}^{m} \xi_i \xi_i^T$ and $\xi_i \in \mathbb{R}^d \sim \mathcal{N}(\theta, \Sigma)$ are i.i.d. with unknown θ. We set $\theta_{true} = 0 \in \mathbb{R}^d$ and consequently $m\Xi$ follows a Wishart distribution $\mathcal{W}(\Sigma, m)$ with m degrees of freedom and covariance matrix Σ under $\mathbb{P}$. We use $\epsilon = 0.05$. The RO formulation for this problem is not readily available while our sampling-based methods are still directly applicable. We thus focus on evaluating the performance of Extended FAST under different hyper-parameters. The results are summarized in Table 8. As we can see, the high dimensions of the problem do not affect the sample size requirement of Extended FAST dramatically, as it increases moderately form $N = 154$ when $d = 5$ to $N = 334$ when $d = 10$ and to $N = 422$ when $d = 15$. Moreover, the average optimal value f_{val} is around -0.85 and feasibility is satisfied ($\hat{\epsilon}$ all within 0.05), showing the consistent effectiveness of our method.

Table 8. Linear objective with quadratic chance constraint for different ϵ, m and d.

	$\epsilon = 0.1, d = 5, m = 5$	$\epsilon = 0.05, d = 10, m = 10$	$\epsilon = 0.05, d = 15, m = 15$
n	80	200	300
N_{exact}	113	371	504
N	154	334	422
$\hat{\epsilon}$	0.0092	0.0050	0.0048
Q_{95}	0.0263	0.0133	0.0128
f_{val}	-0.8393	-0.8576	-0.8672

7 CONCLUSION

We consider data-driven chance constrained problems with limited data. In such situation, standard approaches in SO may not be able to generate statistically feasible solutions. We investigate an approach that uses divergence-based DRO to efficiently incorporate parametric information through a data-driven uncertainty set, and subsequently uses Monte Carlo sampling to generate enough samples to handle the distributionally robust chance constraint. In this way our framework translates the data size requirement in SO into a Monte Carlo requirement, the latter could be much more abundant thanks to cheap modern computational power.

To exploit the full capability of our framework, we have investigated the optimality of the generating distribution in drawing the Monte Carlo samples in the sense of minimizing its required sample size. We have shown that, while the optimal choice is the baseline distribution in the unambiguous and nonparametric DRO cases, this natural choice can be dominated by other

distributions in the parametric DRO case. We proved this by connecting the Neyman-Pearson lemma in statistical hypothesis testing to DRO and SO, which comprises the first such results of its kind as far as we know. We then studied several ways to find better generating distributions by searching for mixtures that enhance distributional variability. Lastly, we showed some numerical results to demonstrate how our approach can give rise to feasible solutions in a wide range of settings where other methods such as RO cannot be utilized directly or give more conservative solutions.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020, CAREER CMMI-1653339/1834710 and IIS-1849280.

REFERENCES

[1] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science* 59, 2 (2013), 341–357.
[2] A. Ben-Tal, L. El Ghaoui, and A.S. Nemirovski. 2009. *Robust Optimization.* Princeton University Press.
[3] Dimitri P Bertsekas. 1971. *Control of Uncertain Systems with a Set-Membership Description of the Uncertainty.* Ph.D. Dissertation. Massachusetts Institute of Technology.
[4] Dimitri P Bertsekas, Angelia Nedi, Asuman E Ozdaglar, et al. 2003. *Convex Analysis and Optimization.* Athena Scientific.
[5] Jose Blanchet, Yang Kang, and Karthyek Murthy. 2016. Robust Wasserstein Profile Inference and Applications to Machine Learning. *arXiv preprint arXiv:1610.05627* (2016).
[6] Jose Blanchet and Karthyek RA Murthy. 2016. Quantifying Distributional Model Risk via Optimal Transport. *arXiv preprint arXiv:1604.01446* (2016).
[7] Giuseppe C Calafiore. 2017. Repetitive Scenario Design. *IEEE Trans. Automat. Control* 62, 3 (2017), 1125–1137.
[8] Giuseppe C Calafiore, Fabrizio Dabbene, and Roberto Tempo. 2011. Research on Probabilistic Methods for Control System Design. *Automatica* 47, 7 (2011), 1279–1293.
[9] Marco C Campi and Algo Carè. 2013. Random Convex Programs with L_1-Regularization: Sparsity and Generalization. *SIAM Journal on Control and Optimization* 51, 5 (2013), 3532–3557.
[10] Marco C Campi and Simone Garatti. 2008. The Exact Feasibility of Randomized Solutions of Uncertain Convex Programs. *SIAM Journal on Optimization* 19, 3 (2008), 1211–1230.
[11] Marco C Campi and Simone Garatti. 2011. A Sampling-and-Discarding Approach to Chance-Constrained Optimization: Feasibility and Optimality. *Journal of Optimization Theory and Applications* 148, 2 (2011), 257–280.
[12] Marco C Campi and Simone Garatti. 2018. Wait-and-Judge Scenario Optimization. *Mathematical Programming* 167, 1 (2018), 155–189.
[13] Algo Carè, Simone Garatti, and Marco C Campi. 2014. FAST-- Fast Algorithm for the Scenario Technique. *Operations Research* 62, 3 (2014), 662–671.
[14] Mohammadreza Chamanbaz, Fabrizio Dabbene, Roberto Tempo, Venkatakrishnan Venkataramanan, and Qing-Guo Wang. 2016. Sequential Randomized Algorithms for Convex Optimization in the Presence of Uncertainty. *IEEE Trans. Automat. Control* 61, 9 (2016), 2565–2571.
[15] Wenqing Chen, Melvyn Sim, Jie Sun, and Chung-Piaw Teo. 2010. From CVaR to Uncertainty Set: Implications in Joint Chance-Constrained Optimization. *Operations research* 58, 2 (2010), 470–485.
[16] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. 2018. Data-Driven Chance Constrained Programs over Wasserstein Balls. *arXiv preprint arXiv:1809.00210* (2018).
[17] Jianqiang Cheng, Erick Delage, and Abdel Lisser. 2014. Distributionally Robust Stochastic Knapsack Problem. *SIAM Journal on Optimization* 24, 3 (2014), 1485–1506.
[18] Frank H Clarke. 1975. Generalized Gradients and Applications. *Trans. Amer. Math. Soc.* 205 (1975), 247–262.
[19] AB Olde Daalhuis. 2010. Confluent Hypergeometric Functions. *NIST Handbook of Mathematical Functions, FWJ Olver, DW Lozier, RF Boisvert, and CW Clark, eds., Cambridge University, New York* (2010), 321–349.
[20] Daniela Pucci De Farias and Benjamin Van Roy. 2004. On Constraint Sampling in the Linear Programming Approach to Approximate Dynamic Programming. *Mathematics of Operations Research* 29, 3 (2004), 462–478.
[21] Erick Delage and Yinyu Ye. 2010. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research* 58, 3 (2010), 595–612.
[22] John Duchi, Peter Glynn, and Hongseok Namkoong. 2016. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *arXiv preprint arXiv:1610.03425* (2016).

[23] Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Petr Plechác. 2016. Path-Space Information Bounds for Uncertainty Quantification and Sensitivity Analysis of Stochastic Dynamics. *SIAM/ASA Journal on Uncertainty Quantification* 4, 1 (2016), 80–111.
[24] E Erdoğan and Garud Iyengar. 2006. Ambiguous Chance Constrained Problems and Robust Optimization. *Mathematical Programming* 107, 1-2 (2006), 37–61.
[25] Peyman Mohajerin Esfahani and Daniel Kuhn. 2015. Data-Driven Distributionally Robust Optimization using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. *Mathematical Programming* (2015), 1–52.
[26] Gerald B Folland. 2013. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons.
[27] Rui Gao and Anton J Kleywegt. 2016. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv preprint arXiv:1604.02199* (2016).
[28] Soumyadip Ghosh and Henry Lam. 2019. Robust Analysis in Stochastic Simulation: Computation and Performance Guarantees. *Operations Research* 67, 1 (2019), 232–249.
[29] Paul Glasserman and Xingbo Xu. 2014. Robust Risk Measurement and Model Risk. *Quantitative Finance* 14, 1 (2014), 29–58.
[30] Joel Goh and Melvyn Sim. 2010. Distributionally Robust Optimization and Its Tractable Approximations. *Operations research* 58, 4-part-1 (2010), 902–917.
[31] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. 2018. Robust Empirical Optimization Is Almost The Same As Mean–Variance Optimization. *Operations Research Letters* 46, 4 (2018), 448–452.
[32] Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. 2015. A Distributionally Robust Perspective on Uncertainty Quantification and Chance Constrained Programming. *Mathematical Programming* 151, 1 (2015), 35–62.
[33] Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. 2017. Ambiguous Joint Chance Constraints under Mean and Dispersion Information. *Operations Research* 65, 3 (2017), 751–767.
[34] Lars Peter Hansen and Thomas J Sargent. 2008. *Robustness*. Princeton university press.
[35] L Jeff Hong, Jun Luo, and Barry L Nelson. 2015. Chance Constrained Selection of the Best. *INFORMS Journal on Computing* 27, 2 (2015), 317–334.
[36] Zhaolin Hu, Jing Cao, and L Jeff Hong. 2012. Robust Simulation of Global Warming Policies using the DICE Model. *Management Science* 58, 12 (2012), 2190–2206.
[37] Zhaolin Hu and L Jeff Hong. 2013. Kullback-Leibler Divergence Constrained Distributionally Robust Optimization. *Available at Optimization Online* (2013).
[38] Zhaolin Hu and L. Jeff Hong. 2015. Robust Simulation of Stochastic Systems with Input Uncertainties Modeled by Statistical Divergences. In *Proceedings of the 2015 Winter Simulation Conference*, L. Yilmaz et al. (Ed.). IEEE, Piscataway, New Jersey, 643–654.
[39] Ran Ji and Miguel Lejeune. 2018. Data-Driven Distributionally Robust Chance-Constrained Optimization with Wasserstein Metric. *Available at SSRN 3201356* (2018).
[40] Ruiwei Jiang and Yongpei Guan. 2016. Data-Driven Chance Constrained Stochastic Program. *Mathematical Programming* 158, 1-2 (2016), 291–327.
[41] Aleksandr Petrovich Korostelev and Olga Korosteleva. 2011. *Mathematical Statistics: Asymptotic Minimax Theory*. American Mathematical Society, Providence, Rhode Island.
[42] Henry Lam. 2016. Robust Sensitivity Analysis for Stochastic Systems. *Mathematics of Operations Research* 41, 4 (2016), 1248–1275.
[43] Henry Lam. 2018. Sensitivity to Serial Dependency of Input Processes: A Robust Approach. *Management Science* 64, 3 (2018), 1311–1327.
[44] Henry Lam. 2019. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research* 67, 4 (2019), 1090–1105.
[45] Henry Lam and Fengpei Li. 2018. Sampling Uncertain Constraints Under Parametric Distributions. In *2018 Winter Simulation Conference (WSC)*. IEEE, 2072–2083.
[46] Henry Lam and Enlu Zhou. 2017. The Empirical Likelihood Approach to Quantifying Uncertainty in Sample Average Approximation. *Operations Research Letters* 45, 4 (2017), 301–307.
[47] Erich Leo Lehmann. 2004. *Elements of Large-sample Theory*. Springer Science & Business Media.
[48] Erich L Lehmann and Joseph P Romano. 2006. *Testing Statistical Hypotheses*. Springer Science & Business Media.
[49] Miguel A Lejeune and François Margot. 2016. Solving Chance-Constrained Optimization Problems with Stochastic Quadratic Inequalities. *Operations Research* 64, 4 (2016), 939–957.
[50] Miguel A Lejeune and Andrzej Ruszczynski. 2007. An Efficient Trajectory Method for Probabilistic Production-Inventory-Distribution Problems. *Operations Research* 55, 2 (2007), 378–394.
[51] Bowen Li, Ruiwei Jiang, and Johanna L Mathieu. 2019. Ambiguous Risk Constraints with Moment and Unimodality Information. *Mathematical Programming* 173, 1-2 (2019), 151–192.

[52] D Love and Güzin Bayraksan. 2015. *Phi-Divergence Constrained Ambiguous Stochastic Programs for Data-Driven Optimization.* Technical Report. Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio.
[53] James Luedtke and Shabbir Ahmed. 2008. A Sample Approximation Approach for Optimization with Probabilistic Constraints. *SIAM Journal on Optimization* 19, 2 (2008), 674–699.
[54] Ahmadreza Marandi, Aharon Ben-Tal, Dick den Hertog, and Bertrand Melenberg. 2017. Extending the Scope of Robust Quadratic Optimization. *Available on Optimization Online* (2017).
[55] Michael R Murr and András Prékopa. 2000. Solution of A Product Substitution Problem Using Stochastic Programming. In *Probabilistic Constrained Optimization*, U. Stanislav (Ed.). Springer, Manhattan, New York, 252–271.
[56] Arkadi Nemirovski and Alexander Shapiro. 2006. Scenario Approximations of Chance Constraints. In *Probabilistic and randomized methods for design under uncertainty.* Springer, 3–47.
[57] Arkadi Nemirovski and Alexander Shapiro. 2007. Convex Approximations Of Chance Constrained Programs. *SIAM Journal on Optimization* 17, 4 (2007), 969–996.
[58] Frank Nielsen and Richard Nock. 2014. On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences. *IEEE Signal Processing Letters* 21, 1 (2014), 10–13.
[59] Leandro Pardo. 2005. *Statistical Inference Based on Divergence Measures.* Chapman and Hall/CRC, New York.
[60] Ian R Petersen, Matthew R James, and Paul Dupuis. 2000. Minimax Optimal Control of Stochastic Uncertain Systems with Relative Entropy Constraints. *IEEE Trans. Automat. Control* 45, 3 (2000), 398–412.
[61] András Prékopa. 2003. Probabilistic Programming. In *Handbooks in Operations Research and Management Science, Volume 10: Stochastic Programming*, A. Ruszczynski and A. Shapiro (Eds.). Elsevier, Amsterdam, Netherlands.
[62] András Prékopa, Tamás Rapcsák, and István Zsuffa. 1978. Serially Linked Reservoir System Design Using Stochastic Programing. *Water Resources Research* 14, 4 (1978), 672–678.
[63] András Prékopa and Tamás Szántai. 1978. Flood Control Reservoir System Design Using Stochastic Programming. In *Mathematical Programming in Use*, M.L. Balinski and C. Lemarechal (Eds.). Springer, Manhattan, New York, 138–151.
[64] Georg Schildbach, Lorenzo Fagiano, and Manfred Morari. 2013. Randomized Solutions to Convex Programs with Multiple Chance Constraints. *SIAM Journal on Optimization* 23, 4 (2013), 2479–2501.
[65] Yuanming Shi, Jun Zhang, and Khaled B Letaief. 2015. Optimal Stochastic Coordinated Beamforming for Wireless Cooperative Networks with CSI Uncertainty. *IEEE Transactions on Signal Processing* 63, 4 (2015), 960–973.
[66] Aad W Van der Vaart. 2000. *Asymptotic Statistics.* Vol. 3. Cambridge University Press.
[67] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. 2014. Distributionally Robust Convex Optimization. *Operations Research* 62, 6 (2014), 1358–1376.
[68] Andreas Winkelbauer. 2012. Moments and Absolute Moments of the Normal Distribution. *arXiv preprint arXiv:1209.4340* (2012).
[69] Weijun Xie and Shabbir Ahmed. 2018. On Deterministic Reformulations of Distributionally Robust Joint Chance Constrained Optimization Problems. *SIAM Journal on Optimization* 28, 2 (2018), 1151–1182.
[70] Yiling Zhang, Ruiwei Jiang, and Siqian Shen. 2016. Ambiguous Chance-Constrained Bin Packing under Mean-Covariance Information. *arXiv preprint arXiv:1610.00035* (2016).
[71] Steve Zymler, Daniel Kuhn, and Berç Rustem. 2013. Distributionally Robust Joint Chance Constraints with Second-Order Moment Information. *Mathematical Programming* (2013), 1–32.

A REGULAR CONDITIONS FOR VERIFYING ASSUMPTION A1

We consider the following conditions:

(C1) $p(x, \theta_1) = p(x, \theta_2)$ for all x implies $\theta_1 = \theta_2$.

(C2) θ_{true} is an inner point of $\Theta \subseteq \mathbb{R}^D$.

(C3) The support of distribution $\{x : p(x, \theta) > 0\}$ does not depend on θ.

(C4) There exists a measurable function $L_1(x)$ such that $\mathbb{E}_{\theta_{true}} L_1^2 < \infty$ and

$$|\log p(x, \theta_1) - \log p(x, \theta_2)| \leq L_1(x)\|\theta_1 - \theta_2\|_2 \tag{A.1}$$

for all θ_1, θ_2 in a neighborhood of θ_{true}.

(C5) $I(\theta_{true})$ is non-singular.

(C6) The density family $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is differentiable in quadratic mean at θ_{true}, i.e., there exists a measurable function $L_2(x) : \mathcal{X} \to R^D$ such that for any $h \in \mathbb{R}^D$ that converges to 0,

$$\int \left(\sqrt{p(x, \theta_{true} + h)} - \sqrt{p(x, \theta_{true})} - \frac{1}{2} h^T L_2(x) \sqrt{p(x, \theta_{true})}\right)^2 dx = o(\|h\|_2^2). \tag{A.2}$$

The consistency and asymptotic normality of MLE in Assumption A1 is guaranteed under conditions (C1)-(C6). See [47, 66].

B ALTERNATE BOUNDS USING χ^2 DISTANCE

Consider the χ^2-based uncertainty set $\mathcal{U}_{data} = \{\mathbb{Q} \in \mathcal{P}_{para} : \chi^2(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}$. Here we provide an alternate upper bound for the function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$, which we call $\tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$. That is, we find $\tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ that satisfies

$$\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(\xi \in A) \leq \tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta), \quad \text{for all } A \text{ such that } \mathbb{P}_0(A) \leq \delta.$$

For any $\mathbb{Q}$ absolutely continuous with respect to $\mathbb{P}_0$, we have

$$\begin{aligned}
\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(\xi \in A) &= \mathbb{P}_0(\xi \in A) + \Big(\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(\xi \in A) - \mathbb{P}_0(\xi \in A) \Big) \\
&= \mathbb{P}_0(\xi \in A) + \sup_{\mathbb{Q}\in\mathcal{U}_{data}} \int \mathbf{1}\{\xi \in A\} \big(\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1 \big) d\mathbb{P}_0(\xi) \\
&\leq \mathbb{P}_0(\xi \in A) + \sup_{\mathbb{Q}\in\mathcal{U}_{data}} \Big(\int \mathbf{1}\{\xi \in A\} d\mathbb{P}_0(\xi) \Big)^{1/2} \cdot \Big(\int \big(\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1 \big)^2 d\mathbb{P}_0(\xi) \Big)^{1/2} \\
&\leq \delta + \delta^{1/2} \cdot \big(\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \chi^2(\mathbb{P}_0, \mathbb{Q}) \big)^{1/2}, \qquad \text{(B.1)}
\end{aligned}$$

where the fourth line follows from the Cauchy-Schwarz inequality. Thus, we can set

$$\tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta) = \delta + \delta^{1/2} \cdot \big(\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \chi^2(\mathbb{P}_0, \mathbb{Q}) \big)^{1/2} = \delta + \delta^{1/2} \cdot \big(\mathcal{D}_{data}(\mathbb{P}_0) \big)^{1/2},$$

which is non-decreasing in δ. By (2.9), we can choose δ_ϵ such that $\delta_\epsilon + \delta_\epsilon^{1/2} \big(\mathcal{D}_{data}(\mathbb{P}_0) \big)^{1/2} \leq \epsilon$, or equivalently,

$$\delta_\epsilon \leq \epsilon + \frac{\mathcal{D}_{data}(\mathbb{P}_0)}{2} - \sqrt{\epsilon \cdot \mathcal{D}_{data}(\mathbb{P}_0) + \frac{1}{4} \big(\mathcal{D}_{data}(\mathbb{P}_0) \big)^2}, \qquad \text{(B.2)}$$

by solving the quadratic equation. In the case where we relax the parametric constraint completely, we have $\mathcal{D}_{data}(\mathbb{P}_0) = \lambda$. Compared to the bound obtained from Theorem 3.2.1 and Corollary 3.2.2, (B.2) is less tight, but the gap can be shown to asymptotically vanish when $\epsilon, \frac{\chi^2_{1-\alpha,D}}{n} \to 0$.

C PROOFS AND OTHER TECHNICAL RESULTS

Proof of Lemma 4.2.1. First, by definition $\mathcal{P}(\Theta)$ is a convex set and, for any $\mu_1, \mu_2 \in \mathcal{P}(\Theta)$ and $0 < t < 1$, we have

$$(1-t)\mu_1 + t\mu_2 \in \mathcal{P}(\Theta).$$

Next, fixing $\theta \in \mathcal{U}_{data}$, the function $L(\cdot, \theta)$ is convex since:

$$\begin{aligned}
L((1-t)\mu_1 + t\mu_2, \theta) &= \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')((1-t)\mu_1 + t\mu_2)(d\theta')} dy \\
&= \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{(1-t)\int_\Theta p(y;\theta')\mu_1(d\theta') + t\int_\Theta p(y;\theta')\mu_2(d\theta')} dy \\
&\leq (1-t) \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu_1(d\theta')} dy + t \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu_2(d\theta')} dy \\
&= (1-t)L(\mu_1, \theta) + tL(\mu_2, \theta)
\end{aligned}$$

for any $0 < t < 1$ where the inequality follows from the convexity of the function $1/x$ for $x > 0$. Thus, as the supremum of convex functions, $l(\mu) \triangleq \sup_{\theta \in \mathcal{U}_{data}} L(\mu, \theta)$ is also convex. □

We provide a version of Danskin' Theorem needed to prove Theorem 4.3.1. Alternately, one can also resort to a generalized version in [18] by verifying the conditions there. Here we opt for the former and provide a self-contained proof, which mostly relies on the techniques from Proposition 4.5.1 of [4] but with some slight modification to handle issues regarding the domain of the involved function. We have:

Lemma C.0.1. *Fix probability measures $\mu_1, \mu_2 \in \mathcal{P}(\Theta)$. Suppose $t_k \downarrow 0$ is a positive sequence such that $(1-t_k)\mu_1 + t_k\mu_2 \in \mathcal{P}(\Theta)$ for all k and $\theta_k \in \Theta^*((1-t_k)\mu_1 + t_k\mu_2)$ is a sequence such that $\theta_k \to \theta_0$ for some $\theta_0 \in \mathcal{U}_{data}$, then we have*

$$\limsup_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \leq \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t},$$

if we assume $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continuous in $0 \leq t \leq 1$ and $\theta \in \Theta$.

Proof. It is known that if $f : \mathbb{I} \to \mathbb{R}$ is a convex function with $\mathbb{I}$ being an open interval containing some point x, we then have the following results [4]:

$$f^+(x) = \lim_{t\downarrow 0} \frac{f(x+t) - f(x)}{t} = \inf_{t>0} \frac{f(x+t) - f(x)}{t}, \tag{C.1}$$

$$f^-(x) = \lim_{t\downarrow 0} \frac{f(x) - f(x-t)}{t} = \sup_{t>0} \frac{f(x) - f(x-t)}{t}, \tag{C.2}$$

and

$$f^+(x) \geq f^-(x). \tag{C.3}$$

In other words, these limits exist and satisfy the above relations for convex functions. Thus, if we define $f_k(t) = L((1-t_k)\mu_1 + t_k\mu_2 + t(\mu_2 - \mu_1), \theta_k)$, it follows from the convexity of $\mathcal{P}(\Theta)$ and $L(\cdot, \theta_k)$ that $f_k(t)$ is convex and well-defined for $-t_k \leq t \leq 1 - t_k$. Using the above results in (C.1), (C.2) and (C.3), we then have

$$\begin{aligned}\frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} &= \frac{f_k(0) - f_k(-t_k)}{t_k} \\ &\leq \sup_{t>0} \frac{f_k(0) - f_k(-t)}{t} = f_k^-(0) \leq f_k^+(0) = \inf_{t>0} \frac{f_k(t) - f_k(0)}{t}.\end{aligned} \tag{C.4}$$

On the other hand, if we define $f_0(t) = L((1-t)\mu_1 + t\mu_2, \theta_0)$, it also follows that $f_0(t)$ is convex and well-defined for $0 \leq t \leq 1$. It follows from the convexity of $f_0(\cdot)$ as well as (C.1) that

$$\begin{aligned}\lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t} &= \lim_{t\downarrow 0} \frac{f_0(t) - f_0(0)}{t} \\ &= \inf_{t>0} \frac{f_0(t) - f_0(0)}{t} = f_0^+(0).\end{aligned} \tag{C.5}$$

Then, it again follows from the convexity of $f_0(\cdot)$ that, given any $\tau > 0$, we can find some $\eta > 0$ such that

$$\frac{f_0(s) - f_0(0)}{s} \leq f_0^+(0) + \tau, \tag{C.6}$$

for all $0 < s < \eta$. It then follows from definitions and (C.6) that

$$\begin{aligned}\frac{L((1-s)\mu_1 + s\mu_2, \theta_0) - L(\mu_1, \theta_0)}{s} =& \frac{L((\mu_1 + s(\mu_2 - \mu_1), \theta_0) - L(\mu_1, \theta_0)}{s} \\ =& \frac{f_0(s) - f_0(0)}{s} \le f_0^+(0) + \tau, \end{aligned} \tag{C.7}$$

for all $0 < s < \eta$. Fixing one such s, since the function $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continuous in $0 \le t \le 1$ and $\theta \in \Theta$, and the sequence satisfies $\theta_k \to \theta_0$, we have

$$\begin{aligned}\lim_{k\to\infty} \frac{f_k(s) - f_k(0)}{s} =& \lim_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2 + s(\mu_2 - \mu_1), \theta_k) - L((1-t_k)\mu_1 + t_k\mu_2, \theta_k)}{s} \\ =& \frac{L((\mu_1 + s(\mu_2 - \mu_1), \theta_0) - L(\mu_1, \theta_0)}{s} = \frac{f_0(s) - f_0(0)}{s} \le f_0^+(0) + 2\tau,\end{aligned}$$

as long as we make $\eta > s > 0$ small enough so that $\eta \le 1 - t_k$ for all k. Then, for k large enough, we have

$$\inf_{t>0} \frac{f_k(t) - f_k(0)}{t} \le \frac{f_k(s) - f_k(0)}{s} \le f_0^+(0) + 2\tau. \tag{C.8}$$

Combining (C.4), (C.5) and (C.8), we have that, for k large enough,

$$\frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \le \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t} + 2\tau.$$

Finally, since τ is arbitrary, we conclude that

$$\limsup_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \le \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t}.$$

□

We now prove the following version of Danskins' Theorem:

Theorem C.0.2. *Fix $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{U}_{data})$. Suppose $t_k \downarrow 0$ is a positive sequence such that $(1-t_k)\mu_1 + t_k\mu_2 \in \mathcal{P}(\Theta)$ for all k and $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continuous in $0 \le t \le 1$ and $\theta \in \Theta$. Then, if we let $\psi(t) = l((1-t)\mu_1 + t\mu_2)$ for $0 \le t \le 1$ with $l(\cdot) = \sup_{\theta\in\mathcal{U}_{data}} L(\cdot, \theta)$ defined as (4.9), we have*

$$\psi^+(0) = \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t} \tag{C.9}$$

Proof. For any $\theta_0 \in \Theta^*(\mu_1)$ and $\theta_t \in \Theta^*((1-t)\mu_1 + t\mu_2)$, we have

$$\begin{aligned}\frac{\psi(t) - \psi(0)}{t} = \frac{l((1-t)\mu_1 + t\mu_2) - l(\mu_1)}{t} =& \frac{L((1-t)\mu_1 + t\mu_2, \theta_t) - \tilde{L}(\mu_1, \theta_0)}{t} \\ \ge& \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t}.\end{aligned}$$

Thus, by taking $t \downarrow 0$ and taking the supremum over all $\theta_0 \in \Theta^*(\mu_1)$, we have

$$\psi^+(0) \ge \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t}. \tag{C.10}$$

Notice that the existence of the several limits above follows from the convexity of related functions. To prove the reverse inequality, we consider a sequence $\{t_k\}$ with $0 < t_k < 1$ and $t_k \downarrow 0$. Then, we pick another sequence $\{\theta_k\} \subseteq \mathcal{U}_{data}$ with $\theta_k \in \Theta^*((1-t_k)\mu_1 + t_k\mu_2)$ for all k. Since $\mathcal{U}_{data}$ is compact, there exist a subsequence of $\{\theta_k\}$ converge to some $\theta_0 \in \mathcal{U}_{data}$. Without loss of generality,

we drop the subsequence and simply assume $\theta_k \to \theta_0$. We first show $\theta_0 \in \Theta^*(\mu_1)$. To do this, pick any $\tilde{\theta}_0 \in \Theta^*(\mu_1)$. Since $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continous in t and θ, we have

$$L(\mu_1, \theta_0) = \lim_{k\to\infty} L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) \geq \lim_{k\to\infty} L((1-t_k)\mu_1 + t_k\mu_2, \tilde{\theta}_0) = L(\mu_1, \tilde{\theta}_0),$$

where the inequality follows from the definition of θ_k. Now, since $\tilde{\theta}_0 \in \Theta^*(\mu_1)$ and $L(\mu_1, \theta_0) \geq L(\mu_1, \tilde{\theta}_0)$, we must have

$$L(\mu_1, \theta_0) = L(\mu_1, \tilde{\theta}_0) \text{ and } \theta_0 \in \Theta^*(\mu_1).$$

Now, using the definition of $\Theta^*(\mu_1)$, we can write

$$\begin{aligned}\psi^+(0) = \inf_{0<t} \frac{\psi(t)-\psi(0)}{t} \leq \frac{\psi(t_k)-\psi(0)}{t_k} =& \frac{l((1-t_k)\mu_1 + t_k\mu_2) - l(\mu_1)}{t_k} \\ =& \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_0)}{t_k} \\ \leq& \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k}. \end{aligned} \tag{C.11}$$

Now we use Lemma C.0.1 to conclude that

$$\begin{aligned}\psi^+(0) \leq& \limsup_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \\ \leq& \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t} \\ \leq& \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t} \end{aligned} \tag{C.12}$$

Finally, we combine (C.10) and (C.12) to conclude the proof

$$\psi^+(0) = \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t}$$

□

Proof of Theorem 4.3.1. The result can be obtained from Leibniz's integral rule (i.e. differentiation under the integral sign). See, for example, Theorem 2.27 in [26]. □

Next we prove Proposition 4.3.2. For convenience, we note that (4.15) can be written in a compact form for exponential family [58]:

$$p(y;\theta) = e^{\langle t(y),\theta\rangle - F(\theta) + k(y)}, \tag{C.13}$$

where $\langle a, b\rangle = a^\intercal b$ represents the usual inner product in the Euclidean space, and $t(\cdot), F(\cdot)$ and $k(\cdot)$ are known functions. In particular, we have

$$F(\theta) = \frac{\theta^\intercal \Sigma^{-1}\theta}{2} \tag{C.14}$$

To facilitate the calculation, we first introduce two lemmas involving the exponential parametric family based on [58].

Lemma C.0.3. *Pick $\theta_1, \theta_2 \in \Theta$. If $2\theta_2 - \theta_1 \in \Theta$, then we have*

$$\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2}{p(y;\theta_1)} dy = e^{F(2\theta_2-\theta_1) - (2F(\theta_2) - F(\theta_1))}.$$

In particular, if $F(\theta) = \frac{\theta^\intercal\Sigma^{-1}\theta}{2}$, then $\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2}{p(y;\theta_1)} dy = e^{(\theta_2-\theta_1)^\intercal\Sigma^{-1}(\theta_2-\theta_1)}$.

Proof. It follows from (C.13) that

$$\begin{aligned}\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2}{p(y;\theta_1)}dy =& e^{<t(y),2\theta_2-\theta_1>-(2F(\theta_2)-F(\theta_1))+k(y)}dy \\ =& e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))}\cdot\int_{\mathcal{Y}} p(y;2\theta_2-\theta_1)dy \\ =& e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))}.\end{aligned}$$

□

Lemma C.0.4. *Pick θ_1,θ_2 and $\theta_3 \in \Theta$. If $2\theta_2 - 2\theta_1 + \theta_3 \in \Theta$, then we have*

$$\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2 p(y;\theta_3)}{(p(y;\theta_1))^2}dy = e^{F(2\theta_2-2\theta_1+\theta_3)-2F(\theta_2)+2F(\theta_1)-F(\theta_3)}.$$

In particular, if $F(\theta) = \frac{\theta^\intercal\Sigma^{-1}\theta}{2}$, then $\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2 p(y;\theta_3)}{(p(y;\theta_1))^2}dy = e^{(\theta_2-\theta_1)^\intercal\Sigma^{-1}(\theta_2-\theta_1)+2(\theta_2-\theta_1)\Sigma^{-1}(\theta_3-\theta_1)}$.

Proof. The proof follows from the same techniques as in Lemma C.0.3. □

Then (4.16) follows from (2.20), (C.14) and Lemma C.0.3 so that

$$\begin{aligned}\mathcal{U}_{data} \triangleq \left\{\theta\in\Theta : \chi^2(\mathbb{P}_{\hat\theta},\mathbb{P}_\theta) \le \frac{\chi^2_{1-\alpha,D}}{n}\right\} &= \left\{\theta\in\Theta : e^{F(2\theta-\hat\theta)-(2F(\theta)-F(\hat\theta))} - 1 \le \frac{\chi^2_{1-\alpha,D}}{n}\right\} \\ &= \left\{\theta\in\Theta : e^{(\theta-\hat\theta)^\intercal\Sigma^{-1}(\theta-\hat\theta)} - 1 \le \frac{\chi^2_{1-\alpha,D}}{n}\right\} \\ &= \left\{\theta : \hat\theta + \Sigma^{\frac{1}{2}}v, \quad \text{for all } \|v\|_2^2 \le \log(1+\frac{\chi^2_{1-\alpha,D}}{n})\right\},\end{aligned}$$

and (4.17) follows. We now prove Proposition 4.3.2:

Proof of Proposition 4.3.2. Following Theorem 4.3.1, Lemma C.0.3 and Lemma C.0.4, we have

$$\begin{aligned}&\sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\int_{\mathcal{Y}} \frac{(p(y;\theta))^2\cdot\int_\Theta p(y;\theta')(\delta_{\hat\theta}-\mu_{prop})(d\theta')}{(\int_\Theta p(y;\theta')\delta_{\hat\theta}(d\theta'))^2}dy \\ =&\sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\left(\int_{\mathcal{Y}}\frac{(p(y;\theta))^2}{p(y;\hat\theta)}dy - \int_{\mathcal{Y}}\frac{(p(y;\theta))^2\cdot\int_\Theta p(y;\theta')(\mu_{prop})(d\theta')}{(p(y;\hat\theta))^2}dy\right) \\ =&\sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\left(\int_{\mathcal{Y}}\frac{(p(y;\theta))^2}{p(y;\hat\theta)}dy - \int_\Theta\int_{\mathcal{Y}}\frac{(p(y;\theta))^2\cdot p(y;\theta')}{(p(y;\hat\theta))^2}dy\cdot\mu_{prop}(d\theta')\right) \\ =&\sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\left(e^{(\theta-\hat\theta)^\intercal\Sigma^{-1}(\theta-\hat\theta)} - \int_\Theta e^{(\theta-\hat\theta)^\intercal\Sigma^{-1}(\theta-\hat\theta)+2(\theta-\hat\theta)^\intercal\Sigma^{-1}(\theta'-\hat\theta)}\mu_{prop}(d\theta')\right) \\ =&(1+\frac{\chi^2_{1-\alpha,D}}{n})\cdot\sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\left(1-\mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat\theta)^\intercal\Sigma^{-1}(\theta'-\hat\theta)}]\right) \\ =&(1+\frac{\chi^2_{1-\alpha,D}}{n})\cdot\left(1-\inf_{\theta\in\Theta^*(\delta_{\hat\theta})}\mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat\theta)^\intercal\Sigma^{-1}(\theta'-\hat\theta)}]\right).\end{aligned} \tag{C.15}$$

Notice the second equality follows from Fubini's theorem. The third equality follows from Lemma C.0.3 and Lemma C.0.4. The fourth equality follows from (4.17). Now, following the last line (C.15), for the search of descent direction, it is sufficient to prove

$$\inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] > 1.$$

However, since $\mu_{prop}(d\theta')$ is a symmetrical distribution around $\hat{\theta}$, we know that

$$\mathbb{E}_{\theta'\sim\mu_{prop}}[2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})] = 0.$$

for any $\theta \in \Theta^*(\delta_{\hat{\theta}})$. Then, it follows from Jensen's inequality that

$$\inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] \geq 1.$$

Now suppose for the sake of contradiction that

$$\inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] = 1.$$

Then, let $\{\theta_k\}_k \subseteq \Theta^*(\delta_{\hat{\theta}})$ be a subsequence such that $\mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta_k-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] \to 1$. Due to the compactness of $\Theta^*(\delta_{\hat{\theta}})$, we can find a subsequence of $\{\theta_k\}_k$ converging to some $\theta_0 \in \Theta^*(\delta_{\hat{\theta}})$. For convenience we drop the subsequence and suppose $\theta_k \to \theta_0$. Then the existence of Y allows us to use dominated convergence theorem:

$$\mathbb{E}[e^{2Y_{\theta_0}}] = \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta_0-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] = \lim_{k\to\infty} \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta_k-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] = 1.$$

However, Jensen's inequality would indicate that $\mathbb{E}[e^{2Y_{\theta_0}}] = 1$ if and only $\mathbb{P}(Y_{\theta_0} = 0) = 1$, which contradicts our assumption. Thus, we know that

$$\inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] > 1,$$

as claimed. □

Proof of Proposition 4.3.3. First we prove (4.18). Letting $c = \frac{1}{(2\pi)^D|\Sigma|}$, we know that

$$\begin{aligned}
\chi^2(\mathbb{P}_0, \mathbb{P}_\theta) =& \int \frac{p^2(y;\theta)}{p_0(y)}dy - 1 \\
=& c\int \frac{e^{-(y-\theta)^T\Sigma^{-1}(y-\theta)}}{p_0(y)}dy - 1 \\
=& ce^{-\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2^2}\int \frac{e^{-(y-\hat{\theta})^T\Sigma^{-1}(y-\hat{\theta})}\cdot e^{-2(y-\hat{\theta})^T\Sigma^{-1}(\hat{\theta}-\theta)}}{p_0(y)}dy - 1 \\
=& c|\Sigma^{1/2}|e^{-\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2^2}\int \frac{e^{-z^Tz}\cdot e^{-2z^T\Sigma^{-1/2}(\hat{\theta}-\theta)}}{p_0(\Sigma^{1/2}z+\hat{\theta})}dz - 1 \\
=& c|\Sigma|e^{-\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2^2}\int \frac{e^{-z^Tz}\cdot e^{-2z^T\Sigma^{-1/2}(\hat{\theta}-\theta)}}{p_Z(z)}dz - 1 \qquad \text{(C.16)}
\end{aligned}$$

where we denote $p_Z(\cdot)$ to be the density function of random variable $Z = \Sigma^{-1/2}(Y-\hat{\theta})$ with $Y \sim \mathbb{P}_0$ and the last two lines follow from a change of variable $z = \Sigma^{-1/2}(y-\hat{\theta})$. Now, since

$\|\Sigma^{-1/2}(\theta_1 - \hat{\theta})\|_2^2 = \|\Sigma^{-1/2}(\theta_2 - \hat{\theta})\|_2^2 = r$ for some r by assumption, it follows from (C.16) that $\chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_1}) = \chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_2})$ if we can show

$$\int \frac{e^{-z^T z} \cdot e^{-2z^T \Sigma^{-1/2}(\hat{\theta} - \theta_1)}}{p_Z(z)} dz = \int \frac{e^{-z^T z} \cdot e^{-2z^T \Sigma^{-1/2}(\hat{\theta} - \theta_2)}}{p_Z(z)} dz.$$

However, since $p_Z(z)$ and $e^{-z^T z}$ are both rotationally invariant functions (i.e. $f(z) = f(Q^\intercal z)$ for all z and rotational matrix Q, with $|Q| = 1$), it can be shown that $\int \frac{e^{-z^T z} \cdot e^{-2z^T \nu}}{p_Z(z)} dz$ holds the same value for any ν such that $\|\nu\|_2^2 = r$. Notice the rotational invariance of $p_Z(z)$ follows from the rotational invariance of Z. This proves (4.18). To prove (4.19), notice that for any $\theta \in \mathcal{U}_{data}$, we can find some $0 \le t \le 1$ such that

$$(((1-t)\hat{\theta} + t\theta^\star) - \hat{\theta})^\intercal \Sigma^{-1} (((1-t)\hat{\theta} + t\theta^\star) - \hat{\theta}) = (\theta - \hat{\theta})^\intercal \Sigma^{-1} (\theta - \hat{\theta})$$

and hence $\chi^2(\mathbb{P}_0, \mathbb{P}_{((1-t)\hat{\theta} + t\theta^*)}) = \chi^2(\mathbb{P}_0, \mathbb{P}_\theta)$ by (4.18). □

Proof of Proposition 4.3.4. To check that $Y \sim \mathbb{P}_t$ with density

$$p_t(y) = \int_\Theta p(y;\theta')((1-t)\delta_{\hat{\theta}} + t\mu_{prop})(d\theta') = (1-t)\mathbb{P}_{\hat{\theta}} + \int_\Theta p(y;\theta')\mu_{prop}(d\theta'),$$

leads to rotationally invariant $Z = \Sigma^{-1/2}(Y - \hat{\theta})$, simply notice that

$$Y \overset{\mathcal{D}}{=} (1 - U_t)(\hat{\theta} + X_1) + U_t(\hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta + X_2),$$

where U_t is an independent Bernoulli variable with success rate t, η is a random vector uniformly distributed on the surface of the D-dimensional unit ball and X_1, X_2 are independent $\mathcal{N}(0, \Sigma)$. Then, it follows that

$$\Sigma^{-1/2}(Y - \hat{\theta}) \overset{\mathcal{D}}{=} (1 - U_t)Z_1 + U_t(\sqrt{\frac{\chi^2_{1-\alpha,D}}{n}}\eta + Z_2)$$

where Z_1, Z_2 are now independent $\mathcal{N}(0, I_D)$. Consequently, the rotational invariance of Z now follows from the rotational invariance of Z_1, Z_2, η and their independence. □

Following the comments after Proposition 4.3.2, we show that $\theta \sim \mu_{prop}$ with $\theta \overset{\mathcal{D}}{=} \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2} \cdot \eta$ provides a descent direction, with an alternate proof using the following lemma and the last line of (C.15).

Lemma C.0.5. *Fixing $\theta_1 \in \Theta^*(\delta_{\hat{\theta}})$, we have*

$$\mathbb{E}_{\theta_2 \sim \mu_{prop}}[e^{2(\theta_1 - \hat{\theta})^T \Sigma^{-1}(\theta_2 - \hat{\theta})}] > 1,$$

for $\theta_2 \sim \mu_{prop}(d\theta)$ where $\theta_2 \overset{\mathcal{D}}{=} \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2} \cdot \eta$ with η following the uniform distribution on the surface of the D-dimensional unit ball.

Proof of Lemma C.0.5. Let $u_1 \in \mathbb{R}^D$ denote an arbitary point on the surface of D dimensional unit ball ($\|u_1\|_2^2 = 1$) and let $\eta = [\eta_1, \eta_2, ..., \eta_D]$ be the random vector in $\mathbb{R}^D$ uniformly distributed on the surface of D dimensional unit ball. Then we claim that $\frac{\mu_1^T \eta + 1}{2} \sim Beta(\frac{D-1}{2}, \frac{D-1}{2})$.

To show this, assume without loss of generality that $u_1 = [1, 0, ..., 0] \in \mathbb{R}^D$. Then for any $t \in [-1, 1]$, it follows that $\mathbb{P}(u_1^T \eta \in dt)$ is proportional to the infinitesimal surface area on the ball corresponding to $\eta_1 \in dt$, which is in turn proportional to the product of the sub-dimension $D-2$ surface area on the belt $x_2^2 + x_3^2 + ... + x_D^2 = 1 - t^2$ with the infinitesimal width of this belt. Specifically, the

sub-dimension $D-2$ surface area around the belt is proportional to $(\sqrt{1-t^2})^{D-2}$. This follows from the fact that points of the form $[0, \sqrt{1-t^2}, 0, ..., 0]$, $[0, 0, \sqrt{1-t^2}, 0, ..., 0]$, ..., $[0, 0, ..., 0, \sqrt{1-t^2}]$ are on this belt. Also, the width of this belt, according to the Pythagorean theorem, is $dt \cdot \sqrt{(\frac{d\sqrt{1-t^2}}{dt})^2 + 1} = \frac{dt}{\sqrt{1-t^2}}$. Thus,

$$\mathbb{P}(u_1^T \eta \in dt) \propto \frac{(\sqrt{1-t^2})^{D-2}}{\sqrt{1-t^2}} dt = (1-t^2)^{\frac{D-3}{2}} dt.$$

Now, we can substitute $\frac{t+1}{2} = s$ with $s \in [0,1]$ to get

$$\mathbb{P}(\frac{u_1^T \eta + 1}{2} \in ds) \propto (s)^{\frac{D-1}{2}-1}(1-s)^{\frac{D-1}{2}-1} ds,$$

which can only be the density function for $Beta(\frac{D-1}{2}, \frac{D-1}{2})$. It now follows from [68] that $\frac{u_1^T \eta + 1}{2}$ has moment generating function

$$\begin{aligned} M(t) \triangleq & \mathbb{E}[e^{t \cdot \frac{u_1^T \eta + 1}{2}}] \\ =& {}_1F_1(\frac{D-1}{2}, D-1, t) = e^{(t/2)} {}_0F_1(; \frac{D}{2}, \frac{t^2}{16}) \geq e^{t/2}(1+ct^2) > e^{(t/2)}. \end{aligned} \tag{C.17}$$

for some $c > 0$ where ${}_1F_1(\cdot,\cdot,\cdot)$ and ${}_0F_1(;,\cdot,\cdot)$ are the confluent hypergeometric function with identity ${}_1F_1(a, 2a, x) = e_0^{x/2} F_1(; a + 1/2, x^2/16)$ (see [19]),

$${}_0F_1(;\alpha,t) \triangleq \sum_{k=0}^{\infty} \frac{t^k}{(\alpha)_k k!} \quad \text{and} \quad {}_1F_1(\alpha,\beta,t) \triangleq \sum_{k=0}^{\infty} \frac{(\alpha)_k t^k}{(\beta)_k k!},$$

with $(\gamma)_k = \frac{\Gamma(\gamma+k)}{\Gamma(\gamma)}$ being the Pochhammer symbol [68]. To conclude the proof, denote $\rho_n = \sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})} \cdot \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}}$ and use (4.16), (4.17) and (C.17) to write

$$\begin{aligned} & \mathbb{E}_{\theta_2 \sim \mu_{prop}}[e^{2(\theta_1-\hat{\theta})^T \Sigma^{-1}(\theta_2-\hat{\theta})}] \\ =& \mathbb{E}_{v\sim\eta}[e^{2\rho_n \cdot \mu_1^T v}] = \mathbb{E}_{X\sim Beta(\frac{D-1}{2},\frac{D-1}{2})}[e^{2\rho_n \cdot (2X-1)}] = M(4\rho_n)/e^{2\rho_n} \geq (1+16c\rho_n^2) > 1. \end{aligned}$$

□

Remark. Following Lemma C.0.5, we discuss the numerical calculations of $\mathcal{D}(\mathbb{P}_0)$ following Proposition 4.3.4. We use $\mathcal{U}_{data} = \{\mathbb{P}_\theta : \|\theta - \hat{\theta}\|_2^2 \leq \frac{\chi^2_{1-\alpha,D}}{n}\}$ where $p(y;\theta) = (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|(y-\theta)\|_2^2}$. Then, for μ_1, the nominal $p_0(y)$ is simply $p(y;\hat{\theta})$ and $\mathcal{D}_{data}(\mathbb{P}_0) = \mathcal{D}_{data}(\mathbb{P}_{\hat{\theta}}) = \max_{\theta\in\mathcal{U}} e^{\|\theta-\hat{\theta}\|_2^2} - 1 = e^{\frac{\chi^2_{1-\alpha,D}}{n}} - 1$ according to (4.16) and Lemma C.0.3. For μ_2, it can be shown that the nominal $\mathbb{P}_0$ follows $\mathcal{N}(\hat{\theta}, (1+\frac{1}{n}) \cdot I_D)$, and a direct computation would show that $\mathcal{D}_{data}(\mathbb{P}_0) = \max_{\theta\in\mathcal{U}} (\frac{(n+1)^2}{n(n+2)})^{\frac{d}{2}} e^{\frac{n}{n+2}\|\theta-\hat{\theta}\|_2^2} - 1 = (\frac{(n+1)^2}{n(n+2)})^{\frac{d}{2}} e^{\frac{n}{n+2}\frac{\chi^2_{1-\alpha,D}}{n}} - 1$. Finally, for μ_3, assume w.l.o.g that $\hat{\theta} = 0$. Then we use the derivation in Lemma C.0.5 that $\frac{\mu_1^T \eta + 1}{2} \sim Beta(\frac{D-1}{2}, \frac{D-1}{2})$ for any u_1 on the D-dimensional unit ball surface to show that, for any $v \in \mathbb{R}^D$,

$$\mathbb{E}_\eta[e^{\eta^T v}] = e^{-\|v\|_2} {}_1F_1(\frac{D-1}{2}, D-1, 2\|v\|_2), \tag{C.18}$$

and consequently

$$p_0(y) = (2\pi)_1^{-\frac{D}{2}} F_1(\frac{d-1}{2}, d-1, 2(\frac{\chi^2_{1-\alpha,D}}{n})^{1/2}\|y\|_2) e^{-\frac{1}{2}(\|y\|_2 + \frac{\chi^2_{1-\alpha,D}}{n})^2}.$$

Then, to calculate $\mathcal{D}_{data}(\mathbb{P}_0)$, we note that

$$\begin{aligned}\mathcal{D}_{data}(\mathbb{P}_0)+1 &= \max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \int \frac{p^2(y;\theta)}{p_0(y)}dy \\ &= \max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|y\|_2^2} \frac{e^{-\|\theta\|_2^2+2\theta^T y+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|y\|_2}}{{}_1F_1(\frac{d-1}{2},d-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|y\|_2)} \\ &= \max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_{Y\sim\mathcal{N}(0,I_D)}\left[\frac{e^{-\|\theta\|_2^2+2\theta^T Y+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|Y\|_2}}{{}_1F_1(\frac{d-1}{2},d-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|Y\|_2)}\right]. \end{aligned} \tag{C.19}$$

Furthermore, through either direct verification or analysis similar to those in Lemma C.0.5, we note that $Y\sim\mathcal{N}(0,I_D)$ shares the same distribution of $L\eta$ where $L\in\mathbb{R}^+$ and $\eta\in\mathbb{R}^D$ are two independent random variables with L being the norm of $\mathcal{N}(0,I_D)$ bearing density $f_L(l)=1_{\{l\ge0\}}\frac{2^{1-\frac{D}{2}}}{\Gamma(\frac{D}{2})}l^{d-1}e^{-\frac{l^2}{2}}$ and η being the random vector on the D-dimensional unit ball surface. Thus, it follows from (C.19) that (C.19) equals

$$\begin{aligned}&\max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_L\left[\mathbb{E}_\eta\left[\frac{e^{-\|\theta\|_2^2+2L\theta^T\eta+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L}}{{}_1F_1(\frac{d-1}{2},d-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L)}\middle|L\right]\right] \\ =&\max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_L\left[e^{-\|\theta\|_2^2+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L-2L\|\theta\|_2}\frac{{}_1F_1(\frac{D-1}{2},D-1,4L\|\theta\|_2)}{{}_1F_1(\frac{D-1}{2},D-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L)}\right] \\ =&\max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_L\left[e^{-\|\theta\|_2^2+\frac{\chi^2_{1-\alpha,D}}{2n}}\frac{{}_0F_1(;\frac{D}{2},L^2\|\theta\|_2^2)}{{}_0F_1(;\frac{D}{2},L^2(\frac{\chi^2_{1-\alpha,D}}{4n})}\right] \\ =&\max_{t\le\frac{\chi^2_{1-\alpha,D}}{n}} e^{-t+\frac{\chi^2_{1-\alpha,D}}{2n}}\int_{l\ge0}\frac{{}_0F_1(;\frac{D}{2},l^2t)}{{}_0F_1(;\frac{D}{2},l^2(\frac{\chi^2_{1-\alpha,D}}{4n})}\frac{2^{1-\frac{D}{2}}}{\Gamma(\frac{D}{2})}l^{d-1}e^{-\frac{l^2}{2}}dl\end{aligned}$$

which is numerically tractable.

Proof of Theorem 4.4.1. It follows from routine calculation that we can find a compact neighborhood of r around 0 such that $\nabla_r L(r,\theta)$ exists and is continuous. Thus we can use the main theorem in [18] to show that

$$\begin{aligned}\lim_{r\downarrow0}\frac{l(r)-l(0)}{r} &= \sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\int_{\mathcal{Y}} -\frac{(p(y;\theta))^2\cdot\lim_{r\downarrow0}\frac{p_r(y)-p(y;\hat\theta)}{r}}{(p(y;\hat\theta'))^2}dy \\ &= \sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\lim_{r\downarrow0}\frac{1}{r}\int_{\mathcal{Y}}\frac{(p(y;\theta))^2(p(y;\hat\theta)-p_r(y))}{(p(y;\hat\theta'))^2}dy \\ &= \sup_{\theta\in\Theta^*(\delta_{\hat\theta})}\lim_{r\downarrow0}\frac{1}{r}\int_{\mathcal{Y}}\frac{(p(y;\theta))^2}{p(y;\hat\theta')}-\frac{(p(y;\theta))^2\int_{\theta'\in\Theta}p(y;\theta')\mu_r(d\theta')}{(p(y;\hat\theta'))^2}dy \\ &= (1+\frac{\chi^2_{1-\alpha,D}}{n})\cdot\lim_{r\downarrow0}\frac{1}{r}\Big(1-\inf_{\theta\in\Theta^*(\delta_{\hat\theta})}\mathbb{E}_{\theta'\sim\mu_r}[e^{2(\theta-\hat\theta)^\intercal\Sigma^{-1}(\theta'-\hat\theta)}]\Big).\end{aligned}$$

□

To prove Corollary 4.4.2, we present two technical Lemmas C.0.6 and C.0.7.

Lemma C.0.6. *For any $\theta \in \Theta^*(\delta_{\hat{\theta}})$, $\lim_{r\downarrow 0} \frac{1}{r}\left(1-\mathbb{E}_{\theta'\sim\mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}]\right)$ is a fixed negative value.*

Proof of Lemma C.0.6. For any $\theta \in \Theta^*(\delta_{\hat{\theta}})$, we have $\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2 = \sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})}$. Denote $\rho_n = \sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})}$. Furthermore, under $\theta' \sim \mu_r^1(d\theta')$, we have $\Sigma^{-1/2}(\theta'-\hat{\theta}) \sim \eta_{\sqrt{r}}$, the uniform distribution inside the D-dimensional unit ball with radius $\sqrt{r}$, which can be viewed as the product of two independent random variables

$$\eta_{\sqrt{r}} \sim U \cdot R,$$

where U is the uniform distribution on the surface of the D-dimensional unit ball and R is the norm of the random vector ranged from 0 to $\sqrt{r}$. For any $0 \le s \le \sqrt{r}$, since $\eta_{\sqrt{r}}$ follows a uniform distribution inside a D-dimensional unit ball, and the volume of a D-dimensional ball with radius s is proportional to s^D, then $f_R(s)$, the density of R, must satisfy

$$f_R(s) \sim \frac{ds^D}{ds} \sim s^{D-1},$$

which is equivalent to saying

$$f_R(s) = \frac{D}{(\sqrt{r})^D} s^{D-1}, \quad \text{for } 0 \le s \le \sqrt{r}.$$

Thus, we have that $\mathbb{E}[R^2] = c_1 r$ for some $c_1 > 0$. Now we let $u_1 = [1, 0, ..., 0] \in \mathbb{R}^D$. We utilize the proof in Lemma C.0.5 as well as the independence of R, U to show that

$$\begin{aligned}
\mathbb{E}_{\theta'\sim\mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] =& \mathbb{E}_{U,R}[e^{2\rho_n\cdot R\cdot u_1^\intercal U}] \\
=& \mathbb{E}_R[\mathbb{E}[e^{2\rho_n\cdot R\cdot u_1^\intercal U}|R]] \\
=& \mathbb{E}_R[M(4\rho_n R)/e^{2\rho_n R}] \\
\ge& \mathbb{E}[1+16c\rho_n^2 R^2] \ge 1+16c\rho_n^2 c_1 r.
\end{aligned}$$

Now it follows that

$$\lim_{r\downarrow 0}\frac{1}{r}\left(1-\mathbb{E}_{\theta'\sim\mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}]\right) \le -16c\rho_n^2 c_1.$$

□

Lemma C.0.7. *For any $\theta \in \Theta^*(\delta_{\hat{\theta}})$, $\lim_{r\downarrow 0} \frac{1}{r}\left(1-\mathbb{E}_{\theta'\sim\mu_r^2}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}]\right)$ is a fixed negative value.*

Proof of Lemma C.0.7. For any $\theta \in \Theta^*(\delta_{\hat{\theta}})$, we have $\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2 = \sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})}$. Denote $\rho_n = \sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})}$. Furthermore, under $\theta' \sim \mu_r^2(d\theta')$, we have $\Sigma^{-1/2}(\theta'-\hat{\theta}) \sim \mathcal{N}(0, rI_D)$. Using the moment generating function for Gaussian random variables, we have

$$\mathbb{E}_{\theta'\sim\mu_r^2}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}] = e^{(2r\cdot(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta}))} = e^{2r\rho_n^2} \ge 1+2r\rho_n^2.$$

Now it follows that

$$\lim_{r\downarrow 0}\frac{1}{r}\left(1-\mathbb{E}_{\theta'\sim\mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}]\right) \le -2\rho_n^2.$$

□

Proof of Corollary 4.4.2. Lemmas C.0.6 and C.0.7 combined with (4.23) indicate that increasing r to positive value would produce a descent direction for $l(r)$ at $r = 0$. □

Proof of Corollary 4.4.3. We proceed the proof as in Proposition 4.3.4. The proof for the case of $\mu_r^1(d\theta)$ is entirely similar. For the proof of the case $\mu_r^2(d\theta)$, we simply notice that if $Y \sim \mathbb{P}_t$, then

$$Y \overset{\mathcal{D}}{=} (1 - U_t)(\hat{\theta} + X_1) + U_t(\hat{\theta} + \sqrt{r}X_2 + X_3),$$

where U_t is an independent Bernoulli variable with success rate t and X_1, X_2, X_3 are independent $\mathcal{N}(0, \Sigma)$. Then, it follows that

$$\Sigma^{-1/2}(Y - \hat{\theta}) \overset{\mathcal{D}}{=} (1 - U_t)Z_1 + U_t(\sqrt{r}Z_2 + Z_3)$$

where Z_1, Z_2, Z_3 are now independent $\mathcal{N}(0, I_D)$. Consequently, the rotational invariance of Z now follows from the rotational invariance of Z_1, Z_2, Z_3 and their independence.

□