Non-Parameteric Conformal Distributionally Robust Optimization

by

Guyang (Kevin) Cao

A dissertation submitted in fulfillment
of the requirements for the
Undergraduate Honors Thesis
(Statistics)
in the University of Michigan
2024

Thesis Advisor:

Professor Ambuj Tewari
Yash Patel

TABLE OF CONTENTS

ABSTRACT

Distributionally robust optimization (DRO) represents a sophisticated approach to making informed decisions in situations where uncertainty plays a significant role, especially in environments where safety cannot be compromised. This methodology offers a structured way to tackle decision-making by preparing for the worst-case scenarios within a defined set of possibilities known as the ambiguity set, aiming to ensure that our decisions remain effective even under the least favorable conditions.

There has been a notable shift towards enriching this traditional framework with data-driven insights in recent years. The goal here is to refine how we define the ambiguity set by leveraging actual data, thereby making DRO more relevant and applicable to real-world scenarios. However, a common limitation of these advancements is their reliance on a globally predefined ambiguity set. This means the set is determined before solving the optimization problem and does not adapt based on new data or insights gained during the decision-making process.

Addressing this limitation, we introduce a novel approach called Conformalized Distributionally Robust Optimization (CRDO). CRDO innovatively combines DRO with conformal prediction, a statistical technique that uses past data to make predictions about future outcomes within a certain level of confidence. Through conformal prediction, CRDO dynamically generates ambiguity sets based on the space of probability measures—mathematically rigorous ways to quantify uncertainties. This approach not only allows for more flexibility in responding to changing data but also enables us to make strong, confidence-backed claims about the performance of the decisions made using CRDO.

To demonstrate the versatility and power of the CRDO framework, we apply it to a series of benchmark tasks within the realm of simulation-based inference to showcase how CRDO can improve decision-making processes by adaptively tuning to the intricacies and uncertainties inherent in each scenario.

In essence, CRDO represents a significant leap forward in the domain of optimization under uncertainty, blending classical principles with modern data-driven methodologies to create a robust, adaptive, and highly applicable approach to tackling some challenging decision-making scenarios in safety-critical settings.

CHAPTER 1

Introduction

Stochastic optimization is a mature field often used in safety-critical situations, such as in the deployment of self-driving cars [4, 17, 5, 30]. Traditionally, problems are framed as seeking decisions $w \in \mathcal{W}$ that minimize $\mathbb{E}_{\mathcal{P}(C)}[f(w, C)]$ for some objective f and random information dictating the decision quality w.

In cases where contextual information is present, this requires explicit modeling of the posterior $\mathcal{P}(C \mid X)$ distribution. However, in many scientific applications, such as in astrophysics, neuroscience, and particle physics [26, 22, 18, 11, 27, 7], decisions are sought over a very large collection of x, typically on the order of 10,000 or more. Thus, even where the likelihood and prior are well specified, exact sampling from the $\mathcal{P}(C \mid X)$ posterior becomes intractable, as doing so would require running on the order of 10,000 separate MCMC chains.

In such scenarios, *amortized variational inference* (amortized VI) is frequently employed, leading to decisions being made based on the expectation $\mathbb{E}_{q_{\varphi(x)}(C)}[f(w, C)]$. Here, amortized VI operates by approximating the posterior distribution of uncertain parameters C given observed data X through a variational approach. This method utilizes a neural network or a similar model with parameters φ, which are learned from the data to map inputs X to the parameters of the distribution $q_{\varphi(x)}(C)$. This process is termed "amortized" because the cost of learning the mapping is spread across all data points, making it efficient for large datasets or complex models where traditional inference methods become computationally prohibitive.

The decision objective $\mathbb{E}_{q_{\varphi(x)}(C)}[f(w, C)]$ signifies evaluating the expected outcome of the decision w under the estimated distribution of C. This objective function f represents the goals or costs associated with different decisions, incorporating the uncertainties captured by $q_{\varphi(x)}(C)$. By employing amortized VI, we seek to minimize or optimize this expectation, thus making informed decisions that account for the underlying uncertainty.

While amortized VI offers a computationally efficient alternative to methods like MCMC for estimating posterior distributions, it's also noted for its limitations, including the potential for introducing bias and lacking theoretical guarantees [6, 24, 39, 38]. This poses challenges, particularly in fields where precise uncertainty quantification is crucial. For instance, in the subfield of likelihood-free inference, a meta-study on widely used algorithms relying on VI revealed that they consistently produce unfaithful, overconfident posterior approximations [19]. Nonetheless, the use of amortized VI in calculating the decision objective $\mathbb{E}_{q_{\varphi(x)}(C)}[f(w, C)]$ represents a pragmatic approach to dealing with complex, high-dimensional data in decision-making processes.

Separately, distributionally robust optimization (DRO) arose, in which solutions of this optimization set are instead sought over an ambiguity set $\mathcal{U}(\mathcal{P})$ of distributions [30, 20, 21]. Significant progress has been achieved in this vein, but DRO requires a priori knowledge of plausible ambiguity sets or noise distributions to produce practically useful answers. An overly conservative ambiguity set will likely result in suboptimal performance in typical circumstances. Towards this end, data-driven DRO has recently become of interest, in which plausible ambiguity sets are learned empirically [12, 23, 10].

Conformal prediction provides a principled framework for producing distribution-free uncertainty quantification with marginal frequentist guarantees [3, 34]. By using conformal prediction on a user-defined score function $s(x, y)$ and obtaining an empirical $1 - \alpha$ quantile $\widehat{q}(\alpha)$ of $s(x, y)$ over a calibration set $\mathcal{D}_C$, prediction regions $\mathcal{C}(x) = \{y \mid s(x, y) \leq \widehat{q}(\alpha)\}$ attain marginal coverage guarantees. Similar to DRO, the utility of such prediction regions is directly related to the nature of the score function: a poor choice of score may result in overly conservative, meaningless prediction sets.

A recent procedure CPO (Conformal-Predict-Then-Optimize) leverages such conformal prediction regions for predict-then-optimize decision-making. In this vein, we extend CPO and propose CDPO (Conformal-Distributional-Predict-Then-Optimize), a procedure that leverages conformal prediction to produce prediction regions over probability measures and thereby produces guarantees on stochastic decision-making algorithms that rely on amortized variational inference, in turn unifying the fields of distributionally robust optimization and predict-then-optimize decision-making. Our main contributions are:

- Proposing a new framework (CDPO) for data-driven distributionally robust optimization that has strong guarantees downstream.
- Demonstrating the use of conformal prediction over non-parametric probability measures.
- Demonstrating the generality of the CDPO framework on a suite of SBI tasks.

CHAPTER 2

Background

2.1 Conformal Prediction

Given a dataset $\mathcal{D}_{\mathcal{C}} = \{(X_1, y_1), \ldots (X_{N_{\mathcal{C}}}, y_{N_{\mathcal{C}}})\}$ of i.i.d. observations from a distribution $\mathcal{P}(Y, X)$, conformal prediction [3, 34] produces prediction regions with distribution-free theoretical guarantees. A prediction region is a mapping from observations of X to sets of possible values for Y. A prediction region $\mathcal{C}$ is said to be marginally calibrated at the $1 - \alpha$ level if $\mathcal{P}(Y \notin \mathcal{U}(X)) \leq \alpha$.

Split conformal is one popular version of conformal prediction. In this approach, marginally calibrated regions $\mathcal{C}$ are designed using a "score function" $s(x, y)$. Intuitively, the score function should have the quality that $s(x, y)$ is smaller when it is more reasonable to guess that $Y = y$ given the observation $X = x$. For example, if one has access to a function $\hat{f}(x)$ which attempts to predict Y from X, one might take $s(x, y) = \|\hat{f}(x) - y\|$. The score function is evaluated on each point of the dataset $\mathcal{D}_{\mathcal{C}}$, called the "calibration dataset," yielding $\mathcal{S} = \{s(x^{(j)}, y^{(j)})\}_{j=1}^{N_{\mathcal{C}}}$. Note that the calibration dataset cannot be used to pick the score function; if data is used to design the score function, it must be independent of $\mathcal{D}_{\mathcal{C}}$. This is how "split conformal" gets its name: in typical cases, data are split into two parts, one used to design s and the other to perform calibration. We then define $\widehat{q}(\alpha)$ as the $\lceil (N_{\mathcal{C}} + 1)(1 - \alpha) \rceil / N_{\mathcal{C}}$ quantile of $\mathcal{S}$. For any future x, the set $\mathcal{U}(x) = \{y \mid s(x, y) \leq \widehat{q}(\alpha)\}$ satisfies $1 - \alpha \leq \mathcal{P}(Y \in \mathcal{U}(X))$. This inequality is known as the coverage guarantee, and it arises from the exchangeability of the score of a test point $s(x', y')$ with $\mathcal{S}$. Those new to conformal prediction may be surprised to note that the coverage guarantee holds regardless of the number of samples $N_{\mathcal{C}}$ used in calibration; conformal guarantees are not asymptotic results.

As noted in Vovk's tutorial [34], while the coverage guarantee holds for any score function, different score functions may lead to more or less informative prediction regions. For example, the score $s(x, y) = 1$ leads to the highly uninformative prediction region of all possible values of Y. Predictive efficiency is one way to quantify informativeness [37, 33]. It is defined as the inverse of the expected Lebesgue measure of the prediction region, i.e. $(\mathbb{E}[|\mathcal{U}(X)|])^{-1}$. Methods employing conformal prediction often seek to identify prediction regions that are efficient as well as calibrated.

2.2 Variational Inference

Bayesian methods aim to sample the posterior distribution $\mathcal{P}(\Theta \mid X)$, typically using either MCMC or VI. VI has risen in popularity recently due to how well it lends itself to amortization. Given an observation X, variational inference transforms the problem of posterior inference into an optimization problem by seeking

$$\varphi^*(X) = \arg\min_{\varphi} D(q_\varphi(\Theta) || \mathcal{P}(\Theta \mid X)), \tag{2.1}$$

where D is a divergence and q_φ is a member of a variational family of distributions $\mathcal{Q}$ indexed by the free parameter φ. Normalizing flows have emerged as a particularly apt choice for $\mathcal{Q}$, as they are highly flexible and perform well empirically [31, 1]. Amortized variational inference expands on this approach by training a neural network to approximate $\varphi^*(X)$. This leads to a variational posterior approximator $q(\Theta \mid X) = q_{\varphi^*(X)}(\Theta)$ that can be rapidly computed for any value X. The characteristics of φ^* depend in part on the variational objective, D. For instance, using a reverse-KL objective, i.e. $D_{KL}(q_\varphi(\Theta) || \mathcal{P}(\Theta \mid X))$, is known to produce mode-seeking posterior approximations, whereas using a forward-KL objective, i.e. $D_{KL}(\mathcal{P}(\Theta \mid X) || q_\varphi(\Theta))$, encourages mode-covering behavior [25]. Popular objectives include the Forward-Amortized Variational Inference (FAVI) objective [2, 8], the Evidence Lower Bound (ELBO), and the Importance Weighted ELBO (IWBO) [9].

2.3 Predict-then-Optimize

We present a summary of predict-then-optimize problems and the application of conformal prediction in this setting, specifically from [28]. Such problems can be formulated as

$$w^*(x) := \min_{w \in \mathcal{W}} \quad \mathbb{E}[f(w, C) \mid x], \tag{2.2}$$

where w are decision variables, C an *unknown* cost parameter, x observed contextual variables, $\mathcal{W}$ a compact feasible region, and $f(w, c)$ an objective function that is convex-concave and L-Lipschitz in c for any fixed w. The predict-then-optimize framework is so called as the nominal approach first predicts $\widehat{c} := f(x)$ and subsequently solves $\min_w f(w, \widehat{c})$. Alternatively, a predictive contextual distribution $\mathcal{P}(C \mid x)$ is assumed, with respect to which the optimization formulation is solved. A more comprehensive review of the predict-then-optimize literature is presented in [13].

The formulation in Equation (2.2), however, is inappropriate in risk-sensitive downstream tasks. For this reason, recent works have begun investigating a risk-sensitive variant or "robust" alternative to this formulation. One way to do so follows in the vein of conditional value at risk as studied in [28, 36], where the goal marginalizes over the observed contexts:

$$\mathbb{E}_X[\min_{w \in \mathcal{W}} \mathrm{VaR}_\alpha[f(w, C) \mid X]]. \tag{2.3}$$

Crucially, this formulation of interest does not lend itself naturally to a solution. For this reason, we consider an equivalent framing as the following robust optimization problem:

$$\begin{aligned} w^*(x) := \min_{w, \mathcal{U}} \max_{\widehat{c} \in \mathcal{U}(x)} & \quad f(w, \widehat{c}) \\ \text{s.t.} \quad & \mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha, \end{aligned} \tag{2.4}$$

where $\mathcal{U} : \mathcal{X} \to \mathcal{F}$ is a uncertainty region predictor. In [28], $\mathcal{U}(x)$ was specifically constructed via conformal prediction to satisfy the desired probabilistic constraint. That is, $\mathcal{U}(x)$ was taken to be the prediction region $\mathcal{C}(x)$ produced by conformalizing a predictor $q : \mathcal{X} \to \mathcal{C}$. The primary consequence of doing so was having a probabilistic guarantee on the suboptimality gap. That is, denoting the suboptimality gap by $\Delta(x, c) := \min_w \max_{\widehat{c} \in \mathcal{C}(X)} f(w, \widehat{c}) - \min_w f(w, c)$, where c is the true parameter corresponding to x, they established that leveraging conformal uncertainty regions guarantees $\mathcal{P}_{X,C} \left(0 \leq \Delta(X, C) \leq L \ \mathrm{diam}(\mathcal{C}(X))\right) \geq 1 - \alpha$.

2.4 Distributionally Robust Optimization

We present much of the background from this section as a review of the corresponding sections from [20]. For a fully comprehensive discussion on the following topics, we refer readers to the relevant sections therein. Consider a decision problem under uncertainty. We have a measurable, extended real-valued loss function denoted as $l(\xi)$ to model each admissible decision result, where $\xi \in \mathbb{R}^m$ is a random vector governed by a probability distribution $\mathcal{P}$. Let the feasible set of l be $\mathcal{L}$. We then have the *risk* of a decision $l \in \mathcal{L}$ defined as the expected loss under $\mathcal{P}$:

$$\mathcal{R}(\mathcal{P}, l) := \mathbb{E}_{\mathcal{P}}[l(\xi)], \tag{2.5}$$

and the *optimal risk* as the lower bound of the risk above within the feasible set $\mathcal{L}$:

$$\mathcal{R}(\mathcal{P}, \mathcal{L}) := \inf_{l \in \mathcal{L}} \mathcal{R}(\mathcal{P}, l). \tag{2.6}$$

2.4.1 Selection of Nominal Distribution

In most decision-making situations, the distribution $\mathcal{P}$ is unknown, meaning we have to find a nominal distribution $\widehat{\mathcal{P}}_N$ as an estimator. Using such an estimator, Equation (2.5) and Equation (2.6) become solvable, simply replacing the respective terms with $\mathcal{R}(\widehat{\mathcal{P}}_N, l)$ and $\mathcal{R}(\widehat{\mathcal{P}}_N, \mathcal{L})$.

(1) For non-parametric models, since there is not enough structural information from the model itself, we have to construct $\widehat{\mathcal{P}}_N$ as the discrete empirical distribution sampling from the unknown $\mathcal{P}$, which is defined as

$$\widehat{\mathcal{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\widehat{\xi}_i},$$

where N denotes the size of training samples and $\delta_{\widehat{\xi}_i}$ denotes the i-th Dirac function whose mass is at the i-th training sample $\widehat{\xi}_i$. Moreover, surprisingly, for the optimization problem $\inf_{Q \in \mathcal{P}_N} W_1(Q, \mathcal{P})$, where $\mathcal{P}_N$ is the family of N-atom discrete probability distributions defined over a compact set $\Xi \in \mathbb{R}^m$, the empirical distribution obtains the optimal convergence rate $N^{-\frac{1}{m}}$ [20].

(2) For parametric models, since there is more structural information presented, we instead set $\widehat{\mathcal{P}}_N$ to be an elliptical distribution

$$\widehat{\mathcal{P}}_N = \mathcal{E}_g(\widehat{\mu}, \widehat{\Sigma}),$$

where g is a generator depending on the specific structure and training samples and $\widehat{\mu}$ and $\widehat{\Sigma}$ are the mean and covariance maximum estimators, respectively. Since we formulated $\widehat{\mathcal{P}}$ with an elliptical structure, finding the estimator $\widehat{\mathcal{P}}_N$ is equivalent to finding an estimator $\widehat{\theta}_N = (\widehat{\mu}, \widehat{\Sigma})$. We know the MLE estimator $\widehat{\theta}_N^{MLE}$ is asymptotically unbiased and efficient as N grows, making it the preferred estimator assuming the structural form of $\widehat{\mathcal{P}}_N$ is well-specified.

2.4.2 Wasserstein Distance

For any $p \in [1, \infty]$, the type-p Wasserstein Distance between two probability Q and Q' on $\mathbb{R}^m$ is defined as:

$$W_p(Q, Q') = \left(\inf_{\pi \in \Pi(Q,Q')} \int_{\mathbb{R}^m \times \mathbb{R}^m} \|\xi - \xi^{'}\|^p \pi(d\xi, d\xi^{'}) \right)^{\frac{1}{p}} \tag{2.7}$$

where $\|\cdot\|$ denotes a norm on $\mathbb{R}^m$ and $\Pi(Q, Q')$ denotes the set of all possible joint probability distributions of ξ and $\xi^{'}$ with marginal distributions Q and $Q^{'}$, respectively.

This problem can be interpreted as the minimum cost of transporting a pile of dirt at the location represented by a distribution Q to the location represented by a distribution $Q^{'}$, where the moving cost of a unit mass is $\|\xi - \xi^{'}\|^p$, and a transportation plan is encoded in π, which means the probability $\pi(A \times B)$ is the amount of mass transport from the initial region A to the target region B.

2.4.3 Strong Duality of Wasserstein Distance

The linear problem above has a strong dual, whose strong duality is proved in [20]:

$$\begin{aligned} W_p^p(Q, Q') = \sup & \left\{ \int_{\mathbb{R}^m} \psi(\xi')Q'(d\xi') - \int_{\mathbb{R}^m} \phi(\xi)Q(d\xi) \right\} \\ \text{s.t.} \quad & \phi \text{ and } \psi \text{ are bounded continuous functions on } \mathbb{R}^m \text{ with} \\ & \psi(\xi) - \phi(\xi') \leq \|\xi - \xi'\|^p \quad \forall \xi, \xi' \in \mathbb{R}^m, \end{aligned} \tag{2.8}$$

where $W_p^p(Q, Q')$ is the p-th power of the p-Wasserstein Distance between Q and $Q^{'}$.

Define the Lipschitz modulus of an extended real-valued function ϕ on $\mathbb{R}^m$ as $\mathrm{Lip}(\phi)$, and

$$\mathrm{Lip}(\phi) := \sup_{\xi \neq \xi'} \frac{|\phi(\xi) - \phi(\xi^{'})|}{\|\xi - \xi^{'}\|}.$$

When $p = 1$, the constraint in problem 2.4 becomes $\mathrm{Lip}(\phi) = \sup_{\xi \neq \xi'} \frac{|\phi(\xi) - \phi(\xi^{'})|}{\|\xi - \xi^{'}\|} \leq 1$, which means problem 2.4 becomes

$$W_1(Q, Q') = \sup_{\mathrm{Lip}(\phi) \leq 1} \int_{\mathbb{R}^m} \psi(\xi')Q'(d\xi') - \int_{\mathbb{R}^m} \phi(\xi)Q(d\xi). \tag{2.9}$$

which is called the Kantorovich-Rubinstein theorem [20]. For a L-Lipschitz continuous loss function $l(\xi)$, and a nominal distribution $\widehat{\mathcal{P}}_N$ with $W_1(\widehat{\mathcal{P}}_N, \mathcal{P}) \leq \varepsilon$, we have

$$\begin{aligned}\left|\mathcal{R}(\widehat{\mathcal{P}}_N, l) - \mathcal{R}(\mathcal{P}, l)\right| &= \mathbb{E}_{\widehat{\mathcal{P}}_N}[l(\xi)] - \mathbb{E}_{\mathcal{P}}[l(\xi)] \\ &= \int_{R^m} l(\xi)\widehat{\mathcal{P}}_N(d\xi) - \int_{R^m} l(\xi)\mathcal{P}(d\xi) \\ &= L \cdot (\int_{R^m} \frac{l(\xi)}{L}\widehat{\mathcal{P}}_N(d\xi) - \int_{R^m} \frac{l(\xi)}{L}\mathcal{P}(d\xi))\end{aligned}$$

Since $l(\xi)$ is L-Lipschitz continuous, $\mathrm{Lip}(l) = \sup_{\xi \neq \xi'} \frac{|l(\xi) - l(\xi')|}{\|\xi - \xi'\|} \leq L$, which means $\mathrm{Lip}(\frac{l}{L}) = \sup_{\xi \neq \xi'} \frac{|\frac{l(\xi)}{L} - \frac{l(\xi')}{L}|}{\|\xi - \xi'\|} \leq \frac{L}{L} = 1$. Therefore, we have that

$$\left|\mathcal{R}(\widehat{\mathcal{P}}_N, l) - \mathcal{R}(\mathcal{P}, l)\right| \leq L \cdot W_1(\widehat{\mathcal{P}}_N, \mathcal{P}) \leq L \cdot \varepsilon.$$

2.4.4 Extra Notations

- *Wasserstein metric.* Let $\Xi \subseteq \mathbb{R}^m$ be a closed set. We denote $\mathcal{P}(\Xi)$ as the family of all probability supported on Ξ. We then define the ambiguity set

  $$\mathbb{B}_{\varepsilon,p}(\widehat{\mathcal{P}}_N) = \left\{Q \in \mathcal{P}(\Xi) : \, W_p(Q, \widehat{\mathcal{P}}_N) \leq \varepsilon\right\}$$

  as the ball centered at the nominal distribution with radius ε in the metric of p-Wasserstein distance. When we do optimization within the feasible set, we can treat ε as the maximum estimation error.

- *Indicator function.* The indicator function for a set $\Xi \in \mathbb{R}^m$ is defined as

  $$\delta_\Xi(\xi) = \begin{cases} 0 & \text{if } \xi \in \Xi, \\ \infty & \text{if } \xi \notin \Xi. \end{cases}$$

- *Conjugate function.* The conjugate of a function $l(\xi)$ on $\mathbb{R}^m$ is defined as

  $$l^*(z) = \sup_{\xi} z^T\xi - l(\xi).$$

  The conjugate of indicator function is

  $$\sigma_\Xi(z) = \delta^*_\Xi(z) = \sup_{\xi} z^T\xi - \delta_\Xi(\xi)$$

  .

Consider the case $\xi \in \Xi$, we obtain the conjugate of the indicator function for the set Ξ is $\sigma_\Xi(z) = \sup_\xi z^T\xi$, which coincides with the definition of *support function* for Ξ. Therefore, on a certain set Ξ, the indicator function and the support function are conjugate of each other.

- *Dual Norm.* If $\|\xi\|$ denotes the norm of $\xi \in \mathbb{R}^m$, then $\|z\|_* = \sup_{\|\xi\|\leq 1} z^T\xi$ denotes the corresponding dual norm.

2.4.5 High-level Analysis of the Optimal Worst-case Risk

Use the notation in section 2.4.4, we have the *worst-case risk*

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathcal{P}}_N, l) := \sup_{Q\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathcal{P}}_N)} \mathcal{R}(Q, l) \tag{2.10}$$

and the *worst-case optimal risk*

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathcal{P}}_N, \mathcal{L}) := \inf_{l\in\mathcal{L}} \mathcal{R}(\widehat{\mathcal{P}}_N, l) \tag{2.11}$$

Equation (2.11) is the general formulation of distributionally robust optimization problems. This problem can often be reformulated as a finite convex program which can be solved in polynomial time, and the detailed process of the reformulation will be shown below.

2.4.6 Computation and Proof

The *worst-case risk* (2.6) for any fixed loss function $l \in \mathcal{L}$ satisfies **strong duality**:

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathcal{P}}_N, l) = \inf_{\gamma>0} \mathbb{E}_{\widehat{\mathcal{P}}_N}[l_\gamma(\xi)] + \gamma\varepsilon^p \tag{2.12}$$

where $l_\gamma(\xi) = \sup_{z\in\Xi} l(z) - \gamma\|z-\xi\|^p$ defined on a convex set Ξ. The proof of this duality proceeds as follows. We first note that

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathcal{P}}_N, l) = \sup_{Q\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathcal{P}}_N)} \mathcal{R}(Q, l) = \sup_{W_p(Q,\widehat{\mathcal{P}}_N)\leq\varepsilon} \mathbb{E}_Q(l(\xi)).$$

This problem can be reformulated as below by the definition of expectation and p-Wasserstein Distance:

$$\begin{cases} \sup\limits_{\Pi,Q} & \displaystyle\int_\Xi l(\xi)Q(d\xi) \\ \text{s.t.} & \displaystyle\int_{\Xi^2} \|\xi-\xi'\|^p\Pi(d\xi, d\xi') \leq \varepsilon^p \end{cases} \tag{2.13}$$

In the non-parametric setting, we know from section 2.4.1 that $\widehat{\mathcal{P}}_N$ should be written as a mean of the sum of N Dirac functions with the mass of each is at i-th training sample $\widehat{\xi_i}$. Therefore, we know the joint probability $\Pi(\xi, \xi')$ can be constructed from the marginal distributions $\widehat{\mathcal{P}}_N$ and Q_i respectively of ξ' and ξ. In turn, we can write $\Pi(d\xi, d\xi') = \frac{1}{N}\sum\limits_{i=1}^{N} \delta_{\widehat{\xi_i}} \otimes Q_i$, where the tensor product denotes the combination of two independent probability distribution $\delta_{\widehat{\xi_i}}$ and Q_i. Let $\mathcal{M}(\Xi)$ denote the set of all Dirac distributions supported on Ξ. Equation (2.13) can then be rewritten as

$$\begin{cases} \sup\limits_{Q_i \in \mathcal{M}(\Xi)} & \dfrac{1}{N}\sum\limits_{i=1}^{N} \displaystyle\int_{\Xi} l(\xi) Q_i(d\xi) \\ \quad \text{s.t.} & \dfrac{1}{N}\sum\limits_{i=1}^{N} \displaystyle\int_{\Xi} \|\xi - \widehat{\xi}_i\|^p Q_i(d\xi) \le \varepsilon^p, \end{cases} \tag{2.14}$$

Using the standard Lagrange dual, Equation (2.14) can be written in its dual form as follows:

$$\sup_{Q_i \in \mathcal{M}(\Xi)} \inf_{\gamma \ge 0} \left\{ \frac{1}{N}\sum_{i=1}^{N} \int_{\Xi} l(\xi) Q_i(d\xi) + \gamma(\varepsilon^p - \frac{1}{N}\sum_{i=1}^{N} \int_{\Xi} \|\xi - \widehat{\xi}_i\|^p Q_i(d\xi)) \right\} \tag{2.15}$$

$$\le \inf_{\gamma \ge 0} \sup_{Q_i \in \mathcal{M}(\Xi)} \left\{ \gamma\varepsilon^p + \frac{1}{N}\sum_{i=1}^{N} \int_{\Xi} (l(\xi) - \gamma\|\xi - \widehat{\xi}_i\|^p) Q_i(d\xi) \right\} \tag{2.16}$$

$$= \inf_{\gamma \ge 0} \gamma\varepsilon^p + \frac{1}{N}\sum_{i=1}^{N} \sup_{\xi \in \Xi} \left\{ l(\xi) - \gamma\|\xi - \widehat{\xi}_i\|^p \right\}, \tag{2.17}$$

where we invoked the Minimax Theorem to perform the second step and noted that the term $\gamma\varepsilon^p$ is invariant in the feasible set and hence could be moved out in the final step. Let $[N]$ denote $\{1, 2, \cdots, N\}$, to maximize the remaining part, in our non-parametric modeling, we have to find the N Dirac distributions $Q_1, \cdots, Q_N$, and each Q_i ($i \in [N]$) will have its mass attained on and only on a $\xi = \xi_i^* \in \Xi$ such that $l(\xi) - \gamma\|\xi - \widehat{\xi}_i\|^p$ attains $\sup_{\xi \in \Xi} \left\{ l(\xi) - \gamma\|\xi - \widehat{\xi}_i\|^p \right\}$. Moreover, since the integral of the Dirac function is 1 due to the definition of the probability mass function, the second equality holds as above.

For the last problem above, we only substitute the variable ξ with z, and it can be written as $\inf_{\gamma \ge 0} \gamma\varepsilon^p + \frac{1}{N}\sum\limits_{i=1}^{N} \sup_{z \in \Xi} \left\{ l(z) - \gamma\|z - \widehat{\xi}_i\|^p \right\}$. Since $\widehat{\mathcal{P}}_N$ is a probability distribution with equal mass at N points $\widehat{\xi}_i$, $i \in [N]$, so we have

$$\mathbb{E}_{\widehat{\mathcal{P}}_N}[l_\gamma(\xi)] = \frac{1}{N}\sum_{i=1}^{N} \sup_{z \in \Xi} \left\{ l(z) - \gamma\|z - \widehat{\xi}_i\|^p \right\}$$

by the definition of expectation, $l_\gamma(\xi)$, and the special property of $\widehat{\mathcal{P}}_N$. Therefore, we have proved that problem Equation (2.10) can be reformulated as its duality form Equation (2.12). Since $l_\gamma(\xi)$ is a Moreau-Yosida regularization [20] of $l(\xi)$, and $l_\gamma(\xi)$ is jointly convex in γ and l for any fixed ξ, therefore the dual form Equation (2.12) is a convex minimization problem whose optimal value is also convex in l, and that's part of the reason why the strong duality is achieved. Equation (2.12) is still an intermediate step of the reformulation, and we will show the remaining necessary assumptions to make the strong duality attained in the process of proving the reformulation to the final finite convex program.

• (*Convexity Assumptions* [14, p.128 Assumption 4.1]) The set $\Xi \in \mathbb{R}^m$ is convex and closed, and $l(\xi) = \max_{j\in[J]} l_j(\xi)$ with the negative constituent functions $-l_j$ proper, convex, lower semicontinous, and not indentically ∞ for all $j \in [j]$. By proposition 3.4 in [35], the inequality of Equation (2.16) can be reduced to equality under the assumptions above. Therefore, theorem Equation (2.12) is fully proved.

Now under the convexity assumption, we continue to reformulate Equation (2.12) into a finite convex program:

$$\begin{aligned} \mathcal{R}_{\varepsilon,p}(\widehat{P}_N, l) = \inf\ & \gamma\varepsilon^p + \frac{1}{N}\sum_{i=1}^{N} s_i. \\ & \text{s.t.}\quad \gamma \in \mathbb{R}_+, s_i \in \mathbb{R}, u_{ij} \in \mathbb{R}^m, v_{ij} \in \mathbb{R}^m \\ & [-l_j]^*(u_{ij} - v_{ij}) + \sigma_\Xi(v_{ij}) - u_{ij}^T\widehat{\xi}_i + \varphi(q)\gamma\left\|\frac{u_{ij}}{\gamma}\right\|_*^q \le s_i \quad \forall i \in [N], j \in [J], \end{aligned} \tag{2.18}$$

Where $\varphi(q) = \frac{(q-1)^{q-1}}{q^q}$ for $q > 1$ and $\varphi(1) = 1$.

Now we start to prove Equation (2.18). Continue the notation of Equation (2.17) and reformulate it as:

$$\begin{cases} \inf\limits_{\gamma,s_i}\ \gamma\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{s.t.}\quad \sup\limits_{\xi\in\Xi}\left(l(\xi) - \gamma\|\xi - \widehat{\xi}_i\|^p\right) \le s_i, \quad \forall i \in [N] \\ \qquad\ \lambda \ge 0 \end{cases} \tag{2.19}$$

$$= \begin{cases} \inf\limits_{\gamma, s_i} \ \gamma\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{s.t.} \sup\limits_{\xi\in\Xi, u_{ij}\in\mathbb{R}^m} \left(l_j(\xi) - u_{ij}^T(\xi - \widehat{\xi}_i) + u_{ij}^T(\xi - \widehat{\xi}_i) - \gamma\|\xi - \widehat{\xi}_i\|^p \right) \leq s_i, \quad \forall i \leq N, \forall j \leq J \\ \lambda \geq 0 \end{cases} \tag{2.20}$$

$$\leq \begin{cases} \inf\limits_{\gamma, s_i} \ \gamma\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{s.t.} \sup\limits_{\xi\in\Xi, u_{ij}\in\mathbb{R}^m} \left(l_j(\xi) - u_{ij}^T(\xi - \widehat{\xi}_i) \right) + \sup\limits_{\xi\in\Xi, u_{ij}\in\mathbb{R}^m} \left(u_{ij}^T(\xi - \widehat{\xi}_i) - \gamma\|\xi - \widehat{\xi}_i\|^p \right) \leq s_i, \quad \forall i \leq N, \forall j \leq J \\ \lambda \geq 0. \end{cases} \tag{2.21}$$

In the first equality, we first exploit the property in the convexity assumptions that $l(\xi)$ can be written as the maximum of a series of constituent concave loss functions. Then we add and delete the term $u_{ij}^T(\xi - \widehat{\xi}_i)$ once in the first constraint, and split it into two parts in the second inequality. Since we have

$$\begin{aligned} &\sup_{\xi\in\Xi, u_{ij}\in\mathbb{R}^m} \left(l_j(\xi) - u_{ij}^T(\xi - \widehat{\xi}_i) + u_{ij}^T(\xi - \widehat{\xi}_i) - \gamma\|\xi - \widehat{\xi}_i\|^p \right) \\ &\leq \sup_{\xi\in\Xi, u_{ij}\in\mathbb{R}^m} \left(l_j(\xi) - u_{ij}^T(\xi - \widehat{\xi}_i) \right) + \sup_{\xi\in\Xi, u_{ij}\in\mathbb{R}^m} \left(u_{ij}^T(\xi - \widehat{\xi}_i) - \gamma\|\xi - \widehat{\xi}_i\|^p \right), \end{aligned}$$

the first constraint in Equation (2.21) is more strict than that in Equation (2.20), which means the minimum value Equation (2.20) can attain is at least as small as that in Equation (2.21).

For

$$\sup_{\xi\in\Xi, u_{ij}\in\mathbb{R}^m} \left(l_j(\xi) - u_{ij}^T(\xi - \widehat{\xi}_i) \right),$$

substitute u_{ij} with $-u_{ij}$, we have

$$\begin{aligned}
&\sup_{\xi\in\Xi,u_{ij}\in\mathbb{R}^m} \left\{l_j(\xi)+u_{ij}^T\xi\right\}-u_{ij}^T\widehat{\xi}_i\\
&=[-l_j+\delta_\Xi]^*-u_{ij}^T\widehat{\xi}_i\\
&=\mathbf{cl}\left([-l_j]^*\,\#\,[\delta_\Xi]^*\right)-u_{ij}^T\widehat{\xi}_i\\
&=\mathbf{cl}\left(\inf_{v_{ij}}\left([-l_j]^*\left(u_{ij}-v_{ij}\right)+[\delta_\Xi]^*\left(v_{ij}\right)\right)\right)-u_{ij}^T\widehat{\xi}_i.
\end{aligned}$$

For the first equality, since the indicator function is defined on Ξ, $\delta_\Xi=0$ for $\xi\in\Xi$, so this term can be added within the conjugate of $-l_j$.

For the second equality, we first introduce an operation called *epi-addition* or *int-convolution* [32, p. 23 1(12)] as below:

For $f_1:\mathbb{R}^m\longrightarrow\mathbb{R}$ and $f_2:\mathbb{R}^m\longrightarrow\mathbb{R}$,

$$(f_1\,\#\,f_2)=\inf_{x_1+x_2=x}\left\{f_1(x_1)+f_2(x_2)\right\}\;=\inf_{w}\left\{f_1(x-w)+f_2(w)\right\},$$

and from [32, p. 493 Theorem 11.23 (a)], for $f=(f_1\,\#\,f_2)$, we will have $f^*=f_1^*+f_2^*$, and if $f=f_1+f_2$, f_1 and f_2 are proper, convex, and lower semicontinous such that **dom** f_1 meets **dom** f_2, then $f^*=\mathbf{cl}(f_1^*\,\#\,f_2^*)$, when **cl** denotes the closure operation of a set. Under the convexity assumption, the second quality holds.

For the third equality, it holds naturally by the definition of *epi-addition* introduced before. Since the closure operation maps any function to its largest lower semicontinuous minorant. As $\mathbf{cl}[f(\xi)]<0$ if and only if $f(\xi)<0$ for any function f [14, p.131], we can conclude that

$$\begin{aligned}
\sup_{\xi\in\Xi,u_{ij}\in\mathbb{R}^m}\left(l_j(\xi)-u_{ij}^T(\xi-\widehat{\xi}_i)\right)&=\sup_{\xi\in\Xi,u_{ij}\in\mathbb{R}^m}\left\{l_j(\xi)+u_{ij}^T\xi\right\}-u_{ij}^T\widehat{\xi}_i\\
&=[-l_j]^*\left(u_{ij}-v_{ij}\right)+\sigma_\Xi(v_{ij})-u_{ij}^T\widehat{\xi}_i. \qquad (2.22)
\end{aligned}$$

For

$$\sup_{\xi\in\Xi,u_{ij}\in\mathbb{R}^m}\left(u_{ij}^T(\xi-\widehat{\xi}_i)-\gamma\|\xi-\widehat{\xi}_i\|^p\right),$$

it equals to

$$\sup_{\xi\in\Xi,u_{ij}\in\mathbb{R}^m}\left(u_{ij}^T\xi-\gamma\|\xi-\widehat{\xi}_i\|^p\right)-u_{ij}^T\widehat{\xi}_i=\varphi(q)\gamma\left\|\frac{u_{ij}}{\gamma}\right\|_*^q. \qquad (2.23)$$

Before giving a vigorous justification for Equation (2.23), we introduce the concept of *Conjugates of Powers of Norms* [40, p.71 Lemma C.9]. Assume $\|\cdot\|$ and $\|\cdot\|_*$ are mutually dual norms, and p, $q \in [1, \infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$, then the conjugate of $h(z) = \frac{1}{p}\|z\|^p$ is given by

$$\begin{aligned} h^*(y) &= \sup_z \left\{ z^T y - \frac{1}{p}\|z\|^p \right\} \\ &\leq \sup_z \left\{ \|z\| \|y\|_* - \frac{1}{p}\|z\|^p \right\} \\ &= \max_{t \geq 0} \left\{ t\|y\|_* - \frac{1}{p} t^p \right\} \end{aligned} \tag{2.24}$$

we can easily find that Equation (2.24) attains its maximum when $t = \|y\|_*^{\frac{1}{p-1}}$, and the maximum value is $(1 - \frac{1}{p})\|y\|_*^{\frac{p}{p-1}} = \frac{1}{q}\|y\|_*^q$, using $\frac{1}{p} + \frac{1}{q} = 1$. When $\|z\| = \|y\|_*^{\frac{1}{p-1}}$, the first inequality is tight, hence we have

$$h^*(y) = \sup_z \left\{ z^T y - \frac{1}{p}\|z\|^p \right\} = \frac{1}{q}\|y\|_*^q. \tag{2.25}$$

Now we can prove Equation (2.23). Since LHS in Equation (2.23) can be written as:

$$\begin{aligned} \sup_{\xi \in \Xi, u_{ij} \in \mathbb{R}^m} \left(u_{ij}^T \xi - \gamma \|\xi - \widehat{\xi}_i\|^p \right) - u_{ij}^T \widehat{\xi}_i &= u_{ij}^T \widehat{\xi}_i + \sup_{\xi} \left\{ u_{ij}^T \xi - \gamma \|\xi\|^p \right\} - u_{ij}^T \widehat{\xi}_i \\ &= \gamma p \sup_{\xi} \left\{ \frac{(u_{ij}/\gamma)^T}{p} \xi - \frac{1}{p}\|\xi\|^p \right\} \\ &= \gamma p \cdot \frac{1}{q} \left\| \frac{u_{ij}}{\gamma} \cdot \frac{1}{p} \right\|_*^q \\ &= \gamma \frac{p^{1-q}}{q} \left\| \frac{u_{ij}}{\gamma} \right\|_*^q \\ &= \gamma \frac{[q/(q-1)]^{1-q}}{q} \left\| \frac{u_{ij}}{\gamma} \right\|_*^q \\ &= \gamma \frac{[(q-1)/q]^{q-1}}{q} \left\| \frac{u_{ij}}{\gamma} \right\|_*^q \\ &= \varphi(q) \gamma \left\| \frac{u_{ij}}{\gamma} \right\|_*^q \end{aligned} \tag{2.26}$$

In Equation (2.26), the first equality use ξ to substitute $\xi - \widehat{\xi}_i$, the third quality uses Equation (2.25), and the last equalities use $\frac{1}{p} + \frac{1}{q} = 1$. Therefore, we have proved Equation (2.23) and the optimal value obtained when $\xi^* = \widehat{\xi}_i + \left\| \frac{u_{ij}}{\gamma} \right\|_*^{\frac{1}{p-1}}$ following from the process of deriving Equation (2.24). Combining Equation (2.22) and Equation (2.23) and the inequality (tight because the derivation of Equation (2.22) is regardless of ξ) right below Equation (2.21), we derived the constraints for Equation (2.18), and hence justify Equation (2.18).

CHAPTER 3

Conformalized Distributionally Robust Optimization

3.1 Method

We now propose CRDO, a method that produces prediction regions over probability measures and thereby enables distribution-free claims to be made downstream. We focus on settings of contextual DRO as in [15], namely where we predict full *distributions* $\mathcal{Q}_{\varphi(x)}(C)$. We assume well-specified prior and likelihood models, respectively $\mathcal{P}(C)$ and $\mathcal{P}(X \mid C)$, with complex posteriors distributions $\mathcal{P}(C \mid X)$, for which amortized variational inference is applied.

3.1.1 CDPO: Score Function

Let $c \in \mathcal{C}$, where $(\mathcal{C}, d)$ is a general metric space, and $\mathcal{F}$ be the σ-field of $\mathcal{C}$. While the standard predict-then-optimize framework assumes a linear objective function $c^T w$, we consider general convex-concave objective functions $f(w, c)$ that are L-Lipschitz in c under the metric d for any fixed w. With this generalization, the robust formulation of predict-then-optimize can be stated as

$$\begin{aligned} w^*(x) := \inf_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} \quad & \mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)] \\ \text{s.t.} \quad & \mathcal{P}_{X, \mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha, \end{aligned} \tag{3.1}$$

where $\mathcal{U} : \mathcal{X} \to \mathcal{M}(\mathcal{F})$ is a uncertainty region predictor over the space of probability measures on $\mathcal{F}$. Exact solution of this problem is intractable, as no practical methods exist to optimize over the measure space $\mathcal{U}$. For any fixed $\mathcal{U}$, this robust counterpart to the stochastic predict-then-optimize problem produces a valid upper bound if we use the following score function:

$$s(x, \mathcal{P}_C) = \mathcal{W}_1(\mathcal{Q}_{\varphi(x)}(C), \mathcal{P}_C), \tag{3.2}$$

where $\mathcal{W}_1$ represents the 1-Wasserstein distance. To compute the quantile $\widehat{q}$ of such a score over $\mathcal{D}_{\mathcal{C}}$, we assume the recovery of samples from the exact posterior $\mathcal{P}(C \mid x)$ for a subset of x, namely via MCMC methods. That is, we assume a dataset of the form $\{x_i, \{c_j^{\mathcal{P}}\}_{j=1}^{N_{\mathcal{P}}}\}$ exists, where each $c_j^{\mathcal{P}} \sim \mathcal{P}(C \mid x_i)$.

From here, $\mathcal{C}(x) = \{\mathcal{Q} \mid s(x, \mathcal{Q}) \leq \widehat{q}(\alpha)\}$ has marginal guarantees in the form $\mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{C}(X)) \geq 1 - \alpha$. Notably, even computing $\mathcal{W}$ for multi-dimensional distributions is a computationally challenging task; however, we can use the well-known equivalence between computing $\mathcal{W}_1$ and the Assignment Problem, which can be solved in $\mathcal{O}(N^3)$ with the Hungarian Algorithm [29]. With this choice of score function, we can bound the nominal stochastic optimal value:

$$\Delta(x, \mathcal{P}_C) := \inf_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} \mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)] - \inf_{w \in \mathcal{W}} \mathbb{E}_{\mathcal{P}_C}[f(w, C)].$$

We clearly see $\Delta(x, \mathcal{P}_C) \geq 0$ if $\mathcal{P}_C \in \mathcal{U}(x)$. This framing makes clear the consequences of leveraging *efficient* prediction regions with guaranteed coverage, formalized below.

Lemma 3.1.1. *Consider any $f(w, c)$ that is L-Lipschitz in c under the metric d for any fixed w. Assume further that $\mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha$ with $\sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} \mathcal{W}_1(\widetilde{\mathcal{Q}}, \mathcal{P}_C) = \operatorname{diam}(\mathcal{U}(x))$. Then, $\mathcal{P}_{X,\mathcal{P}_C}(\Delta(X, \mathcal{P}_C) \leq L \operatorname{diam}(\mathcal{U}(X))) \geq 1 - \alpha$.*

Proof. We consider the event of interest conditionally on a pair $(x, \mathcal{P}_C)$ where $\mathcal{P}_C \in \mathcal{U}(x)$:

$$\begin{aligned}
&|\inf_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} \mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)] - \inf_{w \in \mathcal{W}} \mathbb{E}_{\mathcal{P}_C}[f(w, C)]| \\
&\leq \sup_{w \in \mathcal{W}} |\sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} \mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)] - \mathbb{E}_{\mathcal{P}_C}[f(w, C)]| \\
&\leq \sup_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} |\mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)] - \mathbb{E}_{\mathcal{P}_C}[f(w, C)]| \\
&\leq \sup_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} L\mathcal{W}_1(\widetilde{\mathcal{Q}}, \mathcal{P}_C) = L\operatorname{diam}(\mathcal{U}(x)).
\end{aligned}$$

Since $\mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha$, the result immediately follows. □

Thus, $1 - \alpha$ validity of the prediction region ensures the result of the RO procedure is a valid bound with probability $1 - \alpha$, and greater efficiency of the prediction region translates to a tighter upper bound.

3.1.2 CDPO: Optimization Algorithm

While the statement of Theorem 3.1.1 was made assuming the exact 1-Wasserstein distance could be computed, we note that this is untrue for any distribution of interest, for which this quantity must be estimated with samples drawn respectively from the distributions of interest. That is, to compute Equation (3.2), samples $\{c_j^{\mathcal{Q}}\}_{j=1}^{M_{\mathcal{Q}}} \sim \mathcal{Q}(C)$ are drawn, which, along with the corresponding samples coming from the dataset, can be used to define corresponding empirical distributions, namely as:

$$\widehat{\mathcal{Q}}(C \mid x_i) := \frac{1}{M_{\mathcal{Q}}} \sum_{j=1}^{M_{\mathcal{Q}}} \delta_{c_j^{\mathcal{Q}}} \qquad \widehat{\mathcal{P}}(C \mid x_i) := \frac{1}{M_{\mathcal{P}}} \sum_{j=1}^{M_{\mathcal{P}}} \delta_{c_j^{\mathcal{P}}}. \tag{3.3}$$

For simplicity of computation, we take $M_{\mathcal{P}} = M_{\mathcal{Q}} = M$. Using these empirical distributions, we are then able to estimate the 1-Wasserstein distance using the aforementioned Hungarian Algorithm. That is, with such samples the distance is estimated as:

$$\mathcal{W}_1(\widetilde{\mathcal{Q}}(C \mid x_i), \mathcal{P}(C \mid x_i)) \approx \mathcal{W}_1(\widehat{\mathcal{Q}}(C \mid x_i), \widehat{\mathcal{P}}(C \mid x_i)) = \inf_{\pi} \sum_{j=1}^{M} \left| c_j^{\mathcal{Q}} - c_{\pi(j)}^{\mathcal{P}} \right|, \tag{3.4}$$

where $\pi : [1, ..., M] \to [1, ..., M]$ is a permutation function. We note that this use of an estimate of 1-Wasserstein distance requires a modification to the standard proof of coverage paralleling that presented in [16], but we defer this formalization to future work. Despite a formal presentation, this estimation should not result in significant over- or under-coverage, as the estimand has no one-sided bias with respect to the true probability distance.

We then fix $\alpha \in [0, 1]$ and take $\mathcal{U}(x)$ to be the $1 - \alpha$ prediction region $\mathcal{C}(x)$. We now seek to solve Equation (3.1) for this choice of $\mathcal{U}(x)$. The constraint of the original formulation, therefore, is satisfied by virtue of taking $\mathcal{U}(x) := \mathcal{B}_{\widehat{q}}(\widehat{\mathcal{Q}})$. In turn, we are then left having to solve the optimization problem

$$\inf_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{B}_{\widehat{q}}(\widehat{\mathcal{Q}})} \mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)]. \tag{3.5}$$

We now leverage the insights of [20] to reframe this problem in a tractably solvable manner, as discussed extensively in the background section. That is, we can reformulate this problem simply as a regularized optimization problem in the following sense:

$$w^*(x) := \inf_{w \in \mathcal{W}} \left(\frac{1}{M} \sum_{i=1}^{M} f(w^\top c_i^{\mathcal{Q}}) + \widehat{q} \cdot \mathrm{Lip}(f) \cdot ||w||_\infty \right), \tag{3.6}$$

where we have specifically considered the case where $f(w, c) = f(w^\top c)$ with $c_i^{\mathcal{Q}}$ being samples drawn from $\mathcal{Q}_{\varphi(x)}(C)$. Note that this problem lends itself to an efficient solution algorithm, which we make use of in the experiments of the following section.

3.2 Experiments

We first study the fractional knapsack problem under various complex contextual mappings:

$$\begin{gathered}\inf_{w\in\mathcal{W}} \sup_{\widetilde{\mathcal{Q}}\in\mathcal{B}_{\widehat{q}}(\widehat{\mathcal{Q}})} \mathbb{E}_{\widetilde{\mathcal{Q}}}[-w^\top C] \\ \text{s.t.} w \in [0,1]^n, p^T w \leq B, \mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1-\alpha.\end{gathered} \tag{3.7}$$

where $p \in \mathbb{R}^n$ and $B > 0$. The distributions $\mathcal{P}(C)$ and $\mathcal{P}(X \mid C)$ are taken to be those from various simulation-based inference (SBI) benchmark tasks provided by [19], chosen as they have $\mathcal{P}(C \mid X)$ with complex structure. We specifically study Two Moons, Lotka-Volterra, Gaussian Linear Uniform, Bernoulli GLM, Susceptible-Infected-Recovered (SIR), and Gaussian Mixture. K reference posteriors were provided by the authors of [19] for each task, specifically using a modified rejection sampling scheme and taking $M = 10,000$. The variational family fit in all cases was a normalizing spline flow.

Using the reframing of the previous section, we solved the following equivalent formulation for the setups in question:

$$\begin{gathered}\inf_{w\in\mathcal{W}} \left(\frac{1}{M}\sum_{i=1}^{M} -w^\top c_i^{\mathcal{Q}} + \widehat{q}\cdot ||w||_\infty\right) \\ \text{s.t.} w \in [0,1]^n, p^T w \leq B, \mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1-\alpha,\end{gathered} \tag{3.8}$$

where we note that $f :=$ id has a Lipschitz constant of $\text{Lip}(f) = 1$. We then compute $\widehat{q}$ over the N_C reference-variational posterior pairs taking $\alpha = 0.9$, where $N_C = 10$ in [19]. Solving this problem, by Theorem 3.1.1 then produces a valid upper bound on the nominal stochastic solution, as demonstrated in the following results. We specifically report the expected suboptimality gap proportion, $\Delta_{\%} = \mathbb{E}_X[\Delta(X, C(X))/\min_w f(w, C(X))]$ below.

Table 3.1: Suboptimality gaps ($\Delta_{\%}$) across tasks. Means and standard deviations are reported over 3 test samples.

Task	$\Delta_{\%}$
SLCP	-0.562 (0.041)
Gaussian Linear Uniform	-0.430 (0.048)
Gernoulli GLM	-0.484 (0.169)
Gaussian Mixture	-0.167 (0.024)
Gaussian Linear	-0.805 (0.180)
Bernoulli GLM	-0.456 (0.155)

CHAPTER 4

Conclusion

We have shown that conformal prediction can be practically leveraged in settings of stochastic optimization to provide strong guarantees than those traditionally provided with variation posteriors. This initial work suggests many directions for extension:

First, we solely considered synthetic benchmark tasks in this initial study: future work should demonstrate that the performance and utility of this method is retained in a broader class of applications of practical interest.

Second, this method employed the reframing of the DRO using the empirical distributions of samples from the variational distribution: using the structural forms of the variational distributions would be a similarly interesting extension.

Moreover, the reformulation of the DRO problem in chapter 2.4 is compatible for any $p \in [1, \infty)$, and we only consider the case $p = 1$ as an initial step in our CDPO algorithm and experiments. Making the algorithm for general p is an extension which is potentially exciting.

BIBLIOGRAPHY

[1] Abhinav Agrawal, Daniel R Sheldon, and Justin Domke. Advances in black-box vi: Normalizing flows, importance weighting, and optimization. *Advances in Neural Information Processing Systems*, 33:17358–17369, 2020.

[2] Luca Ambrogioni, Umut Güçlü, Julia Berezutskaya, Eva Borne, Yağmur Güçlütürk, Max Hinne, Eric Maris, and Marcel Gerven. Forward amortized inference for likelihood-free variational marginalization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 777–786. PMLR, 2019.

[3] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

[5] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[7] Jan Boelts, Jan-Matthis Lueckmann, Richard Gao, and Jakob H Macke. Flexible and efficient simulation-based inference for models of decision-making. *Elife*, 11:e77220, 2022.

[8] Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

[9] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[10] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls. *Operations Research*, 2022.

[11] Michael Deistler, Pedro J Goncalves, and Jakob H Macke. Truncated proposals for scalable and hassle-free simulation-based inference. *arXiv preprint arXiv:2210.04815*, 2022.

[12] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[13] Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

[14] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations, 2017.

[15] Adrián Esteban-Pérez and Juan M Morales. Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming*, 195(1-2):1069–1105, 2022.

[16] Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

[17] Virginie Gabrel, Cécile Murat, and Aurélie Thiele. Recent advances in robust optimization: An overview. *European journal of operational research*, 235(3):471–483, 2014.

[18] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

[19] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021.

[20] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[21] Fengming Lin, Xiaolei Fang, and Zheming Gao. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1):159–212, 2022.

[22] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

[23] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[24] Kevin P Murphy. *Probabilistic machine learning: an introduction.* MIT press, 2022.

[25] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics.* MIT Press, 2023.

[26] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

[27] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[28] Yash Patel, Sahana Rayan, and Ambuj Tewari. Conformal contextual robust optimization. *arXiv preprint arXiv:2310.10003*, 2023.

[29] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[30] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[31] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[32] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer, Berlin, 2010.

[33] Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

[34] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[35] Alexander Shapiro. *On Duality Theory of Conic Linear Problems*, volume 57 of *Nonconvex Optimization and Its Applications*, pages 135–165. Springer, Boston, MA, 2001.

[36] Chunlin Sun, Linyu Liu, and Xiaocheng Li. Predict-then-calibrate: A new perspective of robust contextual lp. *arXiv preprint arXiv:2305.15686*, 2023.

[37] Yachong Yang and Arun Kumar Kuchibhotla. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

[38] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR, 2018.

[39] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

[40] Jianzhe Zhen, Daniel Kuhn, and Wolfram Wiesemann. A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle, 2023.