Conformal Contextual Robust Optimization

Yash Patel
University of Michigan

Sahana Rayan
University of Michigan

Ambuj Tewari
University of Michigan

Abstract

Data-driven approaches to predict-then-optimize decision-making problems seek to mitigate the risk of uncertainty region misspecification in safety-critical settings. Current approaches, however, suffer from considering overly conservative uncertainty regions, often resulting in suboptimal decision-making. To this end, we propose Conformal-Predict-Then-Optimize (CPO), a framework for leveraging highly informative, nonconvex conformal prediction regions over high-dimensional spaces based on conditional generative models, which have the desired distribution-free coverage guarantees. Despite guaranteeing robustness, such black-box optimization procedures alone inspire little confidence owing to the lack of explanation of why a particular decision was found to be optimal. We, therefore, augment CPO to additionally provide semantically meaningful visual summaries of the uncertainty regions to give qualitative intuition for the optimal decision. We highlight the CPO framework by demonstrating results on a suite of simulation-based inference benchmark tasks and a vehicle routing task based on probabilistic weather prediction.

1 INTRODUCTION

Predict-then-optimize or contextual robust optimization problems are of long-standing interest in safety-critical settings where decision-making happens under uncertainty (Sun, Liu, and Li, 2023; Elmachtoub and Grigas, 2022; Elmachtoub, Liang, and McNellis, 2020; Peršak and Anjos, 2023). In traditional robust op-

timization, results are made to be robust to distributions anticipated to be present upon deployment (Ben-Tal, El Ghaoui, and Nemirovski, 2009; Beyer and Sendhoff, 2007). Since such decisions are sensitive to proper model specification, recent efforts have sought to supplant this with data-driven uncertainty regions (Cheramin et al., 2021; Bertsimas, Gupta, and Kallus, 2018; Shang and You, 2019; Johnstone and Cox, 2021).

Model misspecification is ever more present in *contextual* robust optimization, spurring efforts to define similar data-driven uncertainty regions (Ohmori, 2021; Chenreddy, Bandi, and Delage, 2022; Sun, Liu, and Li, 2023). Such methods, however, focus on box- and ellipsoid-based uncertainty regions, both of which are necessarily convex and often overly conservative, resulting in suboptimal decision-making.

Conformal prediction provides a principled framework for producing distribution-free prediction regions with marginal frequentist coverage guarantees (Angelopoulos and Bates, 2021; Shafer and Vovk, 2008). 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 $\mathcal{D}_{\mathcal{C}}$, prediction regions $\mathcal{C}(x) = \{y \mid s(x, y) \leq \widehat{q}(\alpha)\}$ attain marginal coverage guarantees. Such prediction regions, however, are notably defined *implicitly*. For simple scores, such as residuals, an explicit expression of such regions can be written, making these the most common approaches used in practice (Tumu et al., 2023; Horwitz and Hoshen, 2022; Angelopoulos et al., 2022; Hu et al., 2022; Mao, Martin, and Reich, 2022).

The disadvantage is that such score functions ignore the structure that is often present in high-dimensional data, such as images. Choices of simplistic scores, thus, tend to be overly conservative and often produce convex prediction regions even when $\mathcal{P}(Y|X)$ is nonconvex. Recent work has demonstrated that defining scores using conditional generative models produces sharper and, hence, more informative prediction regions (Feldman, Bates, and Romano, 2023; Wang et al., 2022; Patel et al., 2023). We, thus, extend the line of data-driven predict-then-optimize work by considering such generative model-based prediction regions.
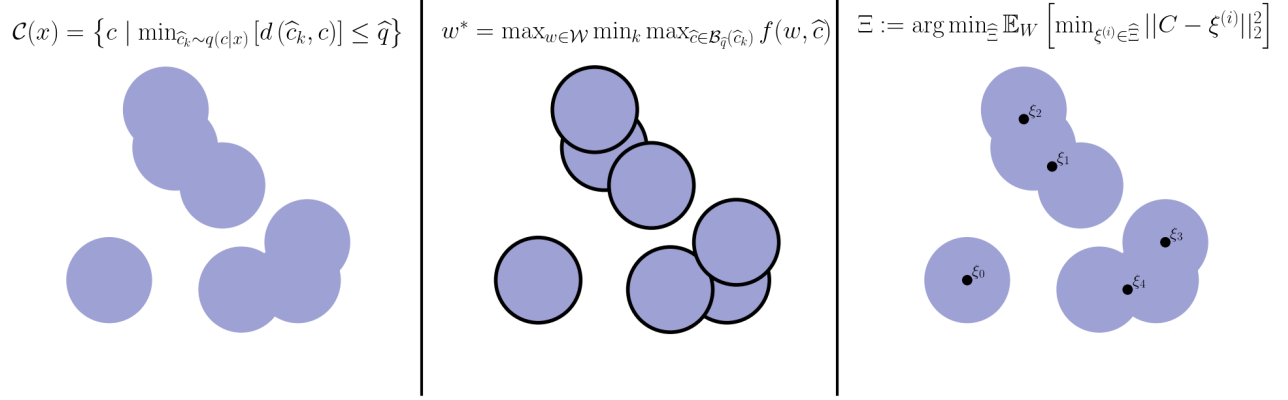
$$\mathcal{C}(x) = \left\{ c \mid \min_{\widehat{c}_k \sim q(c|x)} [d\,(\widehat{c}_k, c)] \leq \widehat{q} \right\} \qquad w^* = \max_{w \in \mathcal{W}} \min_k \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}) \qquad \Xi := \arg\min_{\widehat{\Xi}} \mathbb{E}_W \left[\min_{\xi^{(i)} \in \widehat{\Xi}} ||C - \xi^{(i)}||_2^2 \right]$$

Figure 1: CPO leverages informative, non-convex conformal prediction regions for robust predict-then-optimize decision making. CPO uses a score function such that the resulting prediction regions can be decomposed into convex subregions over which optimization can be carried out efficiently. Visual summaries $\{\xi^{(i)}\}$ of the prediction region can similarly be efficiently sampled to gain intuition on the optimal decision w^*.

In addition to contributing to the predict-then-optimize line of inquiry, we view this work as addressing a concern of the conformal prediction community: how to use implicitly defined non-convex, high-dimensional prediction regions. Works producing such regions have themselves noted the difficulty in their use (Sesia and Romano, 2021; Izbicki, Shimizu, and Stern, 2022). Initial works on coverage for images have framed the utility of their results in highlighting regions of the image with the greatest variability and, hence, uncertainty (Angelopoulos et al., 2022; Horwitz and Hoshen, 2022; Belhasin et al., 2023).

Extending such visualization gives invaluable intuition to the end user. For instance, a black-box optimization procedure for producing drug candidates to robustly bind to a predicted protein structure offers little insight into the decision-making process; however, semantic summaries of the uncertainty region would reveal regions of flexibility of the protein, clarifying why particular structures were deemed optimal in the candidate drug. Such interest in explainable robust decision-making was highlighted in a recent survey (Sadana et al., 2023), especially given the "right to explanation" mandated by the EU's "General Data Protection Regulation" (Doshi-Velez and Kim, 2017; Kaminski, 2019). Our main contributions, thus, are:

- Proposing Conformal-Predict-Then-Optimize (CPO) to leverage informative, non-convex prediction regions for decision-making.

- Providing interpretable visual summaries of uncertainty regions using representative points.

- Demonstrating the generality of CPO across a suite of benchmark tasks and a traffic routing task based on probabilistic weather prediction.

2 BACKGROUND

2.1 Conformal Prediction

Given a dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots (x^{(N)}, y^{(N)})\}$ of i.i.d. observations from a distribution $\mathcal{P}(X, Y)$, conformal prediction (Angelopoulos and Bates, 2021; Shafer and Vovk, 2008) produces prediction regions with distribution-free theoretical guarantees. A prediction region maps from observations of X to sets of possible values for Y and is said to be marginally valid at the $1 - \alpha$ level if $\mathcal{P}_{X,Y}(Y \notin \mathcal{C}(X)) \leq \alpha$.

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 a dataset $\mathcal{D}_{\mathcal{C}}$ called the "calibration dataset," yielding $\mathcal{S} = \{s(x^{(j)}, y^{(j)})\}_{j=1}^{N_{\mathcal{C}}}$, where $N_{\mathcal{C}} := |\mathcal{D}_{\mathcal{C}}|$. Note that the calibration dataset cannot be used to pick the score function; if data is used to design the score function, it must independent of $\mathcal{D}_{\mathcal{C}}$. 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 $\mathcal{C}(x) = \{y \mid s(x, y) \leq \widehat{q}(\alpha)\}$ satisfies $1 - \alpha \leq \mathcal{P}(Y \in \mathcal{C}(X))$. 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 $\mathcal{S}$. The coverage guarantee possesses finite-sample properties.

As noted in Vovk's tutorial (Shafer and Vovk, 2008), while the coverage guarantee holds for any score func-

tion, 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, defined as the inverse of the expected Lebesgue measure of the prediction region, i.e. $(\mathbb{E}[|\mathcal{C}(X)|])^{-1}$ (Yang and Kuchibhotla, 2021; Sesia and Candès, 2020). Methods employing conformal prediction often seek to identify prediction regions that are efficient and calibrated.

2.2 Representative Points

The problem of summarizing the distribution of a random vector with points $\Xi := \{\xi^{(i)}\}_{i=1}^N$ arises in many contexts, such as in optimal stratification (Dalenius, 1950; Dalenius and Gurney, 1951), density estimation (Flury and Tarpey, 1993), and signal quantization (Max, 1960). Such points are known as representative points (RPs). Denoting the space of all sets $\widehat{\Xi}$ such that $|\widehat{\Xi}| \leq n$ as ζ, the RPs of a random variable X are

$$\Xi := \underset{\widehat{\Xi} \in \zeta}{\arg\min} \, \mathbb{E}_X \left[\min_{\xi^{(i)} \in \widehat{\Xi}} ||X - \xi^{(i)}||_2^2 \right]. \qquad (1)$$

For a comprehensive review of representative points, see (Fang and Pan, 2023). Despite extensive study, no general algorithm exists for the efficient construction of representative points for arbitrary distributions. Typical implementations use clustering algorithms, such as Lloyd's algorithm, on $\{x^{(i)}\}_{i=1}^M \sim \mathcal{P}(X)$.

2.3 Predict-Then-Optimize

Predict-then-optimize problems are formulated as

$$w^*(x) := \min_{w \in \mathcal{W}} \quad \mathbb{E}[C^T w \mid x], \qquad (2)$$

where w are decision variables, C an *unknown* cost parameter, x observed contextual variables, and $\mathcal{W}$ a compact feasible region. The predict-then-optimize framework is so called as the nominal approach first predicts $\widehat{c} := f(x)$ and subsequently solves $\min_w \widehat{c}^T w$. Alternatively, a predictive contextual distribution $\mathcal{P}(C \mid x)$ is assumed, with respect to which the optimization formulation is solved. A full review is presented in (Elmachtoub and Grigas, 2022).

This formulation, 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 traditional formulation, namely by replacing $\mathbb{E}[C^T w \mid x]$ with $\max_{\widehat{c} \in \mathcal{U}(x)} \widehat{c}^T w$ (Ohmori, 2021; Chenreddy, Bandi, and Delage, 2022; Sun, Liu, and Li, 2023), where $\mathcal{U}(x)$ is constructed to guarantee coverage of c, precisely stated in Lemma 3.1.

3 METHOD

We now propose CPO, a way to perform robust predict-then-optimize decision-making over informative, non-convex prediction regions based on generative models. We then discuss how to construct visual summaries of the contents of the conformal prediction regions using a collection of N representative points.

3.1 CPO: Problem Formulation

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, as follows:

$$\begin{aligned} w^*(x) := \min_{w, \mathcal{U}} \max_{\widehat{c} \in \mathcal{U}(x)} \quad & f(w, \widehat{c}) \\ \text{s.t.} \quad & \mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha, \end{aligned} \qquad (3)$$

where $\mathcal{U} : \mathcal{X} \to \mathcal{F}$ is a uncertainty region predictor. Exact solution of this problem is intractable, as no practical methods exist to optimize over the predictor function space $\mathcal{U}$. Practical solution of this optimization problem, thus, involves optimizing over several prespecified uncertainty region predictors $\{\mathcal{U}_i\}_{i=1}^N$. For any fixed $\mathcal{U}$, this robust counterpart to the nominal predict-then-optimize problem produces a valid upper bound if $c \in \mathcal{U}(x)$. Denoting the pessimism gap as $\Delta(x, c) := \min_w \max_{\widehat{c} \in \mathcal{U}(x)} f(w, \widehat{c}) - \min_w f(w, c)$, we clearly see $\Delta(x, c) \geq 0$ if $c \in \mathcal{U}(x)$, formalized below.

Lemma 3.1. *Consider any $f(w, c)$ that is L-Lipschitz in c under the metric d for any fixed w. Assume further that $\mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha$. Then,*

$$\mathcal{P}_{X,C}\left(0 \leq \Delta(X, C) \leq L \operatorname{diam}(\mathcal{U}(X))\right) \geq 1 - \alpha. \quad (4)$$

The proof is deferred to Appendix A. Thus, $1 - \alpha$ validity of $\mathcal{U}$ ensures the RO procedure produces a valid bound with probability $1 - \alpha$, with more efficient prediction regions resulting in tighter upper bounds.

3.2 CPO: Score Function

We assume a conditional generative model $q(C \mid X)$ is learned for this prediction task. For most score functions, the min-max optimization problem of Equation 3 is computationally intractable. Crucially, however, we can consider an extension to the score proposed in (Wang et al., 2022), which lends itself to a decomposition under which such optimization becomes tractable. For a fixed K and $\{\widehat{c}_k\}_{k=1}^K \sim q(C \mid x)$, let

$$s(x, c) = \min_k \left[d\left(\widehat{c}_k, c\right)\right]. \qquad (5)$$

We refer to this score as "Generalized Probabilistic Conformal Prediction," (GPCP) whose validity follows from that of the original PCP framework (Wang et al., 2022). We discuss the selection of K in Section 3.4.

3.3 CPO: Optimization Algorithm

We fix $\alpha \in [0,1]$ and take $\mathcal{U}(x)$ to be the $1 - \alpha$ prediction region $\mathcal{C}(x)$. Let $\phi(w) := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$. It follows that $\phi(w)$ is convex by Danskin's Theorem by assumption of the convexity of f in w. Exact solution of the min-max problem, thus, follows using standard gradient-based optimization techniques on $\phi(w)$. By Danskin's Theorem, $\nabla_w \phi(w) = \nabla_w f(w, c^*)$, where $c^* := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$. We follow the standard projected gradient descent optimization scheme, projecting into $\mathcal{W}$ at each iterate, denoted by $\Pi_{\mathcal{W}}$.

Efficient solution of this RO problem, therefore, reduces to being able to efficiently solve the maximization problem over $\mathcal{C}(x)$. While challenging over general nonconvex regions, the GPCP score formulation lends itself to a highly structured prediction region, namely of the form $\mathcal{C}(x) = \bigcup_{k=1}^{K} \mathcal{B}_{\widehat{q}}(\widehat{c}_k)$ with $\mathcal{B}_{\widehat{q}}$ being a ball of radius $\widehat{q}$, the conformal quantile, under the d metric. This decomposition of $\mathcal{C}(x)$ means the maximum can be efficiently computed by aggregating the maxima over the individual balls:

$$\max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c}) = \max_k \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}), \qquad (6)$$

where the maximum over a ball can be efficiently computed with traditional convex optimization techniques. This procedure is summarized in Algorithm 1. The convergence of this procedure proceeds as follows, whose proof is deferred to Appendix B.

Algorithm 1 CPO-OPT

1: **procedure** CPO-OPT
   Inputs: Context x, CGM $q(C \mid X)$, Optimization steps T, Score samples K, Conformal quantile $\widehat{q}$
2: $\quad w \sim U(\mathcal{W}), \{\widehat{c}_k\}_{k=1}^K \sim q(C \mid x)$
3: $\quad$ **for** $t \in \{1, \ldots T\}$ **do**
4: $\quad\quad \left\{ c_k^* \leftarrow \arg\max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}) \right\}_{k=1}^K$
5: $\quad\quad c^* \leftarrow \arg\max_{c_k^*} f(w, c_k^*)$
6: $\quad\quad w \leftarrow \Pi_{\mathcal{W}}(w - \eta \nabla_w f(w, c^*))$
7: $\quad$ **end for**
8: $\quad$ **Return** w
9: **end procedure**

Lemma 3.2. *Let $\phi(w) := \max_{\widehat{c} \in \bigcup_{k=1}^K \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c})$ for $\{\widehat{c}_k\}_{k=1}^K \subset \mathcal{C}$, $\widehat{q} \in \mathbb{R}^+$, and $f(w, c)$ convex-concave and L-Lipschitz in c for any fixed w. Let $w^* \in \mathcal{W}$ be a minimizer of ϕ. For any $\epsilon > 0$, define $T := \frac{L^2 ||w_0 - w^*||}{\epsilon^2}$*

and $\eta := \frac{||w_0 - w^*||}{L\sqrt{T}}$. Then the iterates $\{w_t\}_{t=0}^T$ returned by Algorithm 1 satisfy

$$\phi \left(\frac{1}{T+1} \sum_{t=0}^T w_t \right) - \phi(w^*) \leq \epsilon. \qquad (7)$$

3.4 CPO: K Selection

Crucially, the convergence highlighted in Lemma 3.2 reveals that the number of "outer" iterations (i.e. T) has no dependence on K. This is apparent from the proof, in which the iterate count T hinges upon the Lipschitz constant of $\phi(w) = \max_k \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}) := \max_k \phi_k(w)$, which critically is L-Lipschitz *regardless* of what K is selected, as each $\phi_k(w)$ is L-Lipschitz.

We can, thus, solely focus attention on the impact the choice of K has on the "inner" optimization computational cost, namely $\max_k \phi_k(w)$. This linearly increasing cost with K, however, must be juxtaposed with the improved *statistical* efficiency of such prediction regions. In particular, (Wang et al., 2022) empirically demonstrated region size generally decreased nonlinearly up to a saturation point as a function of K.

Critically, this inflection point can be determined *prior* to performing the optimization, since doing so only requires access to $q(C \mid X)$ and test samples to estimate the prediction region size. As pointed out in (Wang et al., 2022) and proven in (Chan, 2008), estimation of the volume of a union of hyperspheres is complicated by the need to account for overlapped regions. K is, thus, chosen based on Monte Carlo estimates of the prediction region volume using Voronoi cells of the hypersphere centers given by (Edelsbrunner, 1995):

$$\widehat{\ell}(\{\mathcal{B}_{\widehat{q}}(\widehat{c}_k)\}) := |\mathcal{B}_{\widehat{q}}| \sum_{k=1}^K \mathcal{P}_{C \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))}(C \in V(\widehat{c}_k)), \ (8)$$

where $C \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))$ denotes a random variable defined uniformly over the region associated with $\widehat{c}_k$, $|\mathcal{B}_{\widehat{q}}|$ the volume of a hypersphere of radius $\widehat{q}$, and $V(\widehat{c}_k)$ the Voronoi cell of $\widehat{c}_k$, defined as $\{z \in \mathbb{R}^d \mid d(\widehat{c}_k, z) \leq d(\widehat{c}_{k'}, z), k' \neq k\}$. Muller's method enables efficient sampling of $U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))$ (Muller, 1959; Fishman, 2013).

We then choose K^* to be the inflection point, namely the $\arg\min_K |\widehat{\ell}_K - \widehat{\ell}_{K+1}| \leq \epsilon$ for some user-specified ϵ volume tolerance. Critically, these volume estimates must be performed on a distinct subset of the data from $\mathcal{D}_{\mathcal{C}}$ as exchangeability with future test points is otherwise lost in conditioning on $\mathcal{D}_{\mathcal{C}}$ for selecting K^* (Yang and Kuchibhotla, 2021). We, thus, partition $\mathcal{D}_{\mathcal{C}} := \mathcal{D}_{\mathcal{C}_1} \cup \mathcal{D}_{\mathcal{C}_2}$, using $\mathcal{D}_{\mathcal{C}_1}$ for calibration and $\mathcal{D}_{\mathcal{C}_2}$ for volume estimation, detailed in Algorithm 2.

Algorithm 2 CPO

1: **procedure** VOLUMEEST
   Inputs: Context x, CGM $q(C \mid X)$, Conformal quantile $\widehat{q}$
2:    $\{\widehat{c}_k\}_{k=1}^K \sim q(C_{1:K} \mid x)$
3:    $\left\{\{c_{k,m}\}_{m=1}^M \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))\right\}_{k=1}^K$
4:    **Return** $|B_{\widehat{q}}| \sum_{k=1}^K \frac{1}{M} \sum_{m=1}^M \mathbb{1}\left[c_{k,m} \in V(\widehat{c}_k)\right]$
5: **end procedure**

6: **procedure** CPO
   Inputs: Context x, CGM $q(C \mid X)$, Optimization steps T, Desired coverage $1-\alpha$, Max samples $K_{\max}$, Volume Tolerance ϵ, Calibration sets $\mathcal{D}_{\mathcal{C}_1}, \mathcal{D}_{\mathcal{C}_2}$
7:    **for** $K \in \{1, \dots K_{\max}\}$ **do**
8:       $s_K(x,c) \leftarrow \min_{\widehat{c}_k \in \{\widehat{c}_i\} \sim q(C_{1:K}|x)} \left[d\left(\widehat{c}_k, c\right)\right]$
9:       $\mathcal{S}_K \leftarrow \left\{s_K(x^{(i)}, c^{(i)}) \mid (x^{(i)}, c^{(i)}) \in \mathcal{D}_{\mathcal{C}_1}\right\}$
10:      $\widehat{q}_K \leftarrow \frac{\lceil (|\mathcal{D}_{\mathcal{C}_1}|+1)(1-\alpha)\rceil}{|\mathcal{D}_{\mathcal{C}_1}|}$ quantile of $\mathcal{S}_K$
11:      $\widehat{\ell}_K \leftarrow \frac{1}{|\mathcal{D}_{\mathcal{C}_2}|} \sum_{i=1}^{|\mathcal{D}_{\mathcal{C}_2}|}$ VOLUMEEST$(x^{(i)}, q, \widehat{q}_K)$
12:   **end for**
13:   $K^* \leftarrow \arg\min_K \left|\widehat{\ell}_K - \widehat{\ell}_{K+1}\right| \leq \epsilon$
14:   **Return** CPO-OPT$(x, q, T, K^*, \widehat{q}_{K^*})$
15: **end procedure**

3.5 CPO: Representative Points

We now frame the problem of summarizing the prediction region $\mathcal{C}(x)$. We critically note that this issue of interpretability is non-existent in traditional approaches to robust predict-then-optimize, where uncertainty regions are interpretable by construction, being balls around nominal estimates $\mathcal{B}_\epsilon(\widehat{c})$. In other words, there is a fundamental tension in qualitative interpretability and the expressiveness of uncertainty regions, requiring a bespoke method for recovering intuition when leveraging conformal prediction regions. Formally, for a user-specified number of summary points N and query x, we seek

$$\Xi(x) := \underset{\widehat{\Xi} \in \zeta}{\arg\min} \, \mathbb{E}_{C \sim U(\mathcal{C}(x))} \left[\min_{\xi^{(i)} \in \widehat{\Xi}} d(C, \xi^{(i)})\right]. \quad (9)$$

We use the shorthand $d(C, \Xi) := \min_{\xi^{(i)} \in \Xi} d(C, \xi^{(i)})$. In other words, we wish to construct representative points for a uniform sampling of the prediction region. A naive approach would simply involve explicitly gridding the output space $\mathcal{C}$, filtering such points with the rejection criterion of $\mathcal{C}(x)$, and clustering the remaining points per the d metric. This, however, is intractable in high-dimensional cases. Thus, a sampling method is employed to circumvent gridding, paralleling the technique leveraged for volume estimation.

M samples are initially drawn $\{c_{k,m}\}_{m=1}^M \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))$

for each k. Importantly, such uniform sampling of the balls leads to *non*-uniform sampling over $\mathcal{C}(x)$ if naively aggregated across k, as overlapped regions will be more densely sampled. For this reason, we subsample by discarding those samples $c_{k,m}$ for which $c_{k,m} \in V(\widehat{c}_{k'})$ for $k \neq k'$. This results in samples $C := \{c_i\}$ drawn from the desired $U(\mathcal{C}(x))$.

RPs must be aggregated separately for each connected subregion of $\Omega_\ell \subset \mathcal{C}(x)$ to ensure each $\xi^{(i)} \in \mathcal{C}(x)$. That is, we must identify $C_\ell := C \cap \Omega_\ell$. To do so, we determine if two points (c_i, c_j) belong to the same Ω_ℓ by considering the corresponding connected components problem defined on the graph induced by the edge criterion $e_{i,j} = \mathbb{1}[d(c_i, c_j) < \widehat{q}]$. For each C_ℓ, we find a subset $N_\ell := N(|C_\ell|/|C|)$ of the total N representative points, for which we use K-MEANS++ with the d metric. The full procedure is in Algorithm 3.

Algorithm 3 CPO-RPs: QUERYBALL$(\mathcal{T}, x, r)$ is an assumed subroutine that returns all points in the kd tree $\mathcal{T}$ that are within a radius r of x.

1: **procedure** CPO-RPs
   Inputs: Context x, CGM $q(C \mid X)$, RP count N, Conformal quantile $\widehat{q}$
2:    $\{\widehat{c}_k\}_{k=1}^K \sim q(C_{1:K} \mid x)$
3:    $\left\{\{c_{k,m}\}_{m=1}^M \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))\right\}_{k=1}^K$
4:    $C \leftarrow \{c_{k,m} \mid c_{k,m} \in V(\widehat{c}_k)\}_{k=1,m=1}^{K,M}$
5:    $\mathcal{T} \leftarrow$ KD-TREE(C)
6:    $\mathcal{E} \leftarrow \bigcup_i \{c_i \times$ QUERYBALL$(\mathcal{T}, c_i, \widehat{q}) \mid c_i \in \mathcal{T}\}$
7:    $\{C_\ell\} \leftarrow$ CONNECTEDCOMPONENTS$(\mathcal{G}(C, \mathcal{E}))$
8:    $\Xi \leftarrow \bigcup_{\ell=1}^L \{$K-MEANS++$(C_\ell, N\left(\frac{|C_\ell|}{|C|}\right), d)\}$
9:    **Return** Ξ
10: **end procedure**

3.6 CPO: Projection

After obtaining Ξ, further insight can be gleaned by exploring the local projection around each $\xi^{(i)}$. An example of this is visualizing the road-level variability in traffic predictions from uncertainty in upstream weather predictions, shown in Figure 5. To do this, we visualize the extent of the Voronoi cell $V^{(i)} \subset \mathcal{C}(x)$ associated with $\xi^{(i)}$ along the $\mathcal{C}$ space dimensions. That is, for each Voronoi cell, we visualize the Frechet variance along the projections $\{\pi_j\}_{j=1}^J$, where $J = \dim(\mathcal{C})$. Such projections preserve the structure of the objects being modeled, making them visually interpretable. For instance, π_j in the traffic example corresponds to the projection of $V^{(i)}$ to a *single* road j. Similarly, π_j would project to a single atom for a molecular reconstruction task. Formally,

$$\left|V_j^{(i)}\right| := \sum_{c \in V^{(i)}} d^2(\pi_j(c), \pi_j(\xi^{(i)})). \quad (10)$$

4 EXPERIMENT

We now demonstrate the utility of CPO with code at https://github.com/yashpatel5400/csi.

4.1 SBI: Fractional Knapsack

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

$$w^*(x) := \min_{w, \mathcal{U}} \max_{\widehat{c} \in \mathcal{U}(x)} \quad -\widehat{c}^T w \qquad (11)$$

$$\text{s.t.} w \in [0,1]^n, p^T w \leq B, \mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha,$$

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 (Hermans et al., 2021), 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, fully described in Appendix C. We note that, while these particular distributions have little semantic meaning in the traditional context of fractional knapsack, this experiment highlights the capacity for CPO to succeed even for complex distributions, which we leverage in a more semantically meaningful case in Section 4.2.

4.1.1 SBI: Quantitative Assessment

We first demonstrate the quantitative improvement in decision-making from leveraging CPO over the box-(PTC-B) and ellipsoid-based (PTC-E) regions proposed in (Sun, Liu, and Li, 2023), as well as box- and ellipsoid-based sets constructed based solely on observations of $\mathcal{P}(C)$, i.e. where we ignore x, referred to as Box and Ellipsoid. For CPO, we use

$$s(x, c) = \min_k ||\widehat{c}_k - c||_2^2. \qquad (12)$$

$q(\widehat{c} \mid x)$ was taken to be a neural spline normalizing flow (Durkan et al., 2019) trained with FAVI (Ambrogioni et al., 2019). Visualizations of the exact and variational posteriors are provided in Appendix E. Ks were chosen by studying the inflections of the prediction region volume estimate under each distributional setup, with $|\mathcal{D}_{\mathcal{C}_{1,2}}| = 1000$, seen in Figure 2. Inflection points were around $K = 10$ for most setups.

For assessing coverage and the robust objective value, we sampled $|\mathcal{D}_{\mathcal{T}}| = 1000$ test points i.i.d. from $\mathcal{P}(X, C)$. Coverage was assessed across all 1000 samples by measuring the proportion of samples for which $s(x^{(i)}, c^{(i)}) \leq \widehat{q}$. For assessing the objective, optimization was performed across 10 samples, with $p \sim U([0, 1000]^n)$, $u \sim U(0, 1)$, and $B \sim U(\max_i p_i, \sum_i p_i - u \max_i p_i)$ sampled per run.
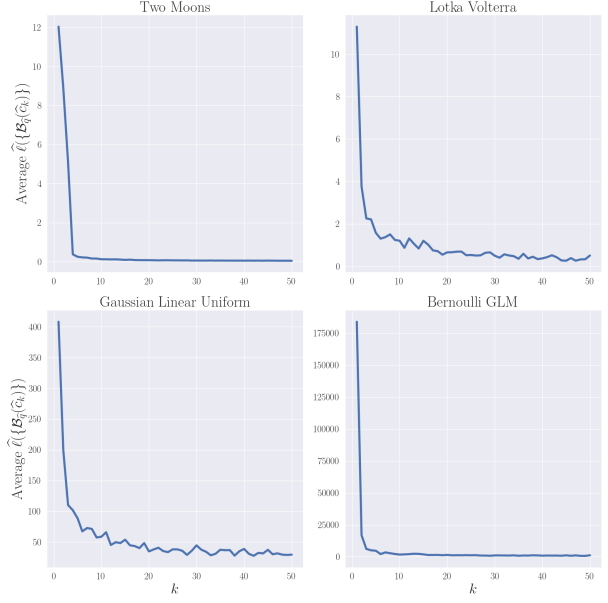

Figure 2: Average volume estimates $\widehat{\ell}(\{\mathcal{B}_{\widehat{q}}(\widehat{c}_k^{(i)})\})$ over $x^{(i)} \in \mathcal{D}_{\mathcal{C}_2}$ across SBI benchmarks.

The results are seen in Table 1. We include the nominal optima as a reference, i.e. $\min_w -c^T w$ for the *true* c. Recall that, by Lemma 3.1, with proper $\mathcal{U}(x)$, the robust objective values should be valid upper bounds on the nominal optima, with more conservative regions resulting in more vacuous bounds. We see this as, although all approaches result in valid coverage guarantees and hence produce valid upper bounds, the overly conservative nature of alternate regions results in their consistent looseness compared to CPO. Notably, these differences are more accentuated in cases where $\mathcal{P}(C|X)$ has complex structure; level sets under the Gaussian Linear, Gaussian Mixture, and Bernoulli GLM cases are roughly ellipsoidal, seen in Appendix E, resulting in comparable performance between CPO and PTC-E. Thus, as discussed and highlighted in Section 4.2, the benefits of CPO primarily manifest under difficult-to-model contextual distributions, where sets for simple geometries become overly large.

4.1.2 SBI: Representative Point Recovery

We next demonstrate that Algorithm 3 can approximately recover RPs for such uncertainty regions, leveraged to glean insights in the modeling task of Section 4.2. Notably, RPs are not unique; for instance, any rigid rotation of Ξ for a uniform distribution over a 2D ball results in a distinct yet optimal set $\widehat{\Xi}$ of RPs. The RP objective minimum, however, is unique, meaning suboptimality can be assessed by measuring

$$\Delta(\Xi, \widehat{\Xi}) := \mathbb{E}_{C \sim U(\mathcal{C}(x))} \left[d(C, \widehat{\Xi}) - d(C, \Xi) \right]. \qquad (13)$$

Table 1: Coverages across tasks for $\alpha = 0.05$ are shown in the left table, where coverage was assessed over a batch of 1,000 i.i.d. test samples. Objective optima are shown in the right table, averaged over a batch of 10 i.i.d. test samples with standard deviations in parentheses. The nominal optima are included as reference points.

	Box	PTC-B	Ellipsoid	PTC-E	CPO	Box	PTC-B	Ellipsoid	PTC-E	CPO	Nominal
Gaussian Uniform	0.94	0.96	0.95	0.95	0.95	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	**-0.27 (0.35)**	**-0.43 (0.4)**	*-4.48 (0.56)*
Gaussian Mixture	0.95	0.93	0.94	0.93	0.94	0.0 (0.0)	**-6.6 (1.67)**	0.0 (0.0)	**-7.38 (1.78)**	**-7.77 (1.87)**	*-11.66 (1.23)*
Bernoulli GLM	0.96	0.95	0.95	0.94	0.94	0.0 (0.0)	**-0.18 (0.49)**	0.0 (0.0)	**-0.06 (0.25)**	**-0.18 (0.37)**	*-3.53 (0.27)*
Lotka Volterra	0.95	0.96	0.94	0.94	0.95	**-0.52 (0.02)**	-0.05 (0.24)	-0.02 (0.0)	-0.22 (0.18)	**-0.68 (0.26)**	*-1.88 (0.01)*
SIR	0.94	0.95	0.93	0.95	0.93	-0.16 (0.02)	-0.22 (0.09)	-0.08 (0.01)	-0.22 (0.06)	**-0.38 (0.05)**	*-0.52 (0.02)*
Two Moons	0.93	0.94	0.94	0.94	0.96	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	**-0.15 (0.11)**	*-0.38 (0.01)*

$N = 5$ representative points were produced per setup. To compute Ξ, a grid discretization over the space was performed followed by a clustering for each connected component of this discretization. That is, the support $\mathcal{C}$ was discretized into 60 bins per dimension. Each discretized point c_k was assessed for membership in $\mathcal{C}(x)$, resulting in a collection of points C, from which we could recover Ξ in the manner described in Section 3.5. Visualizations of the exact and approximate RPs are provided for tasks where $\mathcal{C} \subset \mathbb{R}^2$ in Appendix F.
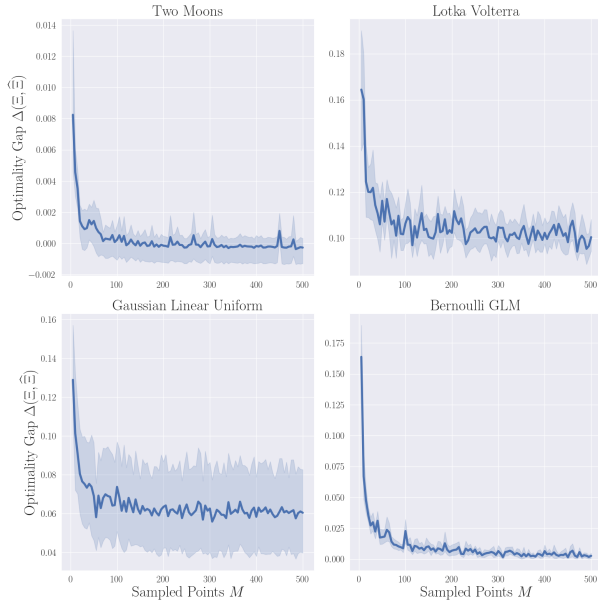

Figure 3: Suboptimality of the approximate representative points $\Delta(\Xi, \widehat{\Xi})$ decreases over increased sampling from the conformal prediction region.

To make explicit discretization possible, problems were projected into lower-dimensional versions, namely $\mathcal{C} \subset \mathbb{R}^4$. Figure 3 demonstrates the suboptimality of $\widehat{\Xi}$ decreases with increasing samples. Of note is that this convergence is slower in higher dimensional problems: for low dimensional cases, recovery of optimal RPs happens for small M, meaning any fluctuations thereafter are noise, as seen in the Two Moons case.

4.2 Robust Vehicle Routing

Optimal routing is a long-standing point of interest in the operations research community, with widespread applications such as in resource distribution and urban traffic flow management (Mor and Speranza, 2022; Saberi and Verbas, 2012; Okulewicz and Mańdziuk, 2019; Kořenář, 2003). We study the traffic flow problem from (Angelelli et al., 2021).

Recent work has demonstrated the utility of generative models in quantifying uncertainty for weather predictions over traditional physics-based approaches (Agrawal et al., 2019; Ayzel, Scheffer, and Heistermann, 2020; Franch et al., 2020; Shi et al., 2017). We specifically leverage a latent diffusion model for such forecasting from (Leinonen et al., 2023). Formally, a forecaster $\mathcal{P}(\widetilde{Y} \mid x)$ maps precipitation readings from radar networks $x \in \mathbb{R}^{T \times W \times H}$, specifically over T time steps with resolutions $W \times H$, to $\widetilde{Y} \in \mathbb{R}^{W \times H}$, the precipitation for some fixed ΔT point beyond x.

We consider the robust traffic flow problem (RTFP) for a source-target pair (s, t) over the network graph of Manhattan, where $|\mathcal{V}| = 4584$ and $|\mathcal{E}| = 9867$. The precipitation $\widetilde{Y}$ was combined with the nominal speed limits to obtain the final travel costs c along edges, fully described in Appendix G. Formally, we seek

$$w^*(x) := \min_{w} \max_{\widehat{c} \in \mathcal{U}(x)} \widehat{c}^T w \tag{14}$$

$$\text{s.t.} \, w \in [0,1]^{\mathcal{E}}, Aw = b, \mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha$$

where w_e represents the proportion of traffic routed along edge e, $C \in \mathbb{R}^{|\mathcal{E}|}$ is the edge weight vector, $A \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{E}|}$ is the node-arc incidence matrix, and $b \in \mathbb{R}^{|\mathcal{V}|}$ has entries $b_s = 1, b_t = -1$, and $b_k = 0$ for $k \notin \{s, t\}$.

We again demonstrate the quantitative improvement in decision-making resulting from using the more informative CPO prediction regions. Experiments were conducted with s and t chosen uniformly at random from $\mathcal{V}$. We take the score as defined in Equation 12 on the edge weight space rather than the initial precipitation map space. Results are shown in Table 2. Again, although all approaches achieve coverage guarantees, bounds resulting from alternate regions are sig-

Table 2: Coverage was assessed over 128 i.i.d. test samples and average objective optima over 10 i.i.d. test samples with standard deviations in parentheses.

	Box	PTC-B	Ellipsoid	PTC-E	CPO	*Nominal*
Coverage	0.94	0.93	0.94	0.92	0.94	—
Objective	7863.45 (0.0)	34559.03 (171.3)	7038.77 (0.0)	8807.68 (4.22)	**4171.22 (321.34)**	*299.50 (0.0)*

nificantly looser compared to those from CPO. This is especially prominent in this task compared to those of Section 4.1 due to the high dimension of the prediction space($\mathbb{R}^{|\mathcal{E}|}$) and complex nature of $\mathcal{P}(C|X)$.
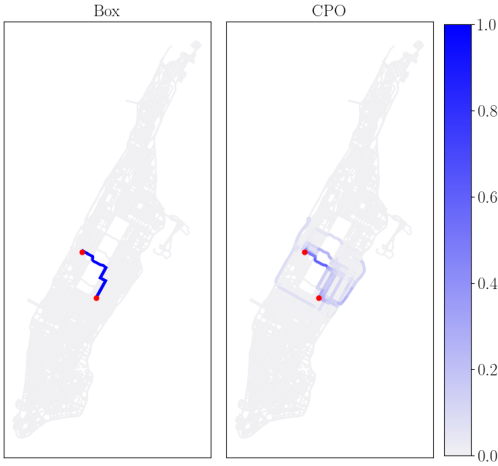
Figure 4: Solutions for the RTFP under the Box (left) and CPO (right) uncertainty regions.

Notably, the formulation in Equation 14 is a relaxation of the standard LP formulation of the robust shortest paths problem (RSPP), in which $\mathcal{W} = \{0,1\}^{\mathcal{E}}$. Given that A is a totally unimodular matrix, the solutions of the *box*-constrained RTFP and RSPP are equivalent, i.e. for both Box and PTC-B; they, however, are *not* equivalent under more general constraint sets (Chaerani, Roos, and Aman, 2005), i.e. Ellipsoid, PTC-E, and CPO, resulting in the observed suboptimality of box constraints. This is highlighted in Figure 4, where the Box constraint results in a fully concentrated allocation of traffic along a single path.

Despite apparent quantitative improvements resulting from the CPO optimal solution, it is difficult to directly understand *why* such allocations were deemed optimal without a qualitative impression of $\mathcal{U}(x)$, as framed in Section 3.5. We, therefore, now construct $N = 5$ representative points and their corresponding projections, two of which are visualized in Figure 5. The RPs highlight the multimodal nature of the edge weights distribution, where $\xi^{(1)}$ exhibits a case of precipitation more heavily concentrating along the north-

east corridor across Manhattan and $\xi^{(2)}$ one where it concentrates on the west. In addition, the projection around $\xi^{(2)}$ reveals especially high uncertainty on the path through Central Park with less on surrounding roads. CPO, thus, hedges its allocation in Figure 4 more evenly across paths, unlike the concentrated allocation under the Box region.
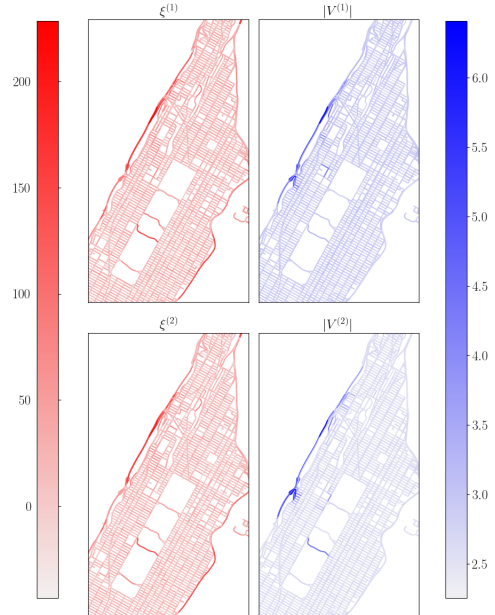

Figure 5: Two RPs for $\mathcal{C}(x)$ for travel time prediction (left) and the extents of their Voronoi cells (right).

5 DISCUSSION

We have presented CPO, a framework to leverage informative, non-convex conformal regions for predict-then-optimize decision-making. This work suggests many directions for future work. We are actively pursuing the extension of CPO to robust LQR control and subsequently to the broader category of robust control. Another interesting extension would be applications of CPO to discrete objects using GFlowNets for conditional sampling (Malkin et al., 2022; Hu et al., 2023). Finally, leveraging CPO over function spaces would enable its use to distributionally robust optimization.

References

[1] Shreya Agrawal et al. "Machine learning for precipitation nowcasting from radar images". In: *arXiv preprint arXiv:1912.12132* (2019).

[2] Luca Ambrogioni et al. "Forward amortized inference for likelihood-free variational marginalization". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 777–786.

[3] Enrico Angelelli et al. "System optimal routing of traffic flows with user constraints using linear programming". In: *European journal of operational research* 293.3 (2021), pp. 863–879.

[4] Anastasios N Angelopoulos and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". In: *arXiv preprint arXiv:2107.07511* (2021).

[5] Anastasios N Angelopoulos et al. "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 717–730.

[6] Georgy Ayzel, Tobias Scheffer, and Maik Heistermann. "RainNet v1. 0: a convolutional neural network for radar-based precipitation nowcasting". In: *Geoscientific Model Development* 13.6 (2020), pp. 2631–2644.

[7] Omer Belhasin et al. "Principal Uncertainty Quantification with Spatial Correlation for Image Restoration Problems". In: *arXiv preprint arXiv:2305.10124* (2023).

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

[9] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. "Data-driven robust optimization". In: *Mathematical Programming* 167 (2018), pp. 235–292.

[10] Hans-Georg Beyer and Bernhard Sendhoff. "Robust optimization–a comprehensive survey". In: *Computer methods in applied mechanics and engineering* 196.33-34 (2007), pp. 3190–3218.

[11] Geoff Boeing. "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks". In: *Computers, Environment and Urban Systems* 65 (2017), pp. 126–139.

[12] Diah Chaerani, Cornelis Roos, and A Aman. "The robust shortest path problem by means of robust linear optimization". In: *Operations Research Proceedings 2004: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR). Jointly Organized with the Netherlands Society for Operations Research (NGB) Tilburg, September 1–3, 2004*. Springer. 2005, pp. 335–342.

[13] Timothy M Chan. "A (slightly) faster algorithm for Klee's measure problem". In: *Proceedings of the twenty-fourth annual symposium on Computational geometry*. 2008, pp. 94–100.

[14] Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. "Data-driven conditional robust optimization". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 9525–9537.

[15] Meysam Cheramin et al. "Data-driven robust optimization using scenario-induced uncertainty sets". In: *arXiv preprint arXiv:2107.04977* (2021).

[16] Tore Dalenius. "The problem of optimum stratification". In: *Scandinavian Actuarial Journal* 1950.3-4 (1950), pp. 203–213.

[17] Tore Dalenius and Margaret Gurney. "The problem of optimum stratification. II". In: *Scandinavian Actuarial Journal* 1951.1-2 (1951), pp. 133–148.

[18] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[19] Conor Durkan et al. "Neural spline flows". In: *Advances in Neural Information Processing Systems* 32 (2019).

[20] Conor Durkan et al. *nflows: normalizing flows in PyTorch*. Version v0.14. Nov. 2020.

[21] H. Edelsbrunner. "The union of balls and its dual shape". In: *Discrete & Computational Geometry* 13.3 (1995), pp. 415–440.

[22] Adam N Elmachtoub and Paul Grigas. "Smart "predict, then optimize"". In: *Management Science* 68.1 (2022), pp. 9–26.

[23] Adam N Elmachtoub, Jason Cheuk Nam Liang, and Ryan McNellis. "Decision trees for decision-making under the predict-then-optimize framework". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2858–2867.

[24] Kai-Tai Fang and Jianxin Pan. "A Review of Representative Points of Statistical Distributions and Their Applications". In: *Mathematics* 11.13 (2023), p. 2930.

[25] Shai Feldman, Stephen Bates, and Yaniv Romano. "Calibrated multiple-output quantile regression with representation learning". In: *Journal of Machine Learning Research* 24.24 (2023), pp. 1–48.

[26] George Fishman. *Monte Carlo: concepts, algorithms, and applications*. Springer Science & Business Media, 2013.

[27] Bernard D Flury and Thaddeus Tarpey. "Representing a large collection of curves: A case for principal points". In: *The American Statistician* 47.4 (1993), pp. 304–306.

[28] Gabriele Franch et al. "Precipitation nowcasting with orographic enhanced stacked generalization: Improving deep learning predictions on extreme events". In: *Atmosphere* 11.3 (2020), p. 267.

[29] Joeri Hermans et al. "Averting a crisis in simulation-based inference". In: *arXiv preprint arXiv:2110.06581* (2021).

[30] Eliahu Horwitz and Yedid Hoshen. "Conffusion: Confidence intervals for diffusion models". In: *arXiv preprint arXiv:2211.09795* (2022).

[31] Edward J Hu et al. "GFlowNet-EM for learning compositional latent variable models". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 13528–13549.

[32] Yuge Hu et al. "Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials". In: *Machine Learning: Science and Technology* 3.4 (2022), p. 045028.

[33] Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. "Cd-split and hpd-split: Efficient conformal regions in high dimensions". In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 3772–3803.

[34] Chancellor Johnstone and Bruce Cox. "Conformal uncertainty sets for robust optimization". In: *Conformal and Probabilistic Prediction and Applications*. PMLR. 2021, pp. 72–90.

[35] Margot E Kaminski. "The right to explanation, explained". In: *Berkeley Technology Law Journal* 34.1 (2019), pp. 189–218.

[36] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[37] Václav Kořenář. "Vehicle routing problem with stochastic demands". In: *ALLOCATION FRAGMENTS OF THE DISTRIBUTED DATABASE* (2003), p. 24.

[38] Jussi Leinonen et al. "Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification". In: *arXiv preprint arXiv:2304.12891* (2023).

[39] Jan-Matthis Lueckmann et al. "Benchmarking simulation-based inference". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 343–351.

[40] Nikolay Malkin et al. "GFlowNets and variational inference". In: *arXiv preprint arXiv:2210.00580* (2022).

[41] Huiying Mao, Ryan Martin, and Brian J Reich. "Valid model-free spatial prediction". In: *Journal of the American Statistical Association* (2022), pp. 1–11.

[42] Joel Max. "Quantizing for minimum distortion". In: *IRE Transactions on Information Theory* 6.1 (1960), pp. 7–12.

[43] Andrea Mor and Maria Grazia Speranza. "Vehicle routing problems over time: a survey". In: *Annals of Operations Research* 314.1 (2022), pp. 255–275.

[44] Mervin E Muller. "A note on a method for generating points uniformly on n-dimensional spheres". In: *Communications of the ACM* 2.4 (1959), pp. 19–20.

[45] Shunichi Ohmori. "A predictive prescription using minimum volume k-nearest neighbor enclosing ellipsoid and robust optimization". In: *Mathematics* 9.2 (2021), p. 119.

[46] Michał Okulewicz and Jacek Mańdziuk. "A metaheuristic approach to solve dynamic vehicle routing problem in continuous search space". In: *Swarm and Evolutionary Computation* 48 (2019), pp. 44–61.

[47] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems* 32 (2019).

[48] Yash Patel et al. "Variational Inference with Coverage Guarantees". In: *arXiv preprint arXiv:2305.14275* (2023).

[49] Egon Peršak and Miguel F Anjos. "Contextual robust optimisation with uncertainty quantification". In: *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*. Springer. 2023, pp. 124–132.

[50] Meead Saberi and İ Ömer Verbas. "Continuous approximation model for the vehicle routing problem for emissions minimization at the strategic level". In: *Journal of Transportation Engineering* 138.11 (2012), pp. 1368–1376.

[51] Utsav Sadana et al. "A Survey of Contextual Optimization Methods for Decision Making under Uncertainty". In: *arXiv preprint arXiv:2306.10374* (2023).

[52] Matteo Sesia and Emmanuel J Candès. "A comparison of some conformal quantile regression methods". In: *Stat* 9.1 (2020), e261.

[53] Matteo Sesia and Yaniv Romano. "Conformal prediction using conditional histograms". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 6304–6315.

[54] Glenn Shafer and Vladimir Vovk. "A Tutorial on Conformal Prediction." In: *Journal of Machine Learning Research* 9.3 (2008).

[55] Chao Shang and Fengqi You. "A data-driven robust optimization approach to scenario-based stochastic model predictive control". In: *Journal of Process Control* 75 (2019), pp. 24–39.

[56] Xingjian Shi et al. "Deep learning for precipitation nowcasting: A benchmark and a new model". In: *Advances in neural information processing systems* 30 (2017).

[57] Chunlin Sun, Linyu Liu, and Xiaocheng Li. "Predict-then-Calibrate: A New Perspective of Robust Contextual LP". In: *arXiv preprint arXiv:2305.15686* (2023).

[58] Renukanandan Tumu et al. "Physics constrained motion prediction with uncertainty quantification". In: *arXiv preprint arXiv:2302.01060* (2023).

[59] Zhendong Wang et al. "Probabilistic conformal prediction using conditional random samples". In: *arXiv preprint arXiv:2206.06584* (2022).

[60] Yachong Yang and Arun Kumar Kuchibhotla. "Finite-sample efficient conformal prediction". In: *arXiv preprint arXiv:2104.13871* (2021).

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No, will be released upon acceptance]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Prediction Region Validity Lemma

Lemma A.1. *Consider any $f(w, c)$ that is L-Lipschitz in c under the metric d for any fixed w. Assume further that $\mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha$. Then,*

$$\mathcal{P}_{X,C}\left(\Delta(X, C) \leq L \operatorname{diam}(\mathcal{U}(X))\right) \geq 1 - \alpha. \tag{15}$$

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

$$\min_w \max_{\widehat{c} \in \mathcal{U}(x)} f(w, \widehat{c}) - \min_w f(w, c)$$
$$\leq \max_w |\max_{\widehat{c} \in \mathcal{U}(x)} f(w, \widehat{c}) - f(w, c)|$$
$$\leq L \max_{\widehat{c} \in \mathcal{U}(x)} d(\widehat{c}, c) \leq L \operatorname{diam}(\mathcal{U}(x)).$$

Since we have the assumption that $\mathcal{P}(C \in \mathcal{U}(X)) \geq 1 - \alpha$, the result immediately follows. $\qquad \square$

B Optimization Convergence Lemma

We first begin by citing a standard result of projected gradient descent, from which the result of interest immediately follows.

Lemma B.1. *Let K be a closed convex set, and $f : K \to \mathbb{R}$ be convex, differentiable, and L-Lipschitz. Let $x^* \in K$ be a minimizer of f, and define $T := \frac{L^2 ||x_0 - x^*||}{\epsilon^2}$ and $\eta := \frac{||x_0 - x^*||}{L\sqrt{T}}$. Then the iterates $\{x_t\}_{t=0}^{T}$ returned by projected gradient descent satisfy*

$$f\left(\frac{1}{T+1} \sum_{t=0}^{T} x_t\right) - f(x^*) \leq \epsilon. \tag{16}$$

Lemma B.2. *Let $\phi(w) := \max_{\widehat{c} \in \bigcup_{k=1}^{K} \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c})$ for $\{\widehat{c}_k\}_{k=1}^{K} \subset \mathcal{C}$, $\widehat{q} \in \mathbb{R}^+$, and $f(w, c)$ convex-concave and L-Lipschitz in c for any fixed w. Let $w^* \in \mathcal{W}$ be a minimizer of ϕ. For any $\epsilon > 0$, define $T := \frac{L^2 ||w_0 - w^*||}{\epsilon^2}$ and $\eta := \frac{||w_0 - w^*||}{L\sqrt{T}}$. Then the iterates $\{w_t\}_{t=0}^{T}$ returned by Algorithm 1 satisfy*

$$\phi\left(\frac{1}{T+1} \sum_{t=0}^{T} w_t\right) - \phi(w^*) \leq \epsilon. \tag{17}$$

Proof. Notice that $\phi(w)$ is convex by Danskin's Theorem by assumption of the convexity of f in w. By Danskin's Theorem, $\nabla_w \phi(w) = \nabla_w f(w, c^*)$, where $c^* := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$. Further notice

$$\phi(w) := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c}) = \max_k \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}). \tag{18}$$

Denote $\phi_k(w) := \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c})$. Clearly, $\phi_k(w)$ is L-Lipschitz by assumption on the structure of f. Further, as the point-wise maximum of L-Lipschitz functions is itself L-Lipschitz, it follows that $\phi(w) = \max_k \phi_k(w)$ is also L-Lipschitz. The conclusion, thus, follows by applying Lemma B.1 to $\phi(w)$. $\qquad \square$

C Simulation-Based Inference Benchmarks

The benchmark tasks are a subset of those provided by (Lueckmann et al., 2021). For convenience, we provide brief descriptions of the tasks curated by this library; however, a more comprehensive description of these tasks can be found in their manuscript.

C.1 Gaussian Linear

10-dimensional Gaussian model with a Gaussian prior:

$$\textbf{Prior}: \mathcal{N}(0, 0.1 \odot I)$$
$$\textbf{Simulator}: x \mid w \sim \mathcal{N}(x \mid w, 0.1 \odot I)$$

C.2 Gaussian Linear Uniform

10-dimensional Gaussian model with a uniform prior:

$$\textbf{Prior}: \mathcal{U}(-1, 1)$$
$$\textbf{Simulator}: x \mid w \sim \mathcal{N}(x \mid w, 0.1 \odot I)$$

C.3 SLCP with Distractors

Simple Likelihood Complex Posterior (SLCP) with Distractors has uninformative dimensions in the observation over the standard SLCP task:

$$\textbf{Prior}: \mathcal{U}(-3, 3)$$
$$\textbf{Simulator}: x \mid w = p(y) \text{ where } p \text{ reorders}$$
$$y \text{ with a fixed random order}$$
$$y_{[1:8]} \sim \mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \begin{bmatrix} w_3^4 & w_3^2 w_4^2 \tanh(w_5) \\ w_3^2 w_4^2 \tanh(w_5) & w_4^4 \end{bmatrix} \right),$$
$$y_{9:100} \sim \frac{1}{20} \sum_{i=1}^{20} t_2(\mu^i, \Sigma^i), \mu^i \sim \mathcal{N}(0, 15^2 I),$$
$$\Sigma_{j,k}^i \sim \mathcal{N}(0, 9), \Sigma_{j,j}^i = 3e^a, a \sim \mathcal{N}(0, 1),$$

C.4 Bernoulli GLM Raw

10-parameter GLM with Bernoulli observations and Gaussian prior. Observations are not sufficient statistics, unlike the standard "Bernoulli GLM" task:

$$\textbf{Prior}: \beta \sim \mathcal{N}(0, 2), f \sim \mathcal{N}(0, (F^T F)^{-1})$$
$$F_{i,i-2} = 1, F_{i,i-1} = -2$$
$$F_{i,i} = 1 + \sqrt{\frac{i-1}{9}}, F_{i,j} = 0; i \leq j$$
$$\textbf{Simulator}: x^{(i)} \mid w \sim \text{Bern}(\eta(v_T^{(i)} f + \beta)),$$
$$\eta(\odot) = \exp(\odot)/(1 + \exp(\odot))$$

C.5 Gaussian Mixture

A mixture of two Gaussians, with one having a much broader covariance structure:

$$\textbf{Prior}: \beta \sim \mathcal{U}(-10, 10)$$
$$\textbf{Simulator}: x \mid w \sim 0.5\mathcal{N}(x \mid w, I) + 0.5\mathcal{N}(x \mid w, .01I)$$

C.6 Two Moons

Task with a posterior that has both global (bimodal) and local (crescent-shaped) structure:

$$\textbf{Prior}: \beta \sim \mathcal{U}(-1, 1)$$
$$\textbf{Simulator}: x \mid w =$$
$$\begin{bmatrix} r\cos(\alpha) + 0.25 \\ r\sin(\alpha) \end{bmatrix} + \begin{bmatrix} -|w_1 + w_2|/\sqrt{2} \\ (-w_1 + w_2)/\sqrt{2} \end{bmatrix}$$
$$\alpha \sim \mathcal{U}(-\pi/2, \pi/2), r \sim \mathcal{N}(0.1, 0.01^2)$$

C.7 SIR

Epidemiology model with S (susceptible), I (infected), and R (recovered). A contact rate β and mean recovery rate of γ are used as follows:

$$\textbf{Prior}: \beta \sim \text{LogNormal}(\log(0.4), 0.5),$$
$$\gamma \sim \text{LogNormal}(\log(1/8), 0.2)$$
$$\textbf{Simulator}: x = (x^{(i)})_{i=1}^{10}; x^{(i)} \mid w \sim \text{Bin}(1000, \frac{I}{N}),$$
$$\text{where } I \text{ is simulated from:}$$
$$\frac{dS}{dt} = -\beta\frac{SI}{N}, \qquad \frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I, \qquad \frac{dR}{dt} = \gamma I$$

C.8 Lotka-Volterra

An ecological model commonly used in describing dynamics of competing species. w parameterizes this interaction as $w = (\alpha, \beta, \gamma, \delta)$:

$$\textbf{Prior}: \alpha \sim \text{LogNormal}(-.125, 0.5)$$
$$\beta \sim \text{LogNormal}(-3, 0.5), \gamma \sim \text{LogNormal}(-.125, 0.5)$$
$$\delta \sim \text{LogNormal}(-3, 0.5)$$
$$\textbf{Simulator}: x = (x^{(i)})_{i=1}^{10},$$
$$x_{1,i} \mid w \sim \text{LogNormal}(\log(X), 0.1),$$
$$x_{2,i} \mid w \sim \text{LogNormal}(\log(Y), 0.1)$$
$$\text{where } X, Y \text{ is simulated from:}$$
$$\frac{dX}{dt} = \alpha X - \beta XY, \qquad \frac{dY}{dt} = -\gamma Y + \delta XY$$

D Training Details

All encoders were implemented in PyTorch (Paszke et al., 2019) with a Neural Spline Flow architecture. The NSF was built using code from (Durkan et al., 2020). Specific architecture hyperparameter choices were taken to be the defaults from (Durkan et al., 2020) and are available in the code. Optimization was done using Adam (Kingma and Ba, 2014) with a learning rate of 10^{-3} over 5,000 training steps. Minibatches were drawn from the corresponding prior $\mathcal{P}(Y)$ and simulator $\mathcal{P}(X \mid Y)$ as specified per task in the preceding section. Training these models required between 10 minutes and two hours using an Nvidia RTX 2080 Ti GPUs for each of the SBI tasks.
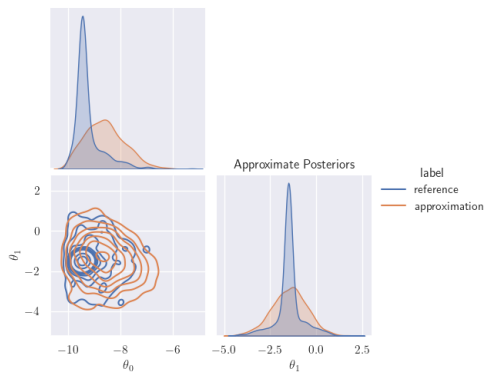
E Posteriors

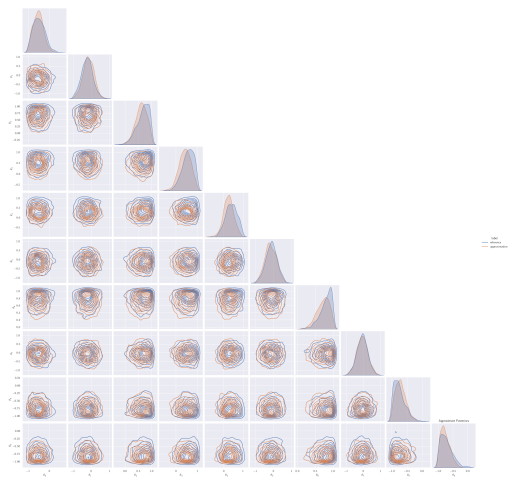We provide visualizations of approximate and reference posteriors (produced with MCMC from (Lueckmann et al., 2021)).
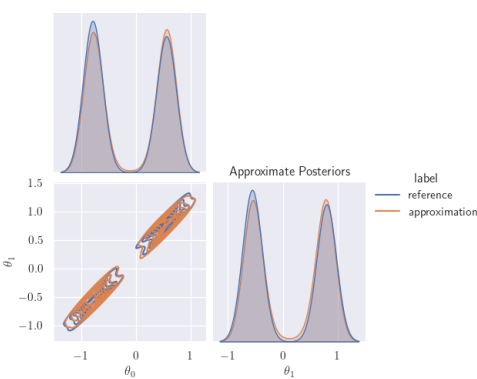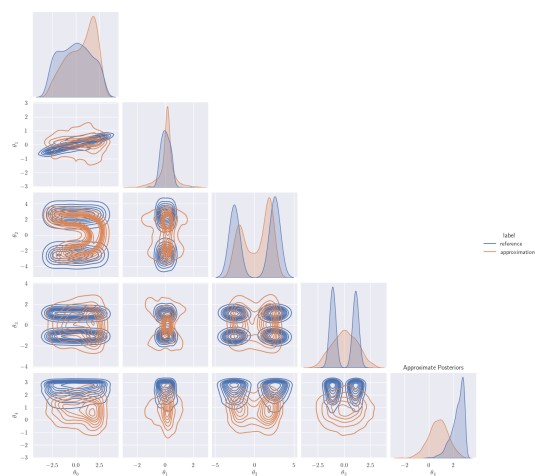
E.1 Gaussian Linear
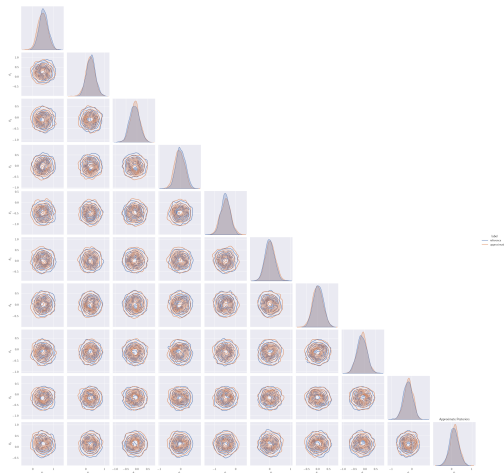
E.2 Gaussian Mixture
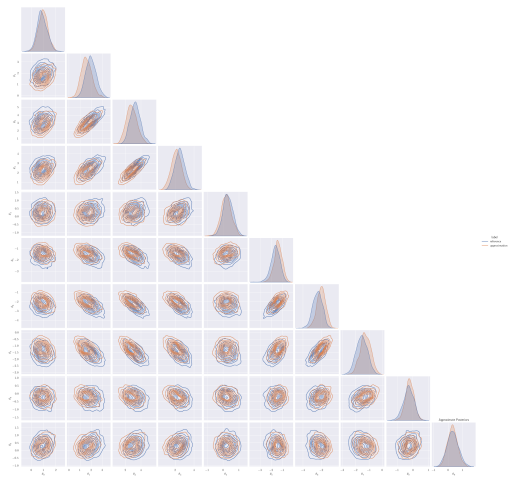
E.3 Gaussian Linear Uniform

E.4 Two Moons

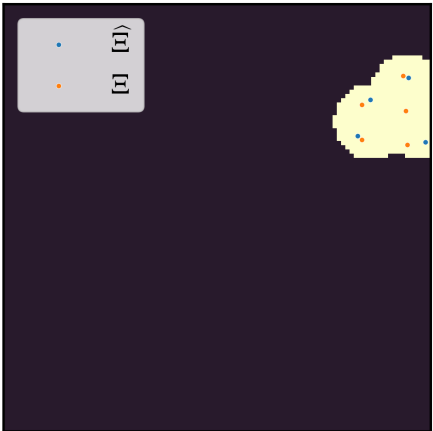

E.5 SLCP

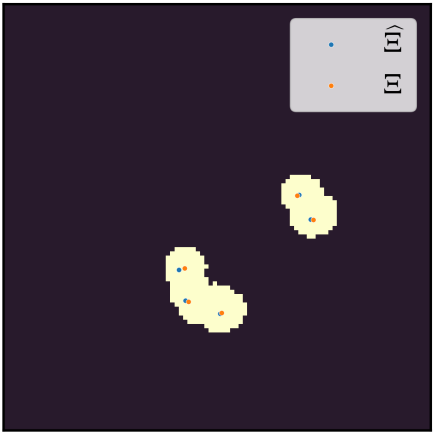### E.6   Bernoulli GLM

F   SBI Representative Points

F.1   Gaussian Mixture
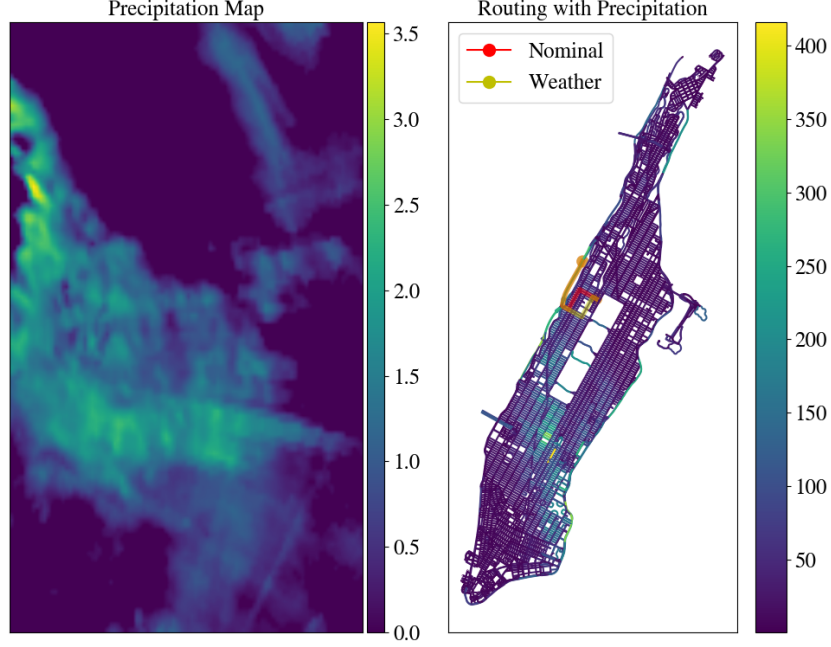

F.2   Two Moons

Figure 6: Precipitation maps (left) are converted to edge weights (right) as per Equation 19. Solving the shortest paths problem (SPP) on this newly weighted graph, therefore, can produce distinct routes from that based on the nominal travel-time SPP, as highlighted by the two distinct paths under the nominal and weather-weighted graphs on the right.

G   Robust Vehicle Routing Setup

The routing graph of Manhattan was extracted using OSMnx, with local highway speeds extracted using Open-StreetMap (Boeing, 2017). Highway speed imputation was performed on edges where such information was not available, specifically by averaging over those highways of comparable categorization, namely "residential," "secondary," or "tertiary." Doing so defined a nominal travel cost $\widetilde{c}$.

We now wish to modify these nominal travel costs to account for the weather predictions made upstream. That is, we wish to account for the precipitation map $\widetilde{Y} \in \mathbb{R}^{W \times H}$ in these edge weights. To do so, we use the global coordinates $(c_x^v, c_y^v) \in \mathbb{R}^2$ of each $v \in \mathcal{V}$ to find the precipitation at the corresponding location. Concretely, we determine the pixel coordinate by scaling the coordinate to the range of the region that was forecasted. So, for a forecast over the window $(c_x^{\min}, c_x^{\max}) \times (c_y^{\min}, c_y^{\max})$, the corresponding pixel lookup is:

$$p_x^v = \lfloor \frac{c_x^v - c_x^{\min}}{c_x^{\max} - c_x^{\min}} \rfloor \times W \qquad p_y^v = \lfloor \frac{c_y^v - c_y^{\min}}{c_y^{\max} - c_y^{\min}} \rfloor \times H.$$

The corresponding precipitation associated with each vertex, therefore, is $\widetilde{Y}_{p_x^v, p_y^v}$. We define the final travel cost for each edge $e \in \mathcal{E}$ with endpoints (e_s, e_t) as:

$$c_e := \widetilde{c}_e \cdot \exp \left\{ \frac{\widetilde{Y}_{p_x^{e_s}, p_y^v} + \widetilde{Y}_{p_x^{e_t}, p_y^{e_t}}}{2} \right\}. \tag{19}$$

We then solve SPP on the weighted directed graph with edge weights c_e. An example of this weighting and the corresponding shortest path is illustrated in Figure 6.