Non-Parameteric Conformal Distributionally Robust Optimization

Anonymous Authors[1]

Abstract

Simulation-based inference leverages amortized variational inference algorithms to perform posterior estimation in scientific domains, often over hundreds or thousands of observations. Such estimated posteriors are often subsequently leveraged in downstream estimation or engineering design. The use of approximated posteriors in these downstream applications, however, ultimately produces results that could be arbitrarily poorly behaved with posterior misspecification. While MCMC could be used to combat this misspecification, doing so limits the number of designs that can be considered within a typical computational budget, translating to lost design efficiency. Toward this end, we propose a distributionally robust formulation, where the problem formulation is specified in a data-driven manner, thereby producing downstream guarantees of interest. In particular, we propose Conformalized Distributionally Robust Optimization (CRDO), a procedure that leverages conformal prediction over the space of *distributions* to produce strong theoretical guarantees on the well-specified problem setup. We then demonstrate that our framework lends itself to an efficient algorithm that we then subsequently highlight on a suite of benchmark problems.

1. Introduction

With increasing compute budget and simulator fidelity, there is growing interest in leveraging simulators to do inference in scientific domains, such as in astrophysics, neuroscience, and particle physics (Cranmer et al., 2020; Zhou et al., 2023; Crisostomi et al., 2023). In these cases, scientific knowledge endows domain experts to specify prior and forward models with great precision, concentrating concern on the ability to then exactly recover the posterior distributions. While MCMC offers a theoretically justified approach to perform such sampling, often posterior distributions $\mathcal{P}(\Theta \mid x)$ are sought over a large collection of x, on the order of 10,000 or more, rendering MCMC computationally intractable.

For this reason, variational inference has increasingly become the de facto approach of posterior estimation in simulation-based inference (Papamakarios & Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019; Deistler et al., 2022; Papamakarios et al., 2019; Boelts et al., 2022). While efforts have gone into improving the calibration of such posteriors (Deistler et al., 2022; Delaunoy et al., 2022; Lemos et al., 2023; Delaunoy et al., 2023), approaches continue to exhibit a lack of consistent calibration, as highlighted in a recent meta-study of likelihood-free inference algorithms (Hermans et al., 2021).

The Θ space in such simulation-based inference settings often parameterizes a dynamical system, i.e. $\dot{x} = f_\theta(x)$. Inference for a queried x, therefore, can be used to parameterize a surrogate dynamics model in engineering workflows, where x characterizes a design of interest (Nguyen et al., 2023a;b; Gupta & Brandstetter, 2022; Shen et al., 2023). For instance, for a car design x, a surrogate model with the parameters $\mathcal{P}(\Theta \mid x)$ would be used for evaluation.

Naively using the posterior distributions produced via variational inference for these design problems, however, results in suboptimal decision-making, as they are generally misspecified. For this reason, several methods of doing robust optimization design have been proposed. For instance, in the space of computational fluid dynamics, robustness is accounted for using Monte Carlo methods or polynomial chaos (Wu et al., 2018; Li et al., 2022; Lee et al., 2020; Liu et al., 2022). Another approach that has become of interest recently is one that formulates the problem as a distributionally robust optimization (DRO) problem, in which solutions of this optimization set are instead sought over an ambiguity set $\mathcal{U}(\mathcal{P})$ of distributions (Chen et al., 2023). DRO, however, requires a priori knowledge of plausible ambiguity sets or noise distributions to produce answers that are practically useful. Towards this end, data-driven DRO has recently become of interest, in which plausible ambiguity sets are learned empirically (Delage & Ye, 2010; Mohajerin Esfahani & Kuhn, 2018; Chen et al., 2022). While these often offer improved empirical performance, they are typically

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SPIGM workshop at ICML 2024. Do not distribute.

specified in an ad-hoc manner, rendering any downstream guarantees thereof nonexistent.

Conformal prediction provides a principled framework for producing distribution-free uncertainty quantification with marginal frequentist guarantees (Angelopoulos & Bates, 2021; Shafer & 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 $\mathcal{D}_C$, 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 work leverages such conformal prediction regions for predict-then-optimize decision-making (Patel et al., 2023). In this vein, we propose Conformal Distributional Predict-Then-Optimize (CDPO), a procedure that leverages conformal prediction to produce prediction regions over probability measures and thereby produces guarantees on solutions of stochastic decision-making problems that rely on amortized variational inference, such as design optimization (Bird et al., 2023; Azad & Herber, 2023; 2022). Our main contributions are:

- Proposing a framework for data-driven distributionally robust optimization that has downstream guarantees.
- Demonstrating the use of conformal prediction over arbitrarily specified probability distributions.

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 (Angelopoulos & Bates, 2021; Shafer & Vovk, 2008) 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 $\mathcal{P}(Y \notin \mathcal{C}(X)) \leq \alpha$.

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 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 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 $\mathcal{C}(x) = \{y \mid s(x, y) \leq \widehat{q}(\alpha)\}$ 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 $\mathcal{S}$. 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 (Shafer & Vovk, 2008), 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 (Yang & Kuchibhotla, 2021; Sesia & Candès, 2020). It is defined as the inverse of the expected Lebesgue measure of the prediction region, i.e. $(\mathbb{E}[|\mathcal{C}(X)|])^{-1}$. 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 $\mathcal{P}(Y \mid X)$, typically using either MCMC or VI. VI has arisen 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 a minimizer $\varphi^*(X) = \arg\min_\varphi D(q_\varphi(Y) || \mathcal{P}(Y \mid X))$, 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 (Rezende & Mohamed, 2015; Agrawal et al., 2020). Amortized variational inference expands on this approach by training a neural network to approximate $\varphi^*(X)$. This leads to a variational posterior approximator $q(Y \mid X) = q_{\varphi^*(X)}(Y)$ that can be rapidly computed for any value X.

2.3. Distributionally Robust Optimization

Distributionally robust optimization (DRO) is a broadly applied framework for framing problems under the ambiguity of distributional specification. We specifically focus on its application to stochastic optimization problems, as discussed in (Gao & Kleywegt, 2023; Bertsimas et al., 2019). For a broader discussion of DRO, we refer readers to (Kuhn et al., 2019). Formally, DRO problems are formulated with the following min-max objective:

$$w^* := \inf_{w \in \mathcal{W}} \sup_{\mathcal{Q} \in \mathcal{B}_q(\mathcal{P})} \mathbb{E}_{\mathcal{Q}(C)} \left[C^T w \right], \tag{1}$$

where w are decision variables, C an *unknown* cost parameter, $\mathcal{W}$ a compact feasible region, and $\mathcal{B}_q(\mathcal{P})$ is a ball in the space of probability distributions of radius q under some prespecified probability metric around a base distribution $\mathcal{P}$. In practice, DRO is often specified using a Wasserstein probability metric. The p-Wasserstein distance for any $p \in [1, \infty]$ between two probability $\mathcal{Q}$ and $\mathcal{Q}'$ defined over $\mathbb{R}^m$ is defined as:

$$W_p^p(\mathcal{Q}, \mathcal{Q}') = \inf_{\pi \in \Pi(\mathcal{Q}, \mathcal{Q}')} \int_{\mathbb{R}^m \times \mathbb{R}^m} \|\xi - \xi'\|^p \pi(d\xi, d\xi')$$

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

In practice, the base distribution is often taken to simply be an empirical distribution $\widehat{P}_M$ defined over M samples drawn $c_i \sim \mathcal{P}$, i.e. $\widehat{P}_M := \frac{1}{M} \sum_{i=1}^M \delta_{c_i}$. In this case, Equation (1) lends itself to a convex reformulation, which can then be solved efficiently using standard convex optimization techniques, specifically as follows:

$$\inf_{w \in \mathcal{W}} \left(\frac{1}{M} \sum_{i=1}^M f(w^\top c_i) + \widehat{q} \cdot \mathrm{Lip}(f) \cdot ||w||_* \right), \tag{2}$$

where $\mathrm{Lip}(f)$ denotes the Lipschitz constant of f and $||\cdot||_*$ denotes the dual norm to the p-norm with respect to which the Wasserstein probability metric was defined.

2.4. Predict-Then-Optimize

Predict-then-optimize problems are formulated as

$$w^*(x) := \min_{w \in \mathcal{W}} \quad \mathbb{E}[C^T w \mid x], \tag{3}$$

where x is an observed context. The predict-then-optimize framework is so-called as the unknown C is typically first predicted from observed contextual variables x. That is, a predictive contextual distribution $\mathcal{P}(C \mid x)$ is assumed, with respect to which the optimization formulation is defined. A full review is presented in (Elmachtoub & Grigas, 2022).

3. CDPO

3.1. Problem Setup

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 (Esteban-Pérez & Morales, 2022), 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.

We additionally assume this setting lends itself to downstream stochastic optimization problems over $\mathcal{P}(C \mid X)$. For example, x may be parametric properties of a car design, such as its chassis length or tire placement, and c the predictions of a parametric fluids surrogate model, such as Reynolds Averaged Navier Stokes. An objective of interest in this case could then be an optimal control scheme w^*, with the cost f corresponding to fuel efficiency. This problem is formalized in the following section.

3.2. 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

$$\begin{aligned} w^*(x) := \inf_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} \quad & \mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)] \\ \text{s.t.} \quad & \mathcal{P}_{X, \mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha, \end{aligned} \tag{4}$$

where $\mathcal{U} : \mathcal{X} \to \mathcal{M}(\mathcal{F})$ is an 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:

$$s(x, \mathcal{P}_C) = \mathcal{W}_1(\mathcal{Q}_{\varphi(x)}(C), \mathcal{P}_C), \tag{5}$$

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 (Peyré et al., 2019).

With this choice of score function, we can characterize the suboptimality gap $\Delta(x, \mathcal{P}_C)$, defined to be:

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

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, shown below. The full proof of this statement is deferred to Appendix A.

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 $\mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha$ with $\sup_{\widetilde{\mathcal{Q}} \in \mathcal{U}(x)} \mathcal{W}_1(\widetilde{\mathcal{Q}}, \mathcal{P}_C) = \operatorname{diam}(\mathcal{U}(x))$. Then,*

$$\mathcal{P}_{X,\mathcal{P}_C}\left(\Delta(X, \mathcal{P}_C) \leq L \operatorname{diam}(\mathcal{U}(X))\right) \geq 1 - \alpha. \quad (6)$$

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.

3.3. Optimization Algorithm

While the statement of Lemma 3.1 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 (5), 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:

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

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:

$$\mathcal{W}_1(\widetilde{\mathcal{Q}}(C \mid x_i), \mathcal{P}(C \mid x_i)) \quad (7)$$
$$\approx \mathcal{W}_1(\widehat{\mathcal{Q}}(C \mid x_i), \widehat{\mathcal{P}}(C \mid x_i)) = \inf_{\pi} \sum_{j=1}^{M} \left| c_j^{\mathcal{Q}} - c_{\pi(j)}^{\mathcal{P}} \right|,$$

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 (Feldman et al., 2023a). We defer this discussion to future work.

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 (4) 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:

$$\inf_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{B}_{\widehat{q}}(\widehat{\mathcal{Q}})} \mathbb{E}_{\widetilde{\mathcal{Q}}}[f(w, C)]. \quad (8)$$

We now leverage the insights of (Kuhn et al., 2019) 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:

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

where $c_i^{\mathcal{Q}}$ are 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.

4. Experiment

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

$$\inf_{w \in \mathcal{W}} \sup_{\widetilde{\mathcal{Q}} \in \mathcal{B}_{\widehat{q}}(\widehat{\mathcal{Q}})} \mathbb{E}_{\widetilde{\mathcal{Q}}}[-w^\top C] \quad (9)$$
$$\text{s.t.} w \in [0, 1]^n, p^T w \leq B, \mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha.$$

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. K reference posteriors were provided by the authors of (Hermans et al., 2021) 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:

$$\inf_{w \in \mathcal{W}} \left(\frac{1}{M} \sum_{i=1}^{M} -w^\top c_i^{\mathcal{Q}} + \widehat{q} \cdot ||w||_\infty \right) \quad (10)$$
$$\text{s.t.} w \in [0, 1]^n, p^T w \leq B, \mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha,$$

where we note that $f := \mathrm{id}$ has a Lipschitz constant of $\operatorname{Lip}(f) = 1$. We then compute $\widehat{q}$ over the N_C reference-variational posterior pairs taking $\alpha = 0.9$, where $N_C = 10$ in (Hermans et al., 2021). Solving this problem, by Lemma 3.1 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, $\Delta_{\%} = \mathbb{E}_X[\Delta(X, C(X)) / \min_w f(w, C(X))]$.

Table 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)

5. Discussion

We have proposed a new methodology to formulate robust predict-then-optimize problems with distributional misspecification in a principled manner along with an approach to solve such problems. This preliminary work suggests many paths forward that would be of interest. The most immediate is formally demonstrating that the coverage guarantees are retained under estimation of the Wasserstein distance; a similar result was established in (Feldman et al., 2023b). Additionally, comparing this to alternate data-driven approaches beyond simple synthetic settings would be of interest.

References

Agrawal, A., Sheldon, D. R., and Domke, J. Advances in black-box vi: Normalizing flows, importance weighting, and optimization. *Advances in Neural Information Processing Systems*, 33:17358–17369, 2020.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Azad, S. and Herber, D. R. Control co-design under uncertainties: formulations. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 86229, pp. V03AT03A008. American Society of Mechanical Engineers, 2022.

Azad, S. and Herber, D. R. An overview of uncertain control co-design formulations. *Journal of Mechanical Design*, 145(9):091709, 2023.

Bertsimas, D., Sim, M., and Zhang, M. Adaptive distributionally robust optimization. *Management Science*, 65(2): 604–618, 2019.

Bird, T. J., Siefert, J. A., Pangborn, H. C., and Jain, N. A set-based approach for robust control co-design. *arXiv preprint arXiv:2310.11658*, 2023.

Boelts, J., Lueckmann, J.-M., Gao, R., and Macke, J. H. Flexible and efficient simulation-based inference for models of decision-making. *Elife*, 11:e77220, 2022.

Chen, L., Rottmayer, J., Kusch, L., Gauger, N. R., and Ye, Y. Data-driven aerodynamic shape design with distributionally robust optimization approaches. *arXiv preprint arXiv:2310.08931*, 2023.

Chen, Z., Kuhn, D., and Wiesemann, W. Data-driven chance constrained programs over wasserstein balls. *Operations Research*, 2022.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Crisostomi, M., Dey, K., Barausse, E., and Trotta, R. Neural posterior estimation with guaranteed exact coverage: The ringdown of gw150914. *Physical Review D*, 108(4): 044029, 2023.

Deistler, M., Goncalves, P. J., and Macke, J. H. Truncated proposals for scalable and hassle-free simulation-based inference. *arXiv preprint arXiv:2210.04815*, 2022.

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

Delaunoy, A., Hermans, J., Rozet, F., Wehenkel, A., and Louppe, G. Towards reliable simulation-based inference with balanced neural ratio estimation. *arXiv preprint arXiv:2208.13624*, 2022.

Delaunoy, A., Miller, B. K., Forré, P., Weniger, C., and Louppe, G. Balancing simulation-based inference for conservative posteriors. *arXiv preprint arXiv:2304.10978*, 2023.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. nflows: normalizing flows in PyTorch, November 2020. URL https://doi.org/10.5281/zenodo.4296287.

Elmachtoub, A. N. and Grigas, P. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

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

Feldman, S., Bates, S., and Romano, Y. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023a.

Feldman, S., Einbinder, B.-S., Bates, S., Angelopoulos, A. N., Gendler, A., and Romano, Y. Conformal prediction is robust to dispersive label noise. In *Conformal and Probabilistic Prediction with Applications*, pp. 624–626. PMLR, 2023b.

Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pp. 2404–2414. PMLR, 2019.

Gupta, J. K. and Brandstetter, J. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. Informs, 2019.

Lee, U., Park, S., and Lee, I. Robust design optimization (rdo) of thermoelectric generator system using non-dominated sorting genetic algorithm ii (nsga-ii). *Energy*, 196:117090, 2020.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-based accuracy testing of posterior estimators for general inference. *arXiv preprint arXiv:2302.03026*, 2023.

Li, J., Du, X., and Martins, J. R. Machine learning in aerodynamic shape optimization. *Progress in Aerospace Sciences*, 134:100849, 2022.

Liu, X., Wei, F., and Zhang, G. Uncertainty optimization design of airfoil based on adaptive point adding strategy. *Aerospace Science and Technology*, 130:107875, 2022.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 343–351. PMLR, 2021.

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

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023a.

Nguyen, T., Jewik, J., Bansal, H., Sharma, P., and Grover, A. Climatelearn: Benchmarking machine learning for weather and climate modeling. *arXiv preprint arXiv:2307.01909*, 2023b.

Papamakarios, G. and Murray, I. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 837–848. PMLR, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Patel, Y., Rayan, S., and Tewari, A. Conformal contextual robust optimization. *arXiv preprint arXiv:2310.10003*, 2023.

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

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Sesia, M. and Candès, E. J. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., et al. Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, 4(8):552–567, 2023.

Wu, X., Zhang, W., and Song, S. Robust aerodynamic shape design based on an adaptive stochastic optimization framework. *Structural and Multidisciplinary Optimization*, 57:639–651, 2018.

Yang, Y. and Kuchibhotla, A. K. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Zhou, K., Wang, L., Pang, L.-G., and Shi, S. Exploring qcd matter in extreme conditions with machine learning. *Progress in Particle and Nuclear Physics*, pp. 104084, 2023.

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 $\mathcal{P}_{X,\mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha$. Then,*

$$\mathcal{P}_{X,\mathcal{P}_C}\left(\Delta(X, \mathcal{P}_C) \leq L \operatorname{diam}(\mathcal{U}(X))\right) \geq 1 - \alpha. \tag{11}$$

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

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

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

B. 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.

B.1. Gaussian Linear

10-dimensional Gaussian model with a Gaussian prior:

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

B.2. Gaussian Linear Uniform

10-dimensional Gaussian model with a uniform prior:

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

B.3. SLCP with Distractors

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

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

B.4. Bernoulli GLM Raw

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

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

B.5. Gaussian Mixture

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

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

B.6. Two Moons

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

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

B.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:

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

B.8. Lotka-Volterra

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

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

C. 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 & 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.