Learning Polyhedral Conformal Sets for Robust Optimization

Shuyi Chen* **Wenbin Zhou*** **Shixiang Zhu**
Carnegie Mellon University

Abstract

Robust optimization (RO) provides a principled framework for decision-making under uncertainty, but its performance critically depends on the choice of the uncertainty set. While large sets ensure reliability, they often lead to overly conservative decisions, whereas small sets risk excluding the true outcome. Recent data-driven approaches, particularly conformal prediction, offer finite-sample validity guarantees but remain largely task-agnostic, ignoring the downstream decision structure. In this paper, we propose a decision-aware conformal framework that learns uncertainty sets tailored to robust optimization objectives. Our approach parameterizes a flexible family of polyhedral sets via data-driven hyperplanes and learns their geometry by directly minimizing the induced robust loss, while preserving statistical validity through conformal calibration. To correct for data-dependent selection, we incorporate a recalibration step on an independent dataset to restore coverage. The resulting sets capture directional and anisotropic uncertainty aligned with the decision objective while remaining computationally tractable. We provide finite-sample coverage guarantees and bounds on the suboptimality gap to an oracle decision. This work bridges the gap between statistical validity and decision optimality, providing a principled framework for data-driven robust optimization.

1 Introduction

Robust optimization (RO) is a fundamental framework for decision-making under uncertainty [Ben-Tal et al., 2009, Ben-Tal and Nemirovski, 1998], with applications across operations research, finance, energy systems, supply chain management, and machine learning. RO prescribes decisions that perform well under worst-case realizations of uncertain parameters, hedging against ambiguity and model misspecification. This is particularly valuable in high-stakes settings: power grid operators must plan infrastructure upgrades under uncertain demand and renewable generation [Jabr, 2013, Abdin et al., 2022, Zhou and Zhu, 2024]; supply chain managers must allocate inventory without knowing future disruptions [Bertsimas and Thiele, 2006]; and healthcare systems must make treatment decisions under uncertain patient outcomes [Mo et al., 2021].

A central component of robust optimization is the specification of the uncertainty set, which characterizes plausible realizations of the unknown parameters. The quality of the resulting decision critically depends on this choice. A well-known challenge is the trade-off between *conservativeness and performance*: overly large uncertainty sets provide strong protection against worst-case scenarios but often lead to overly pessimistic decisions, while overly small sets may exclude the true outcome, resulting in unreliable and potentially catastrophic decisions.

While the literature has extensively studied how to ensure robustness—often by constructing uncertainty sets that are sufficiently large to guarantee feasibility or coverage—it has paid comparatively less attention to the risk of over-conservativeness [Zhou and Zhu, 2025, Wang et al., 2025a]. In

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

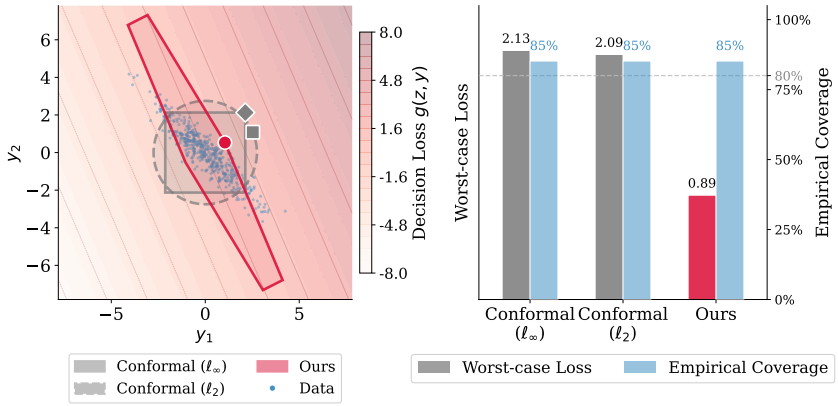


Figure 1: Illustration of decision-aware conformal prediction sets. Standard conformal sets (grey) and our decision-aware set (red), overlaid with the loss $g(z, y) = z^\top y$. All three sets achieve the same empirical coverage (85%), but our set is oriented to avoid high-loss directions, reducing worst-case decision loss.

practice, this imbalance can have significant consequences. For instance, in power systems planning, an overly conservative uncertainty set may trigger unnecessary infrastructure investments, leading to excessive costs for utilities and consumers [Chen et al., 2026b]. In inventory management, it may result in systematic overstocking, tying up capital and increasing waste [Lim and Wang, 2017]. Similarly, in portfolio optimization, overly conservative risk sets can lead to excessively defensive allocations that sacrifice substantial returns [Gregory et al., 2011]. These examples highlight the importance of striking the right balance: uncertainty sets should be large enough to ensure reliability, yet sufficiently tight and structured to avoid overly pessimistic decisions.

Recent data-driven approaches construct uncertainty sets via conformal prediction [Yeh et al., 2024, Patel et al., 2024b], which provides distribution-free, finite-sample guarantees that the true outcome lies in the set with high probability [Vovk et al., 2005]. However, standard conformal prediction is inherently *task-agnostic*: it constructs sets based solely on predictive accuracy, without accounting for the downstream decision problem [Wang et al., 2023], so the resulting sets may be unnecessarily large or poorly shaped for decision-making. Methods that tighten conformal sets or adapt them to the data distribution [Zheng and Zhu, 2024] remain largely disconnected from the structure of the downstream optimization problem. Consequently, there remains a fundamental gap between *statistical validity* and *decision optimality*: existing approaches either guarantee coverage without considering decision performance, or improve efficiency without explicitly optimizing for the decision task.

In this paper, we bridge this gap with a decision-aware conformal prediction framework for robust optimization, in which the geometry of the uncertainty set is explicitly aligned with the downstream decision problem rather than treated as a task-agnostic byproduct of prediction (Fig. 1). We introduce a learnable nonconformity score that induces *polyhedral uncertainty sets* via data-driven hyperplanes, capturing the *directional and anisotropic uncertainty* most relevant to the decision objective while keeping the resulting robust optimization problem tractable. Our method learns the set geometry by minimizing the induced robust loss and maintains statistical validity through conformal calibration, with a recalibration step on an independent dataset that corrects for selection bias. We establish finite-sample coverage for the final, data-dependent set, and derive finite-sample bounds on the suboptimality gap relative to an oracle decision, decomposed into calibration and learning errors, which provide guidance on allocating data between learning and recalibration.

Related work. A growing literature adapts conformal prediction (CP) to construct uncertainty sets for robust optimization (RO), using shapes such as ellipsoids [Johnstone and Cox, 2021] and unions of balls [Patel et al., 2024b], applying CP to RO problems in power systems [Chen et al., 2025, 2026b] and autonomous driving [Lekeufack et al., 2024], and characterizing CP sets under RO from decision-theoretic perspectives [Kiyani et al., 2025, Chenreddy and Delage, 2024]. A parallel line seeks more efficient CP sets via high-density regions [Lei and Wasserman, 2014, Izbicki et al., 2022, Zheng and Zhu, 2024], expressive multidimensional scores [Xu et al., 2024, Tumu et al., 2024,

Fang et al., 2025], model selection [Liang et al., 2024, Yang and Kuchibhotla, 2025], or end-to-end minimization of set size [Bai et al., 2022, Braun et al., 2025, Kiyani et al., 2024]. Closest to our work are decision-focused approaches [Wilder et al., 2019, Mandi et al., 2024] that tailor CP sets to downstream decisions through utility-aware scores [Cortes-Gomez et al., 2024], end-to-end training of neural convex sets [Yeh et al., 2024], or selection among a finite set of conformal predictors [Bao et al., 2025]. In contrast, we directly learn the geometry of a polyhedral conformal set for a general decision objective and restore validity by recalibration; unlike conformal predictive programming [Zhao et al., 2024], which seeks a decision satisfying a conformalized chance constraint, we learn the set geometry itself. Appendix A provides a detailed discussion.

2 Problem Setup and Preliminaries

In many high-stakes decision-making problems, the true outcome $y \in \mathcal{Y} \subseteq \mathbb{R}^d$ is unknown at the time a decision $z \in \mathcal{Z} \subseteq \mathbb{R}^p$ must be made, where $\mathcal{Z}$ denotes the feasible decision region. A standard approach is to first construct an uncertainty (or prediction) set $\mathcal{C}(x) \subseteq \mathcal{Y}$ based on observed features $x \in \mathcal{X}$, and then select a decision that is robust against all realizations within this set [Johnstone and Cox, 2021]. Formally, given a loss function $g(z, u)$ that evaluates the performance of decision z under a potential outcome $u \in \mathcal{Y}$ in context x, the decision is obtained by solving the contextual robust optimization problem

$$\min_{z \in \mathcal{Z}} \max_{u \in \mathcal{C}(x)} g(z, u),$$

where the inner maximization captures adversarial uncertainty over the set $\mathcal{C}(x)$, and the outer minimization selects the decision that minimizes the worst-case loss.

To balance the *trade-off between robustness and informativeness* [Zhou and Zhu, 2025], we seek uncertainty sets that are both *statistically valid* and *decision-aware*. Specifically, we consider a parameterized family of set-valued mappings $\{\mathcal{C}(\cdot;\theta) : \mathcal{X} \to 2^{\mathcal{Y}}\}$ indexed by $\theta \in \Theta$. Let $(X, Y) \sim \mathcal{P}$ denote the joint distribution of features and outcomes. Our goal is to learn a parameter θ such that: (i) the resulting uncertainty set achieves a desired coverage level, *i.e.*, $\mathbb{P}\{Y \in \mathcal{C}(X;\theta)\} \geq 1-\alpha$, and (ii) the induced robust decision incurs low worst-case loss, which leads to the following constrained learning problem:

$$\min_{\theta \in \Theta} \quad \ell(\theta) \coloneqq \min_{z \in \mathcal{Z}} \max_{u \in \mathcal{C}(x;\theta)} g(z, u) \tag{1}$$

$$\text{s.t.} \quad \mathbb{P}\{Y \in \mathcal{C}(X;\theta)\} \geq 1-\alpha. \tag{2}$$

For a fixed context x and parameter θ, the inner min–max problem is a deterministic robust optimization problem that defines the decision induced by $\mathcal{C}(x;\theta)$; the outer optimization then selects, among all sets satisfying (2), the one with the best downstream decision performance. Thus, (1) is a decision-focused formulation that chooses the geometry of the uncertainty set to directly optimize decision quality.

2.1 Preliminaries: Conformal Prediction

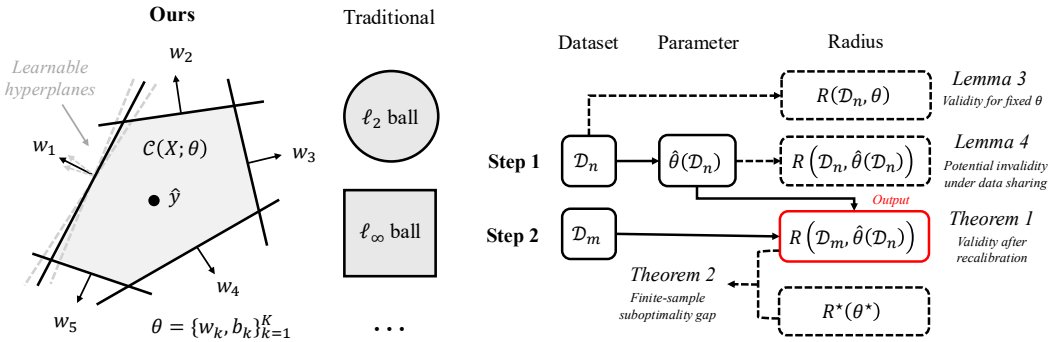
Conformal prediction provides a general, model-agnostic framework for constructing prediction sets with finite-sample, distribution-free coverage guarantees [Vovk et al., 2005, Shafer and Vovk, 2008, Angelopoulos and Bates, 2023]. Suppose the data $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are *exchangeable*. Given calibration data $\mathcal{D}_n \coloneqq \{(X_i, Y_i)\}_{i=1}^{n}$, the goal is to construct a set-valued predictor $\mathcal{C}(x) \subseteq \mathcal{Y}$ such that, for a new test point (X_{n+1}, Y_{n+1}), the coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1};\mathcal{D}_n)\} \geq 1-\alpha \tag{3}$$

holds for a prescribed miscoverage level $\alpha \in (0, 1)$.

Assumption 1 (Exchangeability)**.** *The augmented sample* $\{(X_i, Y_i)\}_{i=1}^{n+1}$ *is exchangeable,* i.e.*, its joint distribution is invariant under any permutation of* $\{1, \ldots, n+1\}$.

A common implementation is split conformal prediction [Vovk et al., 2005, Lei et al., 2018]. The data are partitioned into a training set and a calibration set. A predictive model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ is first trained on the training set, and a nonconformity score function $s(x, y)$ is defined to measure how



(a) The proposed learnable conformal set

(b) Overview of the proposed algorithm

Figure 2: Overview of decision-aware conformal set learning.

well a candidate outcome y conforms to the prediction at $\hat{f}(x)$. Given the calibration data $\mathcal{D}_n$, the empirical quantile of the nonconformity scores is defined as

$$R(\mathcal{D}_n) = \inf\left\{ r : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i) \le r\} \ge \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\}. \tag{4}$$

The prediction set for a new input x is then constructed as

$$\mathcal{C}(x; \mathcal{D}_n) = \{u \in \mathcal{Y} : s(x, u) \le R(\mathcal{D}_n)\}.$$

By construction, the set $\mathcal{C}$ satisfies the marginal coverage guarantee in (3) under the exchangeability assumption, regardless of the choice of the underlying predictive model $\hat{f}$.

3 Decision-Aware Conformal Calibration

We now present our decision-aware conformal framework (Fig. 2). It parameterizes polyhedral sets via data-driven hyperplanes, learns their geometry by minimizing the induced robust decision loss, and ensures statistical validity through conformal calibration, restored after this data-dependent selection by recalibrating on an independent dataset.

3.1 Learnable Polyhedral Conformal Set

Unlike classical uncertainty sets with fixed and isotropic geometry (*e.g.*, ℓ_2 balls) [Ben-Tal and Nemirovski, 1998, 1999, Shafer and Vovk, 2008], our core idea is to construct a parameterized family of polyhedral sets defined as intersections of data-driven halfspaces (supporting hyperplanes), allowing the geometry to adapt to the downstream decision objective (Fig. 2a). We require statistical validity to hold uniformly over all hyperplane configurations, and then optimize the parameters within this valid family to minimize downstream decision loss.

Let $\epsilon(x, y) := y - \hat{f}(x)$ denote the residual, and define its augmented form $\tilde{\epsilon}(x, y) := [\epsilon(x, y)^\top, 1]^\top \in \mathbb{R}^{d+1}$. We parameterize the nonconformity score as

$$s(x, y; \theta) := \max_{k \in [K]} \left\{ \theta_k^\top \tilde{\epsilon}(x, y) \right\}, \tag{5}$$

where $\theta = \{\theta_k := [w_k^\top, b_k]^\top \in \mathbb{R}^{d+1}\}_{k=1}^K$. This score is a max-affine (polyhedral convex) function, which can be interpreted as a polyhedral gauge on the residual space, whose sublevel sets define polyhedral regions shaped by the hyperplanes $\{\theta_k\}_{k=1}^K$; see, *e.g.*, [Niculescu and Persson, 2006]. With symmetric choices of $\{\theta_k\}$, it recovers classical polyhedral norms; for example, $\{\pm e_j\}_{j=1}^d$ yields the ℓ_∞ norm.

Given $s(x, y; \theta)$, we define a parameter-dependent radius via split conformal calibration on the calibration dataset $\mathcal{D}_n$:

$$R(\mathcal{D}_n; \theta) = \inf\left\{ r \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i; \theta) \le r\} \ge \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\}. \tag{6}$$

For any fixed θ and a new input x, the resulting conformal prediction set

$$\mathcal{C}(x; \mathcal{D}_n, \theta) \coloneqq \{u \in \mathcal{Y} : s(x, u; \theta) \leq R(\mathcal{D}_n; \theta)\} \tag{7}$$

satisfies the coverage guarantee in (3). To ensure that $\mathcal{C}(\cdot; \mathcal{D}_n, \theta)$ is a bounded polyhedron (polytope) for all x, we require $K \geq d + 1$ and that the collection of normal vectors $\{w_k\}_{k=1}^K \subset \mathbb{R}^d$—corresponding to the first d components of θ_k—positively spans $\mathbb{R}^d$.

3.2 Proposed Algorithm

We now present a two-stage procedure that integrates decision-aware learning of the proposed conformal set with a post-hoc recalibration step that ensures valid coverage (Fig. 2b).

Step 1: Decision-Aware Conformal Set Learning. Given the parameterized conformal set $\mathcal{C}(\cdot; \mathcal{D}_n, \theta)$ in (7), we learn the parameter θ by directly minimizing the downstream decision loss induced by the corresponding robust optimization problem:

$$\mathcal{L}_x(\theta, r) \coloneqq \min_{z \in \mathcal{Z}} \quad \max_{u \in \mathcal{Y}} \; g(z, u) \tag{8}$$

$$\text{s.t.} \quad s(x, u; \theta) \leq r. \tag{9}$$

We aim to minimize its empirical average over the learning dataset $\mathcal{D}_n$,

$$\hat{\theta}(\mathcal{D}_n) \coloneqq \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{X_i}(\theta, R(\mathcal{D}_n; \theta)) \,.$$

Here, we explicitly denote the dependence of $\hat{\theta}$ on the learning data $\mathcal{D}_n$. In the following sections, we suppress the subscript x from $\mathcal{L}_x$ when clear from context.

Compared to the original formulation (1)–(2), the coverage constraint is enforced implicitly through conformal calibration via $R(\mathcal{D}_n; \theta)$. This allows the probabilistic constraint to be replaced by the deterministic constraint (9), thereby decoupling statistical validity from the optimization procedure. Moreover, the polyhedral structure induced by $s(\cdot, \cdot; \theta)$ implies that the feasible set in (9) can be represented as a finite collection of linear inequalities.

Remark 1 (Connection with sample average approximation). The radius condition $r \geq R(\mathcal{D}_n; \theta)$ is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i; \theta) \leq r\} \geq 1 - \omega_n, \quad \text{where } \omega_n = 1 - \frac{\lceil (n+1)(1-\alpha) \rceil}{n}.$$

Therefore, the radius condition can be viewed as an n-dependent sample average approximation (SAA) constraint with the conformal choice of ω_n [Luedtke and Ahmed, 2008, Zhao et al., 2024].

Step 2: Recalibration for Robust Decision-Making. The conformal calibration in (6) guarantees valid coverage for any *fixed* parameter θ. However, in Step 1, $\hat{\theta}$ is learned from the data $\mathcal{D}_n$, which introduces a data-dependent selection effect [Hegazy et al., 2025, Liang et al., 2024] and violates the exchangeability condition required for conformal validity.

To restore validity, we perform a recalibration step using an independent dataset. Specifically, let $\mathcal{D}_m \coloneqq \{(X_j, Y_j)\}_{j=1}^m$ denote an additional set of exchangeable samples drawn from the same distribution. Given the learned parameter $\hat{\theta}(\mathcal{D}_n)$, we compute the conformal radius on this fresh calibration set, yielding $R(\mathcal{D}_m; \hat{\theta})$. The resulting conformal set $\mathcal{C}(x; \mathcal{D}_m, \hat{\theta})$ is then used to prescribe the final decision

$$\hat{z}(x) = \arg\min_{z \in \mathcal{Z}} \; \max_{u} \; g(z, u), \quad \text{s.t. } u \in \mathcal{C}(x; \mathcal{D}_m, \hat{\theta}(\mathcal{D}_n)). \tag{10}$$

This recalibration step is essential to ensure valid post-selection coverage. Conformal guarantees rely on exchangeability conditional on a fixed nonconformity score; however, optimizing θ using data breaks this condition. Intuitively, reusing the same data for both tuning θ and calibrating the radius leads to overfitting of the score function and overly optimistic (*i.e.*, too small) uncertainty sets, resulting in undercoverage. By recalibrating on an independent dataset, we restore the required independence and recover valid coverage guarantees for the final decision.

3.3 Computational Methods

To operationalize the proposed algorithm, we derive finite-dimensional reformulations of the decision-aware conformal set learning problem in (8)–(9) for linear and more general objective functions g. Two main computational challenges arise: (i) the nonconformity score in (5) introduces a nonsmooth max-affine constraint, and (ii) the conformal radius $R(\mathcal{D}_n;\theta)$ is an empirical quantile that depends on θ through a nonsmooth order statistic. All proofs are included in Appendix B.

We begin by reformulating the max-affine score constraint in (9) as a linear system for fixed conformal radius r. Recall that $\theta_k = [w_k^\top, b_k]^\top$, where $w_k \in \mathbb{R}^d$ and $b_k \in \mathbb{R}$. Define

$$W := [w_1, \ldots, w_K]^\top \in \mathbb{R}^{K\times d}, \qquad b := [b_1, \ldots, b_K]^\top \in \mathbb{R}^K, \qquad c_n := \lceil (n+1)(1-\alpha) \rceil .$$

Lemma 1 (Linear reformulation of the score constraint)**.** *For any radius r, the constraint $s(x,y;W,b) \le r$ is equivalent to the following system of linear inequalities*

$$W\epsilon(x,y) + b \le r1_K,$$

where 1_K denotes the all-ones vector in $\mathbb{R}^K$.

To capture the dependence of the conformal radius $R(\mathcal{D}_n;W,b)$ on (W,b) using the big-M method [Nemhauser and Wolsey, 1988, Williams, 2013], we introduce an auxiliary scalar $r \in \mathbb{R}$ and binary variables $t = [t_1, \ldots, t_n]^\top \in \{0,1\}^n$.

Lemma 2 (Big-M representation of the conformal radius)**.** *Suppose $M > 0$ is such that, for all feasible (W,b) and $i \in [n]$, $R(\mathcal{D}_n;W,b) \ge \underline{r}$ and $W\epsilon_i + b \le (\underline{r} + M)1_K$. Then the empirical conformal radius can be represented as*

$$\begin{aligned} R(\mathcal{D}_n;W,b) = \min_{r,t} \quad & r \\ \text{s.t.} \quad & W\epsilon_i + b \le (r + M(1-t_i))1_K, \quad i \in [n], \\ & 1_n^\top t \ge c_n, \quad t \in \{0,1\}^n. \end{aligned} \tag{11}$$

Here, $t_i = 1$ enforces that the ith nonconformity score is no larger than r, whereas $t_i = 0$ relaxes the corresponding constraint. The constraint $1_n^\top t \ge c_n$ therefore requires at least c_n nonconformity scores to be bounded by r, and minimizing r recovers the empirical conformal quantile.

Combining Lemmas 1 and 2 gives the following exact reformulations to problem (8)–(9).

Proposition 1 (Reformulation of problem (8)–(9))**.** *The problem (8)–(9) is equivalent to*

$$\begin{aligned} \min_{W,b,z,r,t} \quad & \max_{u\in\mathcal{Y}} \; g(z,u) \\ \text{s.t.} \quad & W\left(u - \hat{f}(x)\right) + b \le r1_K, \\ & W\epsilon_i + b \le (r + M(1-t_i))1_K, \qquad i \in [n], \\ & 1_n^\top t \ge c_n, \quad z \in \mathcal{Z}, \quad t \in \{0,1\}^n. \end{aligned} \tag{12}$$

Corollary 1 (Reformulation for linear objective)**.** *Suppose $\mathcal{Y} = \mathbb{R}^d$ and $g(z,y) = g_0(z) + z^\top Qy$. The min–max problem in (12) admits the following equivalent reformulation.*

$$\begin{aligned} \min_{W,b,z,\lambda,r,t} \quad & g_0(z) + \left(r1_K - b + W\hat{f}(x)\right)^\top \lambda \\ \text{s.t.} \quad & W^\top\lambda = Q^\top z, \qquad \lambda \in \mathbb{R}^K_+, \\ & W\epsilon_i + b \le (r + M(1-t_i))1_K, \qquad i \in [n], \\ & 1_n^\top t \ge c_n, \quad z \in \mathcal{Z}, \quad t \in \{0,1\}^n. \end{aligned} \tag{13}$$

The mixed-integer formulation in (13) can be solved exactly using spatial branch-and-bound solvers such as Gurobi [Gurobi Optimization, LLC, 2026] for small K and n. For general concave objectives $g(z,\cdot)$ in (12), we employ Column-and-Constraint Generation (CCG) [Zhao and Zeng, 2012], which alternates between a master problem over (W,b,z,r,t) and an adversarial subproblem over u (Alg. 1 in Appendix B.5). However, the master problem remains a mixed-integer program with exponential complexity in n [Nemhauser and Wolsey, 1988], and when $g(z,\cdot)$ is convex, the adversarial

subproblem requires searching over the vertices of the polytope, which can number $\mathcal{O}(K^{\lfloor d/2 \rfloor})$ in $\mathbb{R}^d$.

For a more efficient solution, we relax the empirical quantile constraint in (11) using a continuous pinball-loss penalty [Steinwart and Christmann, 2011]. This replaces the n binary variables with a piecewise-linear penalty and enables stochastic subgradient methods on minibatches of calibration samples, substantially reducing the per-iteration complexity (Alg. 2 and Appendix B.5).

4 Theoretical Analysis

This section presents two central theoretical properties: (i) the recalibrated parameter yields a valid conformal prediction set, and (ii) the suboptimality gap admits a finite-sample characterization under mild regularity conditions. The proofs are deferred to Appendix C.

4.1 Statistical Validity of the Recalibrated Conformal Set

We begin by recalling the standard validity guarantee of conformal prediction for a fixed nonconformity score, which is well established in the literature [Angelopoulos and Bates, 2023].

Lemma 3 (Validity for Fixed θ)**.** *Under the exchangeability assumption, for any fixed $\theta \in \Theta$, the conformal prediction set $\mathcal{C}(\cdot; \mathcal{D}_n, \theta)$ defined in (6)–(7) satisfies*

$$\mathbb{P}\big\{Y_{n+1} \in \mathcal{C}(X_{n+1}; \mathcal{D}_n, \theta)\big\} \geq 1 - \alpha.$$

When θ is instead fitted to the calibration data, yielding $\hat{\theta}(\mathcal{D}_n)$, the augmented scores are no longer exchangeable, and the prediction set tends to overfit the calibration samples, leading to undercoverage. The following result formalizes this breakdown; its proof follows arguments similar to Theorem 2 by Barber et al. [2023] and Corollary 1 by Zhou et al. [2025]. Let $d_{\mathrm{TV}}(\mathcal{D}_{n+1}, \mathcal{D}_{n+1}^{(i)})$ denote the total variation distance between the nonconformity scores of the augmented sample $\mathcal{D}_{n+1} = \{(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})\}$ computed with $\hat{\theta}$ fitted on the first n entries of $\mathcal{D}_{n+1}$ and on those of the sequence $\mathcal{D}_{n+1}^{(i)}$ obtained by exchanging its $(n+1)$-th and i-th entries. This quantity measures how much exchangeability is violated: it is zero when $\hat{\theta}$ does not depend on $\mathcal{D}_n$ and grows with the dependence between $\hat{\theta}(\mathcal{D}_n)$ and the nonconformity scores.

Lemma 4 (Potential Invalidity Under Data Sharing)**.** *Under the exchangeability assumption, if the learning data are reused in constructing the conformal prediction set $\mathcal{C}(\cdot; \mathcal{D}_n, \hat{\theta}(\mathcal{D}_n))$, then:*

$$\mathbb{P}\big\{Y_{n+1} \in \mathcal{C}(X_{n+1}; \mathcal{D}_n, \hat{\theta}(\mathcal{D}_n))\big\} \geq 1 - \alpha - \tfrac{1}{n+1} \sum\nolimits_{i=1}^{n} d_{\mathrm{TV}}\left(\mathcal{D}_{n+1}, \mathcal{D}_{n+1}^{(i)}\right).$$

Our algorithm prevents such invalidity via a recalibration step using an independent split dataset, thereby restoring finite-sample validity even when θ is data-dependent.

Theorem 1 (Validity of the Recalibrated Set)**.** *Suppose the recalibration data $\{(X_j, Y_j)\}_{j=1}^{m}$, the test point (X_{m+1}, Y_{m+1}), and the data used to obtain $\hat{\theta}$ are mutually exchangeable. Then the recalibrated prediction set satisfies*

$$\mathbb{P}\big\{Y_{m+1} \in \mathcal{C}(X_{m+1}; \mathcal{D}_m, \hat{\theta}(\mathcal{D}_n))\big\} \geq 1 - \alpha.$$

4.2 Suboptimality Gap of the Prescribed Decision

We now quantify the gap between the decision induced by the learned parameter $\hat{\theta}$ and that of an oracle choice $\theta^\star$, establishing finite-sample bounds on how calibration, learning, and optimization errors propagate to the final decision.

To formalize this analysis, we first introduce oracle benchmarks and a radius-explicit formulation of the robust loss. For a fixed x and any θ, define the oracle conformal radius

$$R^\star(\theta) = \inf_r \{r \in \mathbb{R} : \mathbb{P}\{s(X, Y; \theta) \leq r\} \geq 1 - \alpha\},$$

and the corresponding oracle parameter

$$\theta^\star = \arg\min_{\theta} \min_{z \in \mathcal{Z}} \max_{u} \; g(z, u), \quad \text{s.t. } s(x, u; \theta) \leq R^\star(\theta). \tag{14}$$

Recall that the robust decision loss defined in (8) is denoted by $\mathcal{L}(\theta, r)$, where we suppress the subscript x.

We impose the following regularity conditions, which are standard in empirical process analysis and control the estimation of the conformal radius and the stability of the induced robust objective.

Assumption 2 (Regularity Conditions)**.** *Assume all data are independent, and:*

(a) *The function class* $\{(x, y) \mapsto \mathbb{1}\{s(x, y; \theta) \le r\}, \theta \in \Theta, r \in \mathbb{R}\}$ *has VC dimension* V.
(b) *There exists* $\kappa > 0$ *such that for all* $\theta \in \Theta$, $f_\theta(t) \ge \kappa$ *for* t *in a neighborhood of* $R^\star(\theta)$, *where* f_θ *is the density of the nonconformity scores* $s(X, Y; \theta)$.
(c) *There exists* $L_r \ge 0$ *such that for all* $\theta \in \Theta$ *and* $x \in \mathcal{X}$, *there is:*

$$|\mathcal{L}(\theta, r) - \mathcal{L}(\theta, r')| \le L_r \cdot |r - r'|$$

(d) *Problem* (8)–(9) *is solved to optimization tolerance* ϵ_{opt}.

Condition (a) ensures uniform convergence over the score class, (b) guarantees well-behaved quantile estimation near the target radius, (c) ensures stability of the downstream objective with respect to the radius, and (d) accounts for optimization error.

Theorem 2 (Finite-Sample Suboptimality Gap)**.** *Let* $\hat{\theta}$ *be the estimated parameter obtained by solving* (8)–(9), *and let* $\theta^\star$ *be the oracle parameter obtained by solving* (14). *Let* $\mathcal{L}(\cdot)$ *be the robust loss defined in* (8), *and denote the suboptimality gap as:* $\Delta := \mathcal{L}(\hat{\theta}(\mathcal{D}_n), R(\mathcal{D}_m; \hat{\theta}(\mathcal{D}_n))) - \mathcal{L}(\theta^\star, R^\star(\theta^\star))$. *Then, under Assumption 2 and for sufficiently large* m, n, *with probability at least* $1 - \delta$:

$$\Delta \le \frac{L_r}{\kappa}\sqrt{\frac{\log(4/\delta)}{2m}} + \frac{2L_r}{\kappa}\sqrt{\frac{32\left[V\log(en/V) + \log(16/\delta)\right]}{n}} + \frac{2L_r}{\kappa}\left(\frac{1}{m} + \frac{2}{n}\right) + \epsilon_{\text{opt}}.$$

In $\mathcal{O}_{\mathbb{P}}$ notation [Van der Vaart, 2000], the bound simplifies to

$$\Delta \le \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{\log n}{n}}\right) + \epsilon_{\text{opt}}.$$

In addition to the optimization tolerance, the bound contains a calibration error scaling as $m^{-1/2}$ and a learning error scaling as $\sqrt{(\log n)/n}$. Since the calibration error decays faster, it is typically beneficial to allocate more data to learning (Step 1) than to recalibration (Step 2). Because the bound involves unknown constants (V, L_r, ϵ_{opt}, and κ), we use it to guide the data-splitting ratio rather than as a quantitative measure of decision suboptimality. The proof is provided in Appendix C.4.

Remark 2. Theorems 1 and 2 continue to hold when the pinball surrogate is used: validity relies only on the exchangeability of the recalibration data, and the error introduced by the surrogate is absorbed into ϵ_{opt}.

5 Experiments

We evaluate the proposed decision-aware conformal framework on robust decision problems using both synthetic and real-data tasks. Across all experiments, we first fit a predictor $\hat{f} : \mathcal{X} \to \mathbb{R}^d$ on an independent training split and construct residuals $\epsilon_i = Y_i - \hat{f}(X_i)$. The remaining data are split into a learning set $\mathcal{D}_n$, used to learn the set geometry $\hat{\theta}(\mathcal{D}_n)$, and an independent calibration set $\mathcal{D}_m$, used only for the final conformal radius. This yields the final set $\mathcal{C}(x; \mathcal{D}_m, \hat{\theta}(\mathcal{D}_n))$, from which robust decisions are computed on held-out test covariates. Baselines that do not learn set geometry use $\mathcal{D}_n \cup \mathcal{D}_m$ as their calibration set.

We compare against decision-agnostic conformal sets, including Bonferroni boxes [Dunn, 1961], conformal boxes or balls [Vovk et al., 2005], and PCP [Wang et al., 2023], as well as decision-aware alternatives based on F-CROMS [Bao et al., 2025] and PICNN scores [Yeh et al., 2024]. We also include three ablations of our method: a polyhedral set without decision-aware learning, a size-aware polyhedron that uses volume as the loss function, and our decision-aware polyhedron without the post-selection recalibration step. For our method, each geometry-learning problem is first solved exactly as the MIQCP of Corollary 1 with Gurobi under a 30-minute limit; if Gurobi does not certify

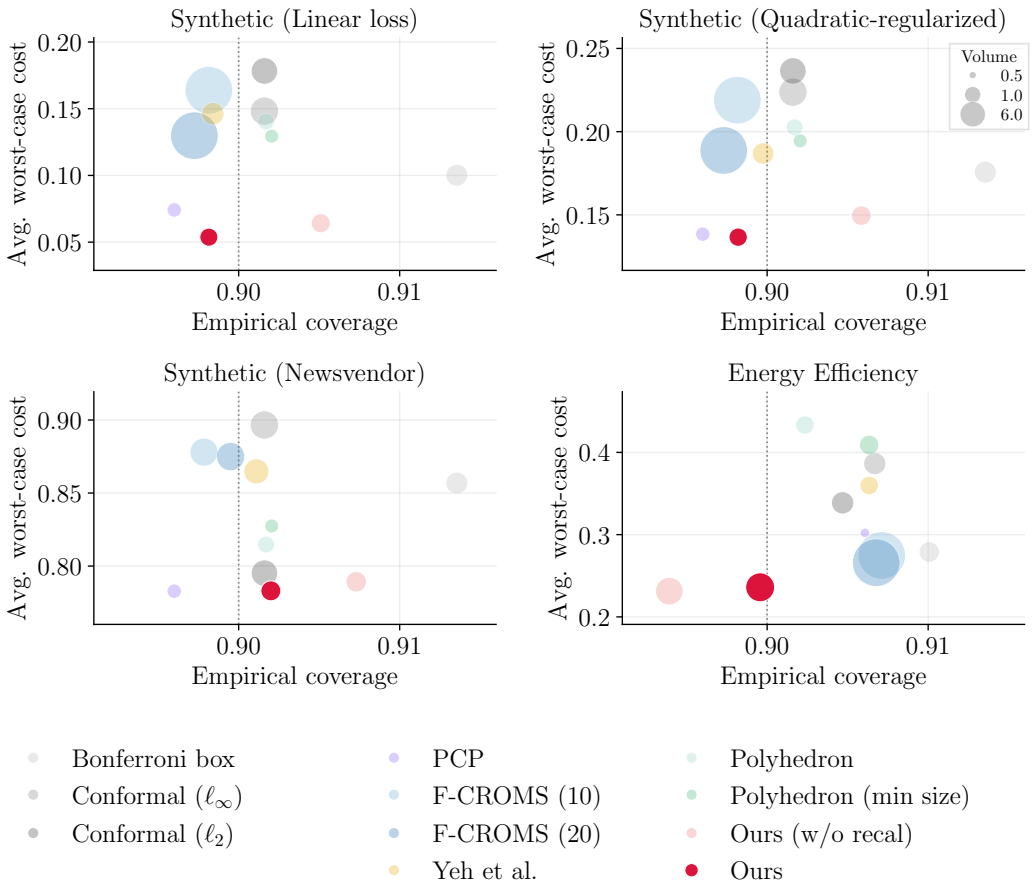


Figure 3: The proposed method in comparison with baselines. The horizontal axis shows the empirical coverage rate; the vertical axis presents the average worst-case cost; and bubble size represents the volume of the uncertainty set.

optimality within this limit, or the reformulation does not apply, we use the pinball-loss gradient method of Section 3.3 (Appendix D.3). We evaluate empirical coverage, average worst-case cost, average regret, set volume, and wall-clock time. Additional details on the objective function, datasets, and implementation are deferred to Appendix D. Results are averaged over 20 random runs, except that we use 80 runs for the real-data Energy Efficiency task and five runs for the sensitivity analysis.

Main results. Fig. 3 and Tables 2 and 3 summarize the results. Within the 30-minute limit, Gurobi certified the MIQCP in 50 of the 80 energy-efficiency runs and 3 of the 20 linear-objective runs; all other runs use the pinball-loss gradient method. Across all tasks, our recalibrated decision-aware polyhedron attains empirical coverage close to the nominal 90% level, confirming that the independent calibration split restores validity after learning the set geometry. Among valid methods, our method consistently achieves competitive or lowest worst-case cost. The comparison with PCP [Wang et al., 2023] and the size-aware polyhedron shows that sharper or smaller uncertainty sets are not necessarily better for downstream robust optimization: although both often produce smaller-volume sets, this does not consistently translate into lower worst-case cost or regret, as minimizing volume can remove directions that are important for lowering the worst-case decision loss. Compared with F-CROMS [Bao et al., 2025], adding more candidate sets of hyperplanes can reduce regret, but its model-selection approach remains less flexible than ours because it selects among a finite candidate class rather than directly learning the conformal-set geometry. The ablation without recalibration highlights the importance of post-selection calibration. Although its worst-case cost is often close to that of the recalibrated method, its coverage can fall below the target level (Fig. 4, $d = 4$) because the same data are used both to select the set geometry and to calibrate the radius, consistent with

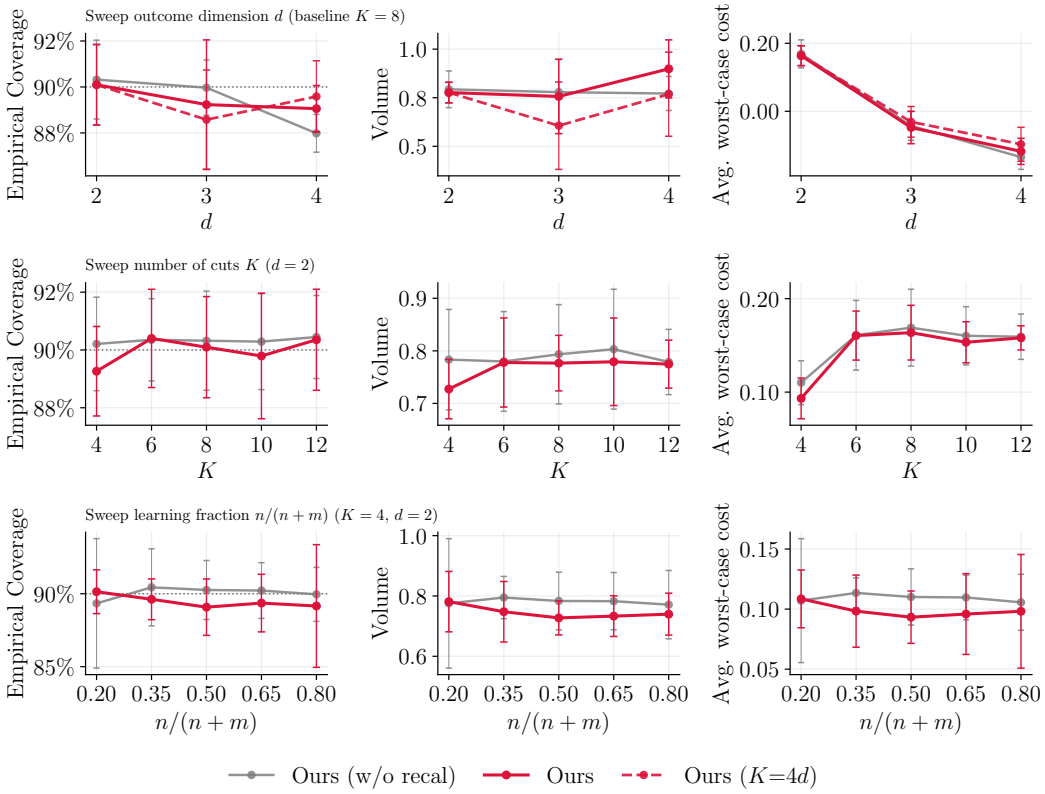


Figure 4: Hyperparameter sensitivity of coverage, volume, and worst-case cost. Ours uses $K = 8$ in the first row and $K = 4$ in the last row.

Lemma 4. By contrast, the proposed two-stage procedure restores the coverage guarantee with little loss in downstream performance.

Sensitivity analysis. Fig. 4 studies the sensitivity of our method to d, K, and $n/(n+m)$. The MIQCP, which applies only at $K = 4$ and $d = 2$, was certified in two of five runs at $n/(n+m) = 0.2$; all other points use the pinball-loss gradient method. As d increases, coverage remains close to the nominal 90% level, while volume grows to satisfy the joint coverage. Scaling the number of hyperplanes with $K = 4d$ helps control the volume growth at a similar worst-case cost, suggesting that the polyhedral template should be more expressive in higher dimensions. Varying K has only a mild effect on coverage. Larger K initially increases volume and worst-case cost, which then remain largely stable. Finally, varying $n/(n+m)$ shows the expected learning–calibration trade-off: allocating more data to $\mathcal{D}_n$ improves geometry learning and can reduce cost, consistent with Theorem 2. But leaving too few samples in $\mathcal{D}_m$ also makes the conformal radius less stable and can lead to coverage below the nominal level.

6 Conclusion and Discussion

We introduced a decision-aware conformal framework for robust optimization that learns uncertainty sets by optimizing downstream robust loss while preserving finite-sample coverage through independent recalibration. The polyhedral parameterization captures directional uncertainty while admitting finite-dimensional reformulations. We established validity for the final learned set and derived a suboptimality-gap bound that separates calibration and learning errors, showing that conformal sets can be designed for both coverage and decision quality. A limitation is the focus on polyhedral geometry, which may be less suitable for highly nonlinear uncertainty. Future work could extend this framework to richer neural or flow-based set families while maintaining calibration validity and computational tractability.

Acknowledgments and Disclosure of Funding

The project was supported by the National Science Foundation (Grant No. CAIG-2425888). We acknowledge partial support from the 2024 Block Center Seed Fund at Carnegie Mellon University and the 2024 GenAI Fellows program offered by the Tepper School of Business.

References

Adam F. Abdin, Aakil Caunhye, Enrico Zio, and Michel-Alexandre Cardin. Optimizing generation expansion planning with operational uncertainty: A multistage adaptive robust approach. *Applied Energy*, 306:118032, 2022. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2021.118032. URL https://www.sciencedirect.com/science/article/pii/S0306261921013271.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pages 136–145. PMLR, 2017.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Yu Bai, Song Mei, Huan Wang, Yingbo Zhou, and Caiming Xiong. Efficient and differentiable conformal prediction with general function classes. In *International Conference on Learning Representations*, 2022.

Yajie Bao, Yang Hu, Haojie Ren, Peng Zhao, and Changliang Zou. Optimal model selection for conformalized robust optimization. *arXiv preprint arXiv:2507.04716*, 2025.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations research letters*, 25(1):1–13, 1999.

Aharon Ben-Tal, Arkadi Nemirovski, and Laurent El Ghaoui. *Robust optimization*. Princeton university press, 2009.

Dimitris Bertsimas and Aurélie Thiele. A robust optimization approach to inventory theory. *Operations Research*, 54(1):150–168, 2006. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/25146954.

Sacha Braun, Liviu Aolaritei, Michael I Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*, 2025.

Giuseppe Carlo Calafiore and Marco C Campi. The scenario approach to robust control design. *IEEE Transactions on automatic control*, 51(5):742–753, 2006.

Abraham Charnes and William W Cooper. Chance-constrained programming. *Management science*, 6(1):73–79, 1959.

Shuyi Chen, Shixiang Zhu, and Ramteen Sioshansi. Enhancing electricity-system resilience with adaptive robust optimization and conformal uncertainty characterization. *arXiv preprint arXiv:2505.11627*, 2025.

Shuyi Chen, Ferdinando Fioretto, Feng Qiu, and Shixiang Zhu. Global-decision-focused neural odes for proactive grid resilience management. *IEEE Transactions on Smart Grid*, 17(3):2506–2516, 2026a. doi: 10.1109/TSG.2025.3642407.

Shuyi Chen, Shixiang Zhu, and Ramteen Sioshansi. Large-scale resilience planning for wildfire-prone electricity-system via adaptive robust optimization. *arXiv preprint arXiv:2604.01232*, 2026b.

Abhilash Chenreddy and Erick Delage. End-to-end conditional robust optimization. *arXiv preprint arXiv:2403.04670*, 2024.

Santiago Cortes-Gomez, Carlos Patino, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. *arXiv preprint arXiv:2410.01767*, 2024.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

Zhenhan Fang, Aixin Tan, and Jian Huang. Contra: Conformal prediction region via normalizing flow transformation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xinbo Geng and Le Xie. Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization. *Annual reviews in control*, 47: 341–363, 2019.

Christine Gregory, Ken Darby-Dowman, and Gautam Mitra. Robust optimization and portfolio selection: The cost of robustness. *European Journal of Operational Research*, 212(2):417–428, 2011.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2026. URL https://www.gurobi.com.

Mahmoud Hegazy, Liviu Aolaritei, Michael I. Jordan, and Aymeric Dieuleveut. Valid selection among conformal sets, 2025. URL https://arxiv.org/abs/2506.20173.

Jessica Hullman, Yifan Wu, Dawei Xie, Ziyang Guo, and Andrew Gelman. Conformal prediction and human decision making. *arXiv preprint arXiv:2503.11709*, 2025.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

R. A. Jabr. Robust transmission network expansion planning with uncertain renewable generation and loads. *IEEE Transactions on Power Systems*, 28(4):4558–4567, 2013. doi: 10.1109/TPWRS.2013.2267058.

Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1):291–327, 2016.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 72–90. PMLR, 2021.

Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024.

Shayan Kiyani, George J Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *International Conference on Machine Learning*, pages 30943–30965. PMLR, 2025.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Jordan Lekeufack, Anastasios N Angelopoulos, Andrea Bajcsy, Michael I Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11668–11675. IEEE, 2024.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection. *arXiv preprint arXiv:2408.07066*, 2024.

Yun Fong Lim and Chen Wang. Inventory management based on target-oriented robust optimization. *Management Science*, 63(12):4409–4427, 2017.

Bo Lin, Erick Delage, and Timothy C Chan. Conformal inverse optimization. *Advances in Neural Information Processing Systems*, 37:63534–63564, 2024.

James Luedtke and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.

James Luedtke, Shabbir Ahmed, and George L Nemhauser. An integer programming approach for linear programs with probabilistic constraints. *Mathematical programming*, 122(2):247–272, 2010.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024.

Bruce L Miller and Harvey M Wagner. Chance constrained programming with joint constraints. *Operations research*, 13(6):930–945, 1965.

Weibin Mo, Zhengling Qi, and Yufeng Liu. Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534):659–674, 2021.

George L Nemhauser and Laurence A Wolsey. *Integer and Combinatorial Optimization*. John Wiley & Sons, New York, 1988. ISBN 978-0-471-82819-8. doi: 10.1002/9781118627372.

Constantin Niculescu and Lars-Erik Persson. *Convex functions and their applications*, volume 23. Springer, 2006.

Yash Patel, Sahana Rayan, and Ambuj Tewari. Conformal robust control of linear systems. *arXiv preprint arXiv:2405.16250*, 2024a.

Yash P Patel, Sahana Rayan, and Ambuj Tewari. Conformal contextual robust optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2485–2493. PMLR, 2024b.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1), February 2011. ISSN 1350-7265. doi: 10.3150/10-bej267. URL http://dx.doi.org/10.3150/10-BEJ267.

David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022.

Athanasios Tsanas and Angeliki Xifara. Energy Efficiency. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C51307.

Renukanandan Tumu, Matthew Cleaveland, Rahul Mangharam, George Pappas, and Lars Lindemann. Multi-modal conformal prediction regions by optimizing convex shape templates. In *6th Annual Learning for Dynamics & Control Conference*, pages 1343–1356. PMLR, 2024.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. ISBN 978-0-387-00152-4. doi: 10.1007/b106715.

Irina Wang, Cole Becker, Bart Van Parys, and Bartolomeo Stellato. Mean robust optimization: I. wang et al. *Mathematical Programming*, 213(1):1235–1277, 2025a.

Prince Zizhuang Wang, Shuyi Chen, Jinhao Liang, Ferdinando Fioretto, and Shixiang Zhu. Gen-dfl: Decision-focused generative learning for robust decision making. *arXiv preprint arXiv:2502.05468*, 2025b.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8814–8836. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wang23n.html.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1658–1665, 2019.

H Paul Williams. *Model building in mathematical programming*. John Wiley & Sons, 2013.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55076–55099, 2024.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120(549):435–447, 2025.

Christopher Yeh, Nicolas Christianson, Alan Wu, Adam Wierman, and Yisong Yue. End-to-end conformal calibration for optimization under uncertainty. *arXiv preprint arXiv:2409.20534*, 2024.

Long Zhao and Bo Zeng. An exact algorithm for two-stage robust optimization with mixed integer recourse problems, 2012.

Yiqi Zhao, Xinyi Yu, Matteo Sesia, Jyotirmoy V Deshmukh, and Lars Lindemann. Conformal predictive programming for chance constrained optimization. *arXiv preprint arXiv:2402.07407*, 2024.

Minxing Zheng and Shixiang Zhu. Generative conformal prediction with vectorized non-conformity scores. *arXiv preprint arXiv:2410.13735*, 2024.

Wenbin Zhou and Shixiang Zhu. Hierarchical probabilistic conformal prediction for distributed energy resources adoption. *arXiv preprint arXiv:2411.12193*, 2024.

Wenbin Zhou and Shixiang Zhu. Calibrating decision robustness via inverse conformal risk control. *arXiv preprint arXiv:2510.07750*, 2025.

Wenbin Zhou, Agni Orfanoudaki, and Shixiang Zhu. Conformalized decision risk assessment. *arXiv preprint arXiv:2505.13243*, 2025.

A Extended Related Work

Adapting conformal prediction (CP) methods to construct valid uncertainty sets in robust optimization (RO) has been a key focus of recent literature development in CP. Some works propose using different shapes, such as using ellipsoids [Johnstone and Cox, 2021] and multiple generated balls [Patel et al., 2024b], to accurately carve the shape of the data and prevent overconservatism. Other works find applications of CP in various real-world RO problems, such as power system resilience [Chen et al., 2025, 2026b] and autonomous driving [Lekeufack et al., 2024]. Another line of literature studies the properties of CP methods in different types of RO problems, such as linear-quadratic regulator systems [Patel et al., 2024a], human decision making [Hullman et al., 2025], inverse optimization

problems [Lin et al., 2024], and decision-theoretic perspectives [Kiyani et al., 2025, Chenreddy and Delage, 2024] to derive optimality conditions for CP methods under RO settings and gain implementation insights. Our work is closely related to this line of research, as we also aim to tailor conformal prediction to better serve robust optimization. However, we differ in our approach: rather than deriving theoretical characterizations of the uncertainty set, we adopt a learning-based framework to estimate its optimal shape.

Tailoring conformal prediction (CP) sets to be more efficient (*i.e.*, tighter) is another important direction of CP research. Previous work has established the optimality of constructing sets in high-density regions (HDR) of the underlying target distribution [Lei and Wasserman, 2014, Izbicki et al., 2022], which motivates a series of studies studying the use of generative models as prediction models to generate multiple fine-grained sets whose union approximates the contour of the HDR [Wang et al., 2023, Zheng and Zhu, 2024]. On the other hand, multidimensional CP focuses on developing more expressive nonconformity scores that allow the prediction set to take flexible shapes, including ellipsoidal sets [Xu et al., 2024], general convex sets [Tumu et al., 2024], and nonconvex sets [Fang et al., 2025]. More recently, several directions of work have explored how conformal prediction set shapes can be explicitly optimized. This can be formulated as a model selection problem [Liang et al., 2024, Yang and Kuchibhotla, 2025], where the task is to select the best performing model from a finite collection of conformal prediction sets. Another line of work formulates this as an end-to-end learning task whose objective is to minimize set size subject to the validity chance-constraint, and can be solved using, *e.g.*, empirical risk minimization [Bai et al., 2022, Braun et al., 2025] and duality [Kiyani et al., 2024]. Our work is closely related to this line of research, as we also formulate a learning problem. However, rather than optimizing an efficiency-oriented objective as in prior works, we consider a general decision-making objective that subsumes efficiency as a special case.

Our work is closely related to decision-focused learning (DFL) [Wilder et al., 2019, Mandi et al., 2024], a class of approaches that train predictive models by optimizing decision-oriented objectives through differentiable optimization layers [Amos and Kolter, 2017, Agrawal et al., 2019], and has been adapted to settings such as generative modeling [Wang et al., 2025b] and sequential modeling [Chen et al., 2026a]. The idea of DFL can be extended to learn conformal prediction models tailored for robust optimization, in a way similar to efficiency-aware conformal prediction [Stutz et al., 2022]. However, this field remains underexplored. For example, Cortes-Gomez et al. [2024] propose decision-aware nonconformity scores that incorporate downstream utility into conformal prediction, yielding sets aligned with decision objectives while preserving coverage guarantees. Yeh et al. [2024] consider a more general framework by parameterizing convex uncertainty sets via neural scores and training them end-to-end using differentiable optimization layers. Bao et al. [2025] instead study a model selection approach, selecting from a discrete set of conformal predictors to minimize downstream decision risk. Our work shares the same motivation as tailoring conformal prediction to downstream decision-making while maintaining validity. However, rather than designing specific scores, relying on end-to-end training, or selecting from a finite model class, we adopt a learning-based approach that directly optimizes the shape of the polyhedral conformal prediction set for a general decision objective to achieve optimality.

Our work is also related to chance-constrained optimization (CCO), which addresses uncertainty by requiring constraints to hold with a prespecified probability, thereby explicitly controlling the risk of constraint violation [Charnes and Cooper, 1959, Miller and Wagner, 1965, Jiang and Guan, 2016]. Since directly evaluating chance constraints is often computationally difficult, classical data-driven approaches replace them with sampled approximations, including scenario programs that enforce all sampled constraints [Calafiore and Campi, 2006] and sample-average approximations (SAA) that permit a controlled fraction of violations [Luedtke and Ahmed, 2008]. For finite-support problems, such empirical chance constraints can also be encoded using mixed-integer formulations [Luedtke et al., 2010]. Geng and Xie [2019] provide a comprehensive review of these methods and their applications to decision-making in power systems. More recently, Zhao et al. [2024] introduce conformal predictive programming, which combines an empirical-quantile reformulation with independent conformal calibration to provide finite-sample a posteriori feasibility guarantees. Our formulation shares a similar setting where we utilize conformal prediction within the constraint of the optimization problem, but our core objective differs: rather than seeking a decision that satisfies a conformalized chance constraint, we aim to learn the geometry of a polyhedral prediction set that minimizes downstream robust decision loss while preserving finite-sample coverage validity.

B Proofs in Section 3.3

B.1 Proof of Lemma 1

Proof. By definition of the max-affine nonconformity score,

$$s(x, y; \theta) = \max_{k \in [K]} \{w_k^\top \epsilon(x, y) + b_k\}.$$

Therefore, for any $r \in \mathbb{R}$,

$$s(x, y; \theta) \le r \quad \Longleftrightarrow \quad w_k^\top \epsilon(x, y) + b_k \le r, \quad \forall k \in [K].$$

Stacking these K inequalities gives

$$W\epsilon(x, y) + b \le r 1_K.$$

□

B.2 Proof of Lemma 2

Proof. By definition in (6),

$$\begin{aligned} R(\mathcal{D}_n; \theta) &= \inf \left\{ r \in \mathbb{R} : \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i; \theta) \le r\} \ge \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\} \\ &= \min\{r \in \mathbb{R} : \sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i; \theta) \le r\} \ge c_n\}. \end{aligned} \tag{15}$$

By Lemma 1, $s(X_i, Y_i; \theta) \le r$ can be expressed as

$$W\epsilon_i + b \le r 1_K = (r + M(1 - t_i)) 1_K$$

when $t_i = 1$, while for $t_i = 0$ the constraint $W\epsilon_i + b \le (r + M)1_K$ holds for all $r \ge \underline{r}$, in particular for $r = R(\mathcal{D}_n; W, b)$, by the choice of M. Therefore,

$$\begin{aligned} R(\mathcal{D}_n; W, b) = \min_{r,t} \quad & r \\ \text{s.t.} \quad & W\epsilon_i + b \le (r + M(1 - t_i))1_K, \quad i \in [n], \\ & 1_n^\top t \ge c_n, \quad t \in \{0, 1\}^n. \end{aligned} \tag{16}$$

□

B.3 Proof of Proposition 1

Define an auxiliary variable r. The worst-case value

$$\max_{u \in \mathcal{Y}} g(z, u) \tag{17}$$

$$\text{s.t.} \quad s(x, u; \theta) \le r \tag{18}$$

is nondecreasing in r, because increasing r enlarges the feasible region of the inner maximization. Hence, problem (8)–(9) is equivalent to:

$$\min_{\substack{\theta, z, r \\ r \ge R(\mathcal{D}_n; \theta)}} \quad \max_{\substack{u \in \mathcal{Y} \\ s(x, u; \theta) \le r}} \quad g(z, u). \tag{19}$$

By Lemma 1, the score constraint in the inner maximization layer admits the linear representation

$$s(x, u; W, b) \le r \quad \Longleftrightarrow \quad W(u - \hat{f}(x)) + b \le r 1_K.$$

It remains to represent the constraint $r \ge R(\mathcal{D}_n; W, b)$ in the outer minimization layer. By Lemma 2, for a valid choice of M:

$$R(\mathcal{D}_n; W, b) = \min_{r,t} \left\{ r : W\epsilon_i + b \le (r + M(1 - t_i))1_K, \ i \in [n], \quad 1_n^\top t \ge c_n, \quad t \in \{0, 1\}^n \right\}.$$

Equivalently, $r \geq R(\mathcal{D}_n; W, b)$ holds if and only if there exists $t \in \{0,1\}^n$ such that

$$W\epsilon_i + b \leq (r + M(1-t_i))1_K, \qquad i \in [n], \qquad 1_n^\top t \geq c_n.$$

Therefore, (19) is equivalent to:

$$\begin{aligned} \min_{W,b,z,r,t} \quad & \max_{u \in \mathcal{Y}} \; g(z,u) \\ \text{s.t.} \quad & W\left(u - \hat{f}(x)\right) + b \leq r1_K, \\ & W\epsilon_i + b \leq (r + M(1-t_i))1_K, \qquad i \in [n], \\ & 1_n^\top t \geq c_n, \quad z \in \mathcal{Z}, \quad t \in \{0,1\}^n. \end{aligned} \tag{20}$$

This is exactly (12).

B.4 Proof of Corollary 1

Proof. By Proposition 1, the robust part of (12) can be written as

$$\min_{W,b,z,r,t} \left\{ \max_u \left\{ g_0(z) + z^\top Q u : W(u - \hat{f}(x)) + b \leq r1_K \right\} \right\},$$

together with the calibration constraints

$$W\epsilon_i + b \leq (r + M(1-t_i))1_K, \qquad i \in [n], \qquad 1_n^\top t \geq c_n, \qquad t \in \{0,1\}^n.$$

Since $g_0(z)$ does not depend on u, the inner maximization is equivalent to

$$g_0(z) + \max_u \left\{ z^\top Q u : W(u - \hat{f}(x)) + b \leq r1_K \right\}.$$

The constraint can be rewritten as

$$Wu \leq r1_K - b + W\hat{f}(x).$$

Therefore, the inner maximization is the linear program

$$\max_u \left\{ (Q^\top z)^\top u : Wu \leq r1_K - b + W\hat{f}(x) \right\}.$$

Introducing dual variables $\lambda \in \mathbb{R}_+^K$ associated with the constraints $Wu \leq r1_K - b + W\hat{f}(x)$, the dual problem is

$$\min_{\lambda \in \mathbb{R}_+^K} \left\{ \left(r1_K - b + W\hat{f}(x)\right)^\top \lambda : W^\top \lambda = Q^\top z \right\}.$$

Under strong duality, which holds whenever the inner linear program is feasible and has a finite optimum, the inner maximization is equal to its dual. Hence, the min–max problem is equivalent to

$$\begin{aligned} \min_{W,b,z,\lambda,r,t} \quad & g_0(z) + \left(r1_K - b + W\hat{f}(x)\right)^\top \lambda \\ \text{s.t.} \quad & W^\top \lambda = Q^\top z, \qquad \lambda \in \mathbb{R}_+^K, \\ & W\epsilon_i + b \leq (r + M(1-t_i))1_K, \qquad i \in [n], \\ & 1_n^\top t \geq c_n, \\ & z \in \mathcal{Z}, \quad t \in \{0,1\}^n. \end{aligned}$$

□

B.5 Polyhedral Conformal Set Learning

In this subsection, write $\mathcal{C}_\theta(x, r) \coloneqq \{y \in \mathcal{Y} : s(x, y; \theta) \leq r\}$ for the set at a specified radius.

This section elaborates on the two optimization algorithms described in the main text: the Column-and-Constraint Generation (CCG) method (Alg. 1) for general concave objectives, and the stochastic pinball-subgradient method (Alg. 2) for scalable training.

Algorithm 1 CCG for the general concave objective (12)

Require: $\mathcal{D}_n$, $\varepsilon > 0$, M

Initialize scenario set $\mathcal{S}_0 \leftarrow \emptyset$, iteration counter $h \leftarrow 0$
repeat
  Solve the restricted master problem MP_h in (21)
  Let $(\theta^h, z^h, r^h, t^h, \eta^h)$ be an optimal master solution
  Solve the separation problem SP_h in (22)
  Let ζ^h be its optimal value and y^{h+1} an optimizer
  if $\zeta^h - \eta^h \le \varepsilon$ **then**
    return $(\theta^h, z^h, r^h, t^h)$
  else
    Add adversarial scenario: $\mathcal{S}_{h+1} \leftarrow \mathcal{S}_h \cup \{y^{h+1}\}$
    Set $h \leftarrow h + 1$
  end if
until termination

Algorithm 2 Stochastic pinball-subgradient method for (23)

Require: Learning data $\mathcal{D}_n$, stepsizes $\{\eta_\ell\}$, minibatch size B, calibration penalty γ, $\tau_n = c_n/n$

Initialize $(\theta^0, r^0, z^0) \in \Theta \times \mathbb{R} \times \mathcal{Z}$, with $\theta^0 = (W^0, b^0)$;
for $\ell = 0, 1, 2, \ldots$ **do**
  Sample minibatch $\mathcal{B}_\ell \subseteq [n]$, $|\mathcal{B}_\ell| = B$;
  $y^\ell \in \arg\max_{y \in \mathcal{C}_{\theta^\ell}(x, r^\ell)} g(z^\ell, y)$; $\widehat{L}_{\mathcal{B}_\ell}(\theta, r) = \frac{1}{B} \sum_{i \in \mathcal{B}_\ell} \rho_{\tau_n}\big(S_i(\theta) - r\big)$;
  Compute a stochastic subgradient $\widehat{G}^\ell$ of $\max_{y \in \mathcal{C}_\theta(x,r)} g(z, y) + \gamma \widehat{L}_{\mathcal{B}_\ell}(\theta, r)$ at $(\theta^\ell, r^\ell, z^\ell)$, using y^ℓ and the multipliers of its constraints for the first term and (24) for the second;
  $(\theta^{\ell+1}, r^{\ell+1}, z^{\ell+1}) \leftarrow \mathrm{Proj}_{\Theta \times \mathbb{R} \times \mathcal{Z}}\big((\theta^\ell, r^\ell, z^\ell) - \eta_\ell \widehat{G}^\ell\big)$;
end for
return $\hat{\theta} = (\hat{W}, \hat{b})$.

CCG learning. Alg. 1 solves the problem (12) by iteratively growing a finite scenario set $\mathcal{S}_h \subseteq \mathcal{Y}$. At CCG iteration h, the restricted master problem is:

$$\begin{aligned} \mathrm{MP}_h: \quad \min_{W,b,z,r,t,\eta} \quad & \eta \\ \text{s.t.} \quad & g(z, y^j) \le \eta, \qquad \forall\, y^j \in \mathcal{S}_h, \\ & W(y^j - \hat{f}(x)) + b \le r 1_K, \qquad \forall\, y^j \in \mathcal{S}_h, \\ & W\epsilon_i + b \le \big(r + M(1 - t_i)\big) 1_K, \quad i \in [n], \\ & 1_n^\top t \ge c_n, \quad z \in \mathcal{Z}, \quad t \in \{0,1\}^n. \end{aligned} \tag{21}$$

Given a master solution $(\theta^h, z^h, r^h, t^h, \eta^h)$ with $\theta^h = (W^h, b^h)$ and conformal radius r^h, the subproblem finds the worst-case outcome in the current conformal set $\mathcal{C}_{\theta^h}(x, r^h)$:

$$\mathrm{SP}_h: \quad \zeta^h = \max_{y \in \mathcal{Y}} g(z^h, y) \quad \text{s.t.} \quad W^h \left(y - \hat{f}(x)\right) + b^h \le r^h 1_K. \tag{22}$$

When $g(z, \cdot)$ is convex, the maximizer over the polyhedral set can be chosen at an extreme point. CCG terminates when $\zeta^h - \eta^h \le \varepsilon$. Because the master problem contains n binary variables $\{t_i\}$, its complexity can grow with n, motivating the pinball surrogate below.

Stochastic learning with pinball loss. The exact formulation in (12) enforces empirical coverage through binary variables that select which calibration residuals are contained in the candidate set. To obtain a scalable continuous approximation, we replace this discrete quantile-selection mechanism with a pinball calibration penalty.

For a fixed set parameter θ, write

$$S_i(\theta) \coloneqq s(X_i, Y_i; \theta), \qquad i \in [n],$$

and let $\tau_n := c_n/n$. The empirical conformal radius is the smallest value of r such that at least c_n nonconformity scores are no larger than r. Equivalently, it is the smallest empirical τ_n-quantile of the scores $\{S_i(\theta)\}_{i=1}^n$. This quantile admits the standard pinball-loss characterization. Define

$$\rho_\tau(u) := \tau u_+ + (1-\tau)(-u)_+, \qquad u_+ := \max\{u, 0\},$$

and

$$L_n(\theta, r) := \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_n}\big(S_i(\theta) - r\big).$$

For fixed θ, the subdifferential with respect to r is

$$\partial_r L_n(\theta, r) = \frac{1}{n}\left[\#\{i : S_i(\theta) < r\} - \tau_n n,\ \#\{i : S_i(\theta) \le r\} - \tau_n n\right].$$

Hence,

$$0 \in \partial_r L_n(\theta, r) \quad \Longleftrightarrow \quad \#\{i : S_i(\theta) < r\} \le \tau_n n \le \#\{i : S_i(\theta) \le r\}.$$

Thus, any minimizer of $L_n(\theta, r)$ is an empirical τ_n-quantile of the nonconformity scores, and the smallest minimizer recovers the conformal radius used in Lemma 2. The pinball loss therefore provides a continuous relaxation of the binary quantile constraint in (12).

We solve the following surrogate problem:

$$\min_{\theta\in\Theta,\ r\in\mathbb{R},\ z\in\mathcal{Z}} \left\{ \max_{y\in\mathcal{C}_\theta(x,r)} g(z, y) + \gamma L_n(\theta, r) \right\}, \tag{23}$$

where $\gamma > 0$ controls the strength of the calibration penalty. The first term learns a polyhedral geometry that is favorable for the downstream robust decision, while the second term keeps the radius aligned with the target empirical conformal quantile. Compared with the exact mixed-integer formulation, this surrogate removes the binary variables and can be optimized with stochastic first-order methods.

For use in Alg. 2, the subgradient of the pinball term with respect to its argument is

$$\xi_i \in \partial\rho_{\tau_n}\big(S_i(\theta) - r\big) = \begin{cases} \tau_n, & S_i(\theta) > r, \\ [-(1-\tau_n),\ \tau_n], & S_i(\theta) = r, \\ -(1-\tau_n), & S_i(\theta) < r. \end{cases} \tag{24}$$

Consequently, scores above the current radius push r upward, while scores below the current radius push it downward. When differentiating with respect to θ, the same coefficients ξ_i weight the subgradients of the scores $S_i(\theta)$, allowing the set geometry and radius to be updated jointly.

C Proofs in Section 4

C.1 Proof for Lemma 3

Proof. Since the nonconformity score $s(\cdot, \cdot; \theta)$ is deterministic for fixed θ, and we assume that the augmented sample $\{(X_i, Y_i)\}_{i=1}^{n+1}$ is exchangeable, the induced score sequence $\{s(X_i, Y_i; \theta)\}_{i=1}^{n+1}$ is also exchangeable. Therefore, the rank of $s(X_i, Y_i; \theta)$ among the $n+1$ scores (with ties broken uniformly at random) is uniformly distributed over $\{1, \ldots, n+1\}$. By the definition of the conformal radius $R(\mathcal{D}_n; \theta)$ in (6), there is

$$\mathbb{P}\left\{s(X_{n+1}, Y_{n+1}; \theta) \le R(\mathcal{D}_n; \theta)\right\} \ge 1 - \alpha.$$

Using the definition of $\mathcal{C}(x; \mathcal{D}_n, \theta)$ completes the proof:

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}; \mathcal{D}_n, \theta)\right\} \ge 1 - \alpha.$$

□

C.2 Proof for Lemma 4

Proof. Since $\hat{\theta}(\mathcal{D}_n)$ is learned from $\mathcal{D}_n$, conditioning on $\hat{\theta}(\mathcal{D}_n)$ reveals information about the calibration samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, but not about the test point (X_{n+1}, Y_{n+1}). Consequently, the conditional score sequence

$$s(X_1, Y_1; \hat{\theta}), \ldots, s(X_n, Y_n; \hat{\theta}), s(X_{n+1}, Y_{n+1}; \hat{\theta}) \mid \hat{\theta} = \hat{\theta}(\mathcal{D}_n)$$

is generally no longer exchangeable. The loss of exchangeability arises from the asymmetric role of the test point relative to the calibration samples, which is precisely the type of nonexchangeability studied in Barber et al. [2023].

Following the argument of Theorem 2 in Barber et al. [2023], call an entry of a sequence of $n+1$ data points *strange* if its nonconformity score, computed with $\hat{\theta}$ fitted on the first n entries of that sequence, exceeds the c_n-th smallest score of the other n entries. At most $n + 1 - c_n \leq \alpha(n+1)$ entries of any sequence are strange, and the test point is not covered exactly when the $(n+1)$-th entry of $\mathcal{D}_{n+1}$ is strange. Since $\mathcal{D}_{n+1}^{(i)}$ has the same distribution as $\mathcal{D}_{n+1}$, the probability that the $(n+1)$-th entry of $\mathcal{D}_{n+1}^{(i)}$, namely (X_i, Y_i), is strange equals the miscoverage probability. Computing the scores with $\hat{\theta}$ fitted on the first n entries of $\mathcal{D}_{n+1}$ instead changes this probability by at most $d_{\mathrm{TV}}(\mathcal{D}_{n+1}, \mathcal{D}_{n+1}^{(i)})$, and turns it into the probability that the i-th entry of $\mathcal{D}_{n+1}$ is strange. Averaging over the $n+1$ entries,

$$\begin{aligned}
\mathbb{P}\{Y_{n+1} \notin \mathcal{C}(X_{n+1}; \mathcal{D}_n, \hat{\theta}(\mathcal{D}_n))\} &\leq \frac{1}{n+1}\,\mathbb{E}\big[\text{number of strange entries of } \mathcal{D}_{n+1}\big] \\
&\quad + \frac{1}{n+1}\sum_{i=1}^{n} d_{\mathrm{TV}}\left(\mathcal{D}_{n+1}, \mathcal{D}_{n+1}^{(i)}\right) \\
&\leq \alpha + \frac{1}{n+1}\sum_{i=1}^{n} d_{\mathrm{TV}}\left(\mathcal{D}_{n+1}, \mathcal{D}_{n+1}^{(i)}\right).
\end{aligned}$$

Taking complements gives the claim. When $\hat{\theta}$ does not depend on $\mathcal{D}_n$, both fits coincide, the distance is zero, and the usual conformal guarantee is recovered. This completes the proof. □

C.3 Proof for Theorem 1

Proof. Conditional on the learning data $\mathcal{D}_n$, the learned parameter $\hat{\theta}(\mathcal{D}_n)$ is fixed. By assumption, the recalibration dataset $\mathcal{D}_m = \{(X_j, Y_j)\}_{j=1}^{m}$ and the test point (X_{m+1}, Y_{m+1}) are exchangeable with $\mathcal{D}_n$. Therefore, conditional on $\mathcal{D}_n$, the augmented score sequence

$$s(X_1, Y_1; \theta), \ldots, s(X_m, Y_m; \theta), s(X_{m+1}, Y_{m+1}; \theta) \mid \theta = \hat{\theta}(\mathcal{D}_n)$$

is exchangeable. Using Lemma 3, there is:

$$\mathbb{P}\left\{Y_{m+1} \in \mathcal{C}(X_{m+1}; \mathcal{D}_m, \hat{\theta}(\mathcal{D}_n)) \mid \hat{\theta}(\mathcal{D}_n)\right\} \geq 1 - \alpha.$$

Taking expectation with respect to $\mathcal{D}_n$ yields

$$\mathbb{P}\left\{Y_{m+1} \in \mathcal{C}(X_{m+1}; \mathcal{D}_m, \hat{\theta}(\mathcal{D}_n))\right\} \geq 1 - \alpha.$$

□

C.4 Proof for Theorem 2

Proof of Theorem 2. We begin by defining the calibration error and the parameter error in (25) and (26), respectively:

$$\Delta(\mathcal{D}_m) = \mathcal{L}\left(\hat{\theta}(\mathcal{D}_n), R(\mathcal{D}_m; \hat{\theta}(\mathcal{D}_n))\right) - \mathcal{L}\left(\hat{\theta}(\mathcal{D}_n), R^\star(\hat{\theta}(\mathcal{D}_n))\right), \tag{25}$$

$$\Delta(\mathcal{D}_n) = \mathcal{L}\left(\hat{\theta}(\mathcal{D}_n), R^\star(\hat{\theta}(\mathcal{D}_n))\right) - \mathcal{L}\Big(\theta^\star, R^\star(\theta^\star)\Big). \tag{26}$$

Then, the suboptimality gap is upper bounded by the sum of two errors:

$$\Delta = \Delta(\mathcal{D}_m) + \Delta(\mathcal{D}_n) \leq |\Delta(\mathcal{D}_m)| + \Delta(\mathcal{D}_n).$$

We proceed with analyzing the bound by expanding the two terms respectively.

Calibration error $\Delta(\mathcal{D}_m)$ For any $\theta \in \Theta$, note that the data-driven conformal radius $R(\mathcal{D}_m;\theta)$ is the empirical counterpart of the population quantile $R^\star(\theta)$. Denote the cumulative distribution function (CDF) that they each represent by $\hat{F}_m : \mathbb{R} \to [0,1]$ and $F^\star : \mathbb{R} \to [0,1]$, respectively. Using the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [Dvoretzky et al., 1956], there is:

$$\mathbb{P}\left\{\|\hat{F}_m - F^\star\|_\infty \geq \epsilon\right\} \leq 2e^{-2m\epsilon^2} \Rightarrow \mathbb{P}\left\{\|\hat{F}_m - F^\star\|_\infty \leq \sqrt{\frac{\log(2/\delta)}{2m}}\right\} \geq 1-\delta.$$

Using (b) in Assumption 2, we know that $F^\star$ has a derivative at least κ near $R^\star(\theta)$. Since $R(\mathcal{D}_m;\theta)$ is the empirical quantile at level $c_m/m \in [1-\alpha, 1-\alpha+2/m)$, with $c_m = \lceil (m+1)(1-\alpha) \rceil$, inverting the CDFs gives, with probability greater than $1-\delta$ and for m large enough that the radii lie in the neighborhood in (b):

$$|R(\mathcal{D}_m;\theta) - R^\star(\theta)| \leq \frac{1}{\kappa}\left(\|\hat{F}_m - F^\star\|_\infty + \frac{2}{m}\right) \leq \frac{1}{\kappa}\left(\sqrt{\frac{\log(2/\delta)}{2m}} + \frac{2}{m}\right).$$

By using the Lipschitz condition (c) in Assumption 2 and applying this bound conditionally on $\mathcal{D}_n$, which is independent of $\mathcal{D}_m$, there is:

$$|\Delta(\mathcal{D}_m)| \leq L_r \cdot |R(\mathcal{D}_m;\hat{\theta}(\mathcal{D}_n)) - R^\star(\hat{\theta}(\mathcal{D}_n))| \leq \frac{L_r}{\kappa}\left(\sqrt{\frac{\log(2/\delta)}{2m}} + \frac{2}{m}\right).$$

Parameter error $\Delta(\mathcal{D}_n)$ The parameter error $\Delta(\mathcal{D}_n)$ can be further decomposed into three error terms:

$$\Delta(\mathcal{D}_n) = \Delta_1 + \Delta_2 + \Delta_3 \leq |\Delta_1| + \Delta_2 + |\Delta_3|,$$

where

$$\begin{aligned}
\Delta_1 &= \mathcal{L}\left(\hat{\theta}(\mathcal{D}_n), R^\star(\hat{\theta}(\mathcal{D}_n))\right) - \mathcal{L}\left(\hat{\theta}(\mathcal{D}_n), R(\mathcal{D}_n;\hat{\theta}(\mathcal{D}_n))\right), \\
\Delta_2 &= \mathcal{L}\left(\hat{\theta}(\mathcal{D}_n), R(\mathcal{D}_n;\hat{\theta}(\mathcal{D}_n))\right) - \mathcal{L}\left(\theta^\star, R(\mathcal{D}_n;\theta^\star)\right), \\
\Delta_3 &= \mathcal{L}\left(\theta^\star, R(\mathcal{D}_n;\theta^\star)\right) - \mathcal{L}\left(\theta^\star, R^\star(\theta^\star)\right).
\end{aligned}$$

The second error term Δ_2 can be bounded using condition (d) of Assumption 2:

$$\Delta_2 \leq \mathcal{L}\left(\hat{\theta}(\mathcal{D}_n), R(\mathcal{D}_n;\hat{\theta}(\mathcal{D}_n))\right) - \inf_{\theta\in\Theta} \mathcal{L}\left(\theta, R(\mathcal{D}_n;\theta)\right) \leq \epsilon_{\text{opt}}.$$

Using condition (c) of Assumption 2, the first and third error terms share the same upper bound:

$$\begin{aligned}
|\Delta_1| &\leq L_r \cdot \left|R^\star(\hat{\theta}(\mathcal{D}_n)) - R(\mathcal{D}_n;\hat{\theta}(\mathcal{D}_n))\right| \leq L_r \cdot \sup_{\theta\in\Theta} |R^\star(\theta) - R(\mathcal{D}_n;\theta)| \\
|\Delta_3| &\leq L_r \cdot |R^\star(\theta^\star) - R(\mathcal{D}_n;\theta^\star)| \leq L_r \cdot \sup_{\theta\in\Theta} |R^\star(\theta) - R(\mathcal{D}_n;\theta)|.
\end{aligned}$$

For any θ, the conformal radius $R(\mathcal{D}_n;\theta)$ is an empirical approximation to the population quantile $R^\star(\theta)$ using n samples. We redefine the cumulative distribution function (CDF) that they each represent by $\hat{F}_n(\cdot;\theta) : \mathbb{R} \to [0,1]$ and $F^\star(\cdot;\theta) : \mathbb{R} \to [0,1]$, respectively. Combined with the condition (a) of Assumption 2, using the uniform convergence of VC classes [Vapnik and Chervonenkis, 2015], there is:

$$\mathbb{P}\left\{\sup_{\theta\in\Theta,\, r\in\mathbb{R}} \left|\hat{F}_n(r;\theta) - F^\star(r;\theta)\right| \leq \epsilon\right\} \geq 1 - 8\left(\frac{en}{V}\right)^V \exp\left(-\frac{n\epsilon^2}{32}\right).$$

By reorganizing the terms, we conclude that with probability greater than $1-\delta$:

$$\sup_{\theta\in\Theta} \left\|\hat{F}_n(\cdot;\theta) - F^\star(\cdot;\theta)\right\|_\infty \leq \sqrt{\frac{32\left[V\log(en/V) + \log(8/\delta)\right]}{n}}.$$

Again, using (b) in Assumption 2 and the same argument with level $c_n/n \in [1-\alpha, 1-\alpha+2/n)$, we obtain that with probability $1-\delta$:

$$\sup_{\theta\in\Theta} |R(\mathcal{D}_n;\theta) - R^\star(\theta)| \leq \frac{1}{\kappa}\left(\sqrt{\frac{32\left[V\log(en/V) + \log(8/\delta)\right]}{n}} + \frac{2}{n}\right).$$

Therefore, combining the three error terms, there is with probability greater than $1-\delta$:

$$\Delta(\mathcal{D}_n) \leq \frac{2L_r}{\kappa}\left(\sqrt{\frac{32\left[V\log(en/V) + \log(8/\delta)\right]}{n}} + \frac{2}{n}\right) + \epsilon_{\text{opt}}.$$

Combining $\Delta(\mathcal{D}_m)$ and $\Delta(\mathcal{D}_n)$ Combining the calibration and parameter errors and using a union bound adjustment by setting δ to $\delta/2$ in the previous terms, there is, with probability greater than $1-\delta$:

$$\Delta \leq \frac{L_r}{\kappa}\sqrt{\frac{\log(4/\delta)}{2m}} + \frac{2L_r}{\kappa}\sqrt{\frac{32\left[V\log(en/V)+\log(16/\delta)\right]}{n}} + \frac{2L_r}{\kappa}\left(\frac{1}{m}+\frac{2}{n}\right) + \epsilon_{\text{opt}}.$$

Using the big-$\mathcal{O}_{\mathbb{P}}$ notation [Van der Vaart, 2000], we conclude that:

$$\Delta \leq \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{1}{m}} + \sqrt{\frac{\log n}{n}}\right) + \epsilon_{\text{opt}}.$$

□

D Additional Experimental Details and Results

D.1 Synthetic Experimental Details

Data splits and evaluation. For the synthetic experiments, we use a fixed four-way split: 1,200 samples for fitting the predictor $\hat{f}$, a learning set $\mathcal{D}_n$ of 300 samples for learning the set geometry θ, a calibration set $\mathcal{D}_m$ of 300 samples for determining the final conformal radius, and a held-out test set of 1,000 samples used only for evaluation. Results are averaged over 20 random seeds (five for the sensitivity analysis), each corresponding to an independent draw of the full dataset. Baselines that do not require a separate learning split (Bonferroni and conformal box/ball) use the combined pool $\mathcal{D}_n \cup \mathcal{D}_m$ of 600 samples for calibration, ensuring a fair comparison. We set the nominal miscoverage level to $\alpha = 0.10$ (corresponding to a target coverage of $1-\alpha = 0.90$) and use $K = 4$ half-planes for all polyhedral sets.

Synthetic data generation and predictor. For the synthetic tasks, covariates $X \in \mathbb{R}^4$ are drawn from $\mathcal{N}(0, I_4)$, and outcomes $Y \in \mathbb{R}^2$ follow a heteroscedastic nonlinear model:

$$Y = \mu(X) + \varepsilon(X),$$

where $\mu_1(X) = 0.75X_1 + 0.35\sin(X_2) - 0.20X_3$, $\mu_2(X) = -0.45X_1 + 0.55X_2^2 + 0.30X_4$, and $\varepsilon(X)$ is drawn from a zero-mean Gaussian with a covariate-dependent scale factor $0.80 + 0.45\,\sigma(X_1)$ (where σ is the sigmoid function) applied to a fixed anisotropic root-covariance matrix. An additional 6% heavy-tail contamination is added to each draw. This design produces residuals that are both anisotropic and heteroscedastic, so that the set shape has a meaningful effect on coverage and downstream cost.

We fit the predictor $\hat{f}$ by ridge regression on a nonlinear feature map $\phi(x) = [1,\, x^\top,\, x_1^2,\, x_2^2,\, \sin(x_1),\, \sin(x_2)]^\top \in \mathbb{R}^9$, with ridge penalty $\lambda = 10^{-3}$. Residuals $\epsilon_i = Y_i - \hat{f}(X_i)$ on the learning and calibration splits serve as the input to all conformal methods.

Synthetic objectives. We evaluate three two-dimensional robust decision tasks. The decision variable $z \in \Delta_2$ lives on the probability simplex, i.e. $z_1 + z_2 = 1$, $z_j \geq 0$.

Linear objective.

$$g(z,u) = (c+u)^\top z, \qquad c = (-0.08,\, 0).$$

The constant $c_1 = -0.08$ gives the first action a slight structural advantage; the sign of the realized outcome determines whether this advantage holds, making set orientation consequential.

Quadratic-regularized objective.

$$g(z,u) = (c+u)^\top z + \beta(z_1 - z_2)^2, \qquad c = (-0.08,\, 0), \quad \beta = 0.08.$$

The quadratic term penalizes unequal allocation, introducing a smoothly curved Pareto frontier between the two actions.

Newsvendor objective.

$$g(z,u) = \sum_{j=1}^{2}\left[c_o\max(z_j - u_j, 0) + c_u\max(u_j - z_j, 0)\right], \qquad c_o = 0.3, \quad c_u = 0.7.$$

Here $z \in \Delta_2$ is the order quantity and u is the realized demand. Overage ($c_o = 0.3$) and underage ($c_u = 0.7$) costs are asymmetric. The loss is piecewise linear and nonsmooth; its worst-case value depends on which tail of the residual distribution the uncertainty set exposes, making the learned set shape particularly consequential.

D.2 Real Data Experimental Details

Data splits. We use the UCI Energy Efficiency dataset [Tsanas and Xifara, 2012], also used by Wang et al. [2023] for conformal prediction evaluation. We partition the data without replacement into 320 training, 160 learning, 144 recalibration, and 144 test observations for each seed. The four subsets are mutually disjoint within each run, while the 80 seeds correspond to independently repeated random partitions of the same dataset; we use more partitions than for the synthetic tasks because each one has only 144 calibration and 144 test observations.

Evaluation and objectives. Unlike prior evaluations of this dataset that report only coverage and set volume, we measure downstream worst-case cost and regret. The covariates $X \in \mathbb{R}^8$ encode eight building-design attributes (relative compactness, surface area, wall area, roof area, height, orientation, glazing area, glazing-area distribution), and $Y = (Y_1, Y_2)$ are heating and cooling loads (kWh/m^2). We pair the dataset with the linear objective $g(z, u) = (c + u)^\top z$, $c = (-0.08, 0)$, where $z \in \Delta_2$ allocates a unit budget between heating (z_1) and cooling (z_2) control. We fit $\hat{f}$ by the same ridge feature regression as in the synthetic experiments.

D.3 Baseline Implementation Details

We compare against decision-agnostic conformal baselines, including Bonferroni boxes [Dunn, 1961], conformal boxes or balls [Vovk et al., 2005], and PCP [Wang et al., 2023]. We also include decision-aware baselines based on F-CROMS [Bao et al., 2025] and PICNN scores [Yeh et al., 2024]. For ablations, "Polyhedron" calibrates a polyhedral set without using the downstream loss, "Polyhedron (min size)" learns the polyhedral set by minimizing volume, and "Ours (w/o recal)" uses the same learned geometry $\hat{\theta}(\mathcal{D}_n)$ as our method but sets the radius on $\mathcal{D}_n$ itself, without an independent post-selection calibration step. All methods are evaluated over multiple random seeds using the same train–learn–calibrate–test splits whenever applicable.

Ours. For each learning problem, we first attempt the exact mixed-integer quadratically constrained program (MIQCP) of Corollary 1 with Gurobi [Gurobi Optimization, LLC, 2026] under a 30-minute wall-clock limit. The MIQCP averages the robust loss over the first 120 learning contexts, i.e., it approximates the empirical objective defining $\hat{\theta}(\mathcal{D}_n)$, and is posed over the compact class of paired normals $W = (a_1, a_2, -a_1, -a_2)^\top$ with $A = [a_1; a_2]$ symmetric, $2\|A\|_F^2 = 1$, $\lambda_{\min}(A) \geq 0.15$, and $b = 0$; we substitute $\mu = r\lambda$ for the dual variables, and the quadratic-regularized objective only adds the convex term $\beta(z_1 - z_2)^2$ to $g_0(z)$. We use the MIQCP solution only if Gurobi certifies optimality (relative gap at most 0.5% or absolute gap at most 10^{-3}). Otherwise, and whenever the reformulation does not apply (newsvendor loss, $K \neq 4$, or $d > 2$), we use the pinball-loss gradient method: for $d = 2$, gradient descent on the smoothed pinball surrogate over the rotation, aspect ratio, and radius of a regular K-gon (five starts, 80 iterations); for $d > 2$, the stochastic pinball-subgradient method of Alg. 2 over $W = DM$, where D is a fixed positively spanning set of K directions and $M \in \mathbb{R}^{d \times d}$ is learned. No variant selects from a fixed library of candidate sets. In the main experiments, Gurobi certified the MIQCP within 1–29 minutes in 50 of the 80 energy-efficiency runs and within 3–9 minutes in 3 of the 20 linear-objective runs; all quadratic-regularized runs and the remaining linear and energy-efficiency runs reached the limit with absolute gaps between 0.001 and 0.15, and the newsvendor task has no MIQCP reformulation. In the sensitivity study, the MIQCP was certified within 1–2 minutes in two of the five runs with $n = 120$ and reached the limit at all other $K = 4$, $d = 2$ points. In all cases the final radius is recalibrated on $\mathcal{D}_m$, and reported times include the full geometry-learning time, including the 30-minute MIQCP attempt whenever it is made.

F-CROMS. F-CROMS [Bao et al., 2025] selects a conformal prediction set from a finite library of candidate sets, each induced by a different nonconformity score. In the original paper the library consists of model-induced scores; here we instantiate the library with randomly generated halfspace cuts to give the method a fair chance to discover polyhedral geometries similar to ours: each candidate

intersects K halfspaces with random unit normals, $s_k(x, y) = \max_{j \le K} w_{kj}^\top (y - \hat{f}(x))$. For each k we calibrate the threshold on the learning split $\mathcal{D}_n$ via split conformal prediction, obtaining a polyhedral prediction set. F-CROMS then picks the candidate $k^\star$ whose set minimizes the empirical worst-case robust objective on $\mathcal{D}_n$ and recalibrates it on $\mathcal{D}_m$. We evaluate two library sizes: F-CROMS (10) with 10 candidates and F-CROMS (20) with 20. This implementation is faithful to the original method's spirit: it can only select among a discrete pool of pre-specified directions and cannot optimize the halfspace orientation continuously. Our method, by contrast, optimizes the set geometry over a continuous parameter space rather than selecting it from a finite pool.

PCP. PCP [Wang et al., 2023] draws N samples $\hat{y}_1, \ldots, \hat{y}_N$ from a conditional generative model of Y given $X = x$, uses the distance to the nearest sample, $s(x, y) = \min_{k \le N} \|y - \hat{y}_k\|_2$, as the nonconformity score, and returns the union of the N balls of radius $\hat{q}$ centered at the samples. Following the authors' setting for two-dimensional outcomes, the generative model is a mixture density network (10 Gaussian components with diagonal covariances, two hidden layers of width 16) for $Y - \hat{f}(X)$ given $\hat{f}(X)$, fitted by maximum likelihood on $\mathcal{D}_n$ with early stopping on a held-out fifth. We use the high-density version proposed in the same paper (HD-PCP), which draws $\lceil N/0.8 \rceil$ samples per covariate and keeps the $N = 1000$ with the highest estimated density; without this filtering, the far-tail samples dominate the worst case. The radius $\hat{q}$ is calibrated on $\mathcal{D}_m$. Because the losses are convex in y, the worst case over the union of balls equals the worst case over its convex hull, whose support function is $\max_k a^\top \hat{y}_k + \hat{q}\|a\|_2$. We report the area of the union, averaged over the calibration covariates.

Yeh et al. Yeh et al. [2024] train the uncertainty set end-to-end by parameterizing the nonconformity score via a partially input-convex neural network (PICNN), which guarantees that the sublevel set $\{y : s(x, y; \theta) \le \tau\}$ is convex in y for every x. Gradients with respect to θ flow through the adversarial argmax via KKT conditions and implicit differentiation, enabling joint learning of set shape and size. We implement this baseline following the authors' code, using their PICNN score class with a two-hidden-layer architecture (width 64, ReLU activations) that takes $\hat{f}(x)$ as input and the standardized residual as its convex argument, plus the term $\|y\|_\infty$ that the authors propose to keep the sublevel sets bounded (without it, the learned sets were unbounded on the energy data). The score is pretrained with the authors' estimate-then-optimize energy-based fit (Adam, step size 10^{-2}, 50 epochs) and then trained end-to-end on the realized task loss: in each minibatch, half of the points set the conformal threshold, and the robust decisions for the other half are differentiated through the dual linear program with `cvxpylayers` [Agrawal et al., 2019] (for the newsvendor loss, each affine piece is dualized separately). We use Adam with the step size selected from $\{10^{-3}, 10^{-4}\}$ and early stopping (at most 100 epochs) on a held-out fifth of $\mathcal{D}_n$. The conformal radius is then calibrated on the same held-out split as ours so that the final coverage guarantee is comparable; since these sets vary with x, we report their volume averaged over the calibration covariates. This design is sensible when $g(z, \cdot)$ is differentiable and the gradient signal can propagate through the PICNN; it produces general convex sets and is thus potentially more expressive than our polyhedral parameterization.

D.4 Comparison of the Pinball-Loss Approximation and Exact MIQCP

We additionally compare the exact MIQCP formulation with its pinball-loss approximation as the learning-set size $n = |\mathcal{D}_n|$ increases. This experiment isolates scalability with respect to sample size at a fixed outcome dimension. For each value of n, we run both formulations using the same five random seeds and report their average runtimes and worst-case objectives. The performance gap is the percentage increase in the worst-case objective of the pinball solution relative to the solution returned by the MIQCP formulation. Both solutions undergo the same independent conformal recalibration before evaluation.

This study uses the fixed-context specialization of the MIQCP at the mean learning context and the stochastic pinball-subgradient method of Alg. 2, both over the compact paired-normal class of Appendix D.3. Each MIQCP is run with a 30-minute limit, within which Gurobi certified global optimality (relative gap at most 0.5%) in all 30 runs; Table 1 reports the number of certified runs for each n.

The exact MIQCP is faster for very small learning sets but becomes substantially more expensive as n increases. Its average runtime grows from 0.10 seconds at $n = 20$ to 325.69 seconds at $n = 600$

Table 1: Comparison of the exact MIQCP formulation and the pinball-loss approximation. Runtimes are averaged over five seeds. The time ratio is MIQCP runtime divided by pinball runtime, so a value greater than one indicates that the pinball approximation is faster.

n	Certified runs	MIQCP time (s)	Pinball time (s)	Time ratio	Performance gap
20	$5/5$	0.10	1.68	$0.06\times$	1.0%
40	$5/5$	0.13	1.72	$0.08\times$	9.3%
80	$5/5$	0.87	2.04	$0.43\times$	7.3%
160	$5/5$	4.00	2.05	$1.95\times$	7.6%
300	$5/5$	134.12	2.00	$67.22\times$	4.3%
600	$5/5$	325.69	2.00	$163.20\times$	4.7%

(up to 1234 seconds for a single run), whereas the pinball runtime remains stable between 1.68 and 2.05 seconds. At $n = 300$ and $n = 600$, the pinball approximation is approximately 67 and 163 times faster, respectively. Meanwhile, the increase in the worst-case objective remains below 10% in all settings and falls below 5% for the two largest instances. These results indicate that the exact formulation remains practical for small instances, while the pinball approximation offers substantially more stable computation with a modest loss in objective performance.

D.5 Additional Results

Tables 2 and 3 provide additional results. Across the three synthetic objectives, the proposed recalibrated method maintains empirical coverage close to the nominal 90% level while achieving the best or near-best worst-case cost. In the linear and quadratic-regularized objectives, our method substantially improves worst-case cost relative to standard conformal boxes, conformal balls, and fixed polyhedral baselines; it also improves on PCP in the linear case and matches it in the quadratic-regularized case. For the newsvendor task, PCP attains the lowest worst-case cost, but our method remains competitive while using a decision-aware learned geometry. These results further show that smaller volume alone does not necessarily imply better downstream performance: PCP and the minimum-size polyhedron often produce compact sets but can incur larger worst-case costs.

The real-data energy-efficiency experiment in Table 3 shows a similar pattern. Our method achieves the lowest worst-case cost among all methods with a coverage guarantee, while maintaining coverage near the target level within sampling variability. Compared with Ours (w/o recal), the recalibrated version has nearly identical regret and runtime, but provides the statistically valid two-stage construction used in the main method. Overall, these results support the conclusion that learning the geometry of the conformal set in a decision-aware manner improves robust downstream performance without sacrificing empirical coverage.

Table 2: Comparison across synthetic objectives. Mean $\pm$ std over 20 seeds; $n = m = 300$, $K = 4$, $1-\alpha = 0.90$. WC cost and regret are in units of $\times 10^{-1}$; time is total wall-clock seconds per run. For Ours and Ours (w/o recal), time includes the 30-minute MIQCP attempt whenever it is made (Appendix D.3).

Task	Method	Coverage	Volume	WC cost	Regret	Time (s)
Linear objective	Bonferroni box	0.914 ± 0.019	0.86 ± 0.17	1.00 ± 0.53	0.27 ± 0.13	< 0.01
	Conformal (ℓ_∞)	0.902 ± 0.016	1.13 ± 0.17	1.48 ± 0.57	0.15 ± 0.09	< 0.01
	Conformal (ℓ_2)	0.902 ± 0.019	1.05 ± 0.07	1.78 ± 0.39	0.34 ± 0.09	< 0.01
	PCP	0.896 ± 0.023	0.60 ± 0.11	0.74 ± 0.75	0.44 ± 0.21	1.3
	F-CROMS (10)	0.898 ± 0.019	3.38 ± 4.64	1.64 ± 0.45	0.69 ± 0.41	0.28
	F-CROMS (20)	0.897 ± 0.022	2.30 ± 3.49	1.30 ± 0.55	0.62 ± 0.50	0.55
	Yeh et al.	0.898 ± 0.014	0.88 ± 0.18	1.46 ± 0.74	0.34 ± 0.19	881
	Polyhedron	0.902 ± 0.020	0.68 ± 0.15	1.40 ± 0.84	0.49 ± 0.31	< 0.01
	Polyhedron (min size)	0.902 ± 0.020	0.59 ± 0.08	1.29 ± 1.06	0.49 ± 0.32	0.05
	Ours (w/o recal)	0.905 ± 0.020	0.76 ± 0.15	$\underline{0.64} \pm 0.61$	0.34 ± 0.24	1599
	Ours	0.898 ± 0.021	0.72 ± 0.11	$\mathbf{0.54} \pm 0.48$	0.32 ± 0.23	1599
Quadratic-regularized objective	Bonferroni box	0.914 ± 0.019	0.86 ± 0.17	1.76 ± 0.53	0.27 ± 0.13	< 0.01
	Conformal (ℓ_∞)	0.902 ± 0.016	1.13 ± 0.17	2.24 ± 0.57	0.16 ± 0.08	< 0.01
	Conformal (ℓ_2)	0.902 ± 0.019	1.05 ± 0.07	2.36 ± 0.38	0.33 ± 0.08	< 0.01
	PCP	0.896 ± 0.023	0.60 ± 0.11	$\underline{1.38} \pm 0.78$	0.42 ± 0.21	1.7
	F-CROMS (10)	0.898 ± 0.019	3.38 ± 4.64	2.19 ± 0.43	0.63 ± 0.34	1.1
	F-CROMS (20)	0.897 ± 0.022	2.20 ± 3.43	1.89 ± 0.54	0.57 ± 0.43	2.2
	Yeh et al.	0.900 ± 0.016	0.85 ± 0.16	1.87 ± 0.54	0.27 ± 0.14	1387
	Polyhedron	0.902 ± 0.020	0.68 ± 0.15	2.03 ± 0.84	0.48 ± 0.29	< 0.01
	Polyhedron (min size)	0.902 ± 0.020	0.59 ± 0.08	1.94 ± 1.06	0.48 ± 0.30	0.04
	Ours (w/o recal)	0.906 ± 0.022	0.77 ± 0.16	1.50 ± 0.48	0.34 ± 0.28	1967
	Ours	0.898 ± 0.022	0.73 ± 0.12	$\mathbf{1.37} \pm 0.40$	0.33 ± 0.28	1967
Newsvendor	Bonferroni box	0.914 ± 0.019	0.86 ± 0.17	8.57 ± 0.38	0.18 ± 0.07	< 0.01
	Conformal (ℓ_∞)	0.902 ± 0.016	1.13 ± 0.17	8.97 ± 0.37	0.19 ± 0.07	< 0.01
	Conformal (ℓ_2)	0.902 ± 0.019	1.05 ± 0.07	7.95 ± 0.28	0.15 ± 0.05	< 0.01
	PCP	0.896 ± 0.023	0.60 ± 0.11	$\mathbf{7.83} \pm 0.39$	0.21 ± 0.05	2.0
	F-CROMS (10)	0.898 ± 0.022	1.12 ± 0.25	8.78 ± 0.73	0.19 ± 0.07	0.37
	F-CROMS (20)	0.899 ± 0.022	1.13 ± 0.26	8.75 ± 0.71	0.21 ± 0.10	0.71
	Yeh et al.	0.901 ± 0.016	0.99 ± 0.23	8.65 ± 0.63	0.19 ± 0.08	1782
	Polyhedron	0.902 ± 0.020	0.68 ± 0.15	8.15 ± 0.44	0.19 ± 0.05	< 0.01
	Polyhedron (min size)	0.902 ± 0.020	0.59 ± 0.08	8.27 ± 0.47	0.21 ± 0.06	0.03
	Ours (w/o recal)	0.907 ± 0.019	0.81 ± 0.12	7.89 ± 0.39	0.16 ± 0.05	25
	Ours	0.902 ± 0.021	0.78 ± 0.14	$\underline{7.83} \pm 0.41$	0.16 ± 0.05	25

Table 3: Energy Efficiency (real data). Mean $\pm$ std over 80 seeds; $n = 160$, $m = 144$, $K = 4$, $1-\alpha = 0.90$. WC cost and regret are in units of $\times 10^{-1}$; time is total wall-clock seconds per run. For Ours and Ours (w/o recal), time includes the 30-minute MIQCP attempt whenever it is made (Appendix D.3).

Method	Coverage	Volume	WC cost ($\times 10^{-1}$)	Regret ($\times 10^{-1}$)	Time (s)
Bonferroni box	0.910 ± 0.029	0.81 ± 0.09	2.79 ± 1.10	0.47 ± 0.09	< 0.01
Conformal (ℓ_∞)	0.907 ± 0.026	0.87 ± 0.10	3.86 ± 1.09	0.38 ± 0.07	< 0.01
Conformal (ℓ_2)	0.905 ± 0.027	0.88 ± 0.08	3.39 ± 1.04	0.76 ± 0.06	< 0.01
PCP	0.906 ± 0.032	0.44 ± 0.08	3.02 ± 1.11	0.50 ± 0.09	1.2
F-CROMS (10)	0.907 ± 0.032	2.87 ± 4.98	2.75 ± 1.15	0.68 ± 0.14	0.23
F-CROMS (20)	0.907 ± 0.032	2.93 ± 7.53	2.66 ± 1.10	0.64 ± 0.13	0.44
Yeh et al.	0.906 ± 0.037	0.74 ± 0.35	3.60 ± 1.32	0.44 ± 0.09	408
Polyhedron	0.902 ± 0.037	0.73 ± 0.13	4.33 ± 1.15	0.79 ± 0.11	< 0.01
Polyhedron (min size)	0.906 ± 0.034	0.77 ± 0.17	4.09 ± 1.31	0.71 ± 0.20	0.02
Ours (w/o recal)	0.894 ± 0.036	1.09 ± 0.26	$\mathbf{2.31} \pm 1.07$	0.56 ± 0.10	860
Ours	0.900 ± 0.033	1.15 ± 0.35	$\underline{2.36} \pm 1.09$	0.56 ± 0.10	860

NeurIPS Paper Checklist

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We discuss the main claims in abstract and introduction.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss limitations in the conclusion.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, we provide details of assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details of experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No, but we plan to make the implementation open access.

   Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include train-calibrate-test details in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, the tables in the appendix report the mean and standard deviation over random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report computation time and implementation details in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We find no clear path to negative social impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper has no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Properly cited for existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, clearly communicated.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper has no human subject.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No need for IRB for our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This paper does not involve LLM as an important component.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.