Conformalized Decision Risk Assessment

Wenbin Zhou
Heinz College of Information Systems and Public Policy, Carnegie Mellon University, wenbinz2@andrew.cmu.edu

Agni Orfanoudaki
Saïd Business School, University of Oxford, agni.orfanoudaki@sbs.ox.ac.uk

Shixiang Zhu
Heinz College of Information Systems and Public Policy, Carnegie Mellon University, shixianz@andrew.cmu.edu

In many operational settings, decision-makers must commit to actions before uncertainty resolves, but existing optimization tools rarely quantify how consistently a chosen decision remains optimal across plausible scenarios. This paper introduces CREDO—Conformalized Risk Estimation for Decision Optimization, a distribution-free framework that quantifies the probability that a prescribed decision remains (near-)optimal across realizations of uncertainty. CREDO reformulates decision risk through the inverse feasible region—the set of outcomes under which a decision is optimal—and estimates its probability using inner approximations constructed from conformal prediction balls generated by a conditional generative model. This approach yields finite-sample, distribution-free lower bounds on the probability of decision optimality. The framework is model-agnostic and broadly applicable across a wide range of optimization problems. Extensive numerical experiments demonstrate that CREDO provides accurate, efficient, and reliable evaluations of decision optimality across various optimization settings.

Key words: Conformal prediction, Inverse optimization, Risk analysis, Human-centered decision-making

1. Introduction

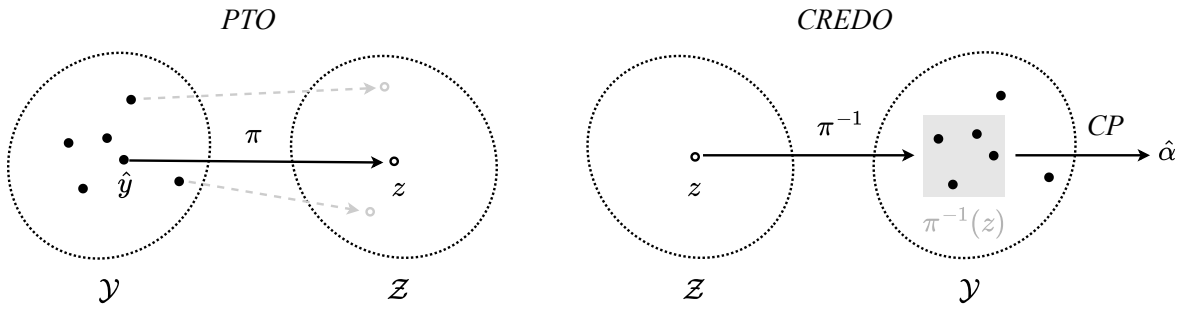
Organizations routinely make consequential decisions in the presence of substantial uncertainty (Kochenderfer 2015). Utilities must plan infrastructure upgrades without knowing future load growth or the intensity of extreme weather events (Chen et al. 2025); hospitals must allocate limited staff and beds before patient needs are realized (Kim and Mehrotra 2015); and public agencies must design policies whose impacts depend on uncertain behavioral and socioeconomic responses (Zhu et al. 2022). In such settings, decision-makers are tasked with selecting the best action by solving an optimization problem that will remain effective across a range of possible scenarios. Classical approaches address this challenge by representing uncertainty through forecasts or sampled scenarios and subsequently optimizing with respect to this surrogate representation. Two common instantiations of this paradigm are predict-then-optimize (Bertsimas and Kallus 2020, Elmachtoub and Grigas 2022), where decisions are optimized against a point forecast, and scenario-based stochastic programming (Bertsimas et al. 2018), where decisions are chosen to minimize expected cost across empirical samples. These tools have formed the backbone of prescriptive analytics in operations and have been successfully deployed across numerous application domains (Bertsimas et al. 2021, Tian et al. 2023).

Despite their widespread use, these approaches provide a limited view of the decision landscape. Many operational problems admit multiple actions that perform nearly equivalently across plausible realizations of uncertainty (Topkis 1998, Bertsimas and Sim 2004). For example, several investment plans may harden the electric grid to similar degrees under different storm patterns (Lombardi et al. 2025), or multiple staff schedules may achieve comparable service quality across demand scenarios (DeCarolis 2011, Palmintier 2014). Yet classical optimization pipelines return only a single recommended action, offering little visibility into whether this choice is robust or whether alternative decisions are nearly as effective (Li and Zhu 2024, Zhang et al. 2025). As a result, decision-makers must often rely on informal heuristics or domain intuition to assess the reliability of the prescribed solution and to understand how frequently it would remain optimal as the environment varies (Delarue et al. 2025). This gap between the deterministic nature of traditional models and the practical need to evaluate the reliability of decisions highlights the need for new rigorous tools that more directly quantify decision reliability.

Our work addresses this gap by adopting a complementary perspective to the established literature. Rather than collapsing uncertainty into a single deterministic representation, we focus on *quantifying the likelihood that a given decision remains (near-)optimal under different realizations of the environment*. The central question shifts from "What does the model recommend?" to "How reliably will this prescribed decision perform under varying conditions?" Under this view, analytical tools serve not as automatic decision engines but as decision-support systems that assess how robust a candidate solution is to underlying data variability. Such information is valuable in operational human-in-the-loop contexts where human judgment, regulatory oversight, or organizational risk tolerance ultimately shape the chosen action (Dietvorst et al. 2018, Grand-Clément and Pauphilet 2024). By quantifying the reliability of a decision, the framework provides interpretable evidence that supports accountability, stakeholder communication, and more informed strategic planning.

To operationalize this perspective, we introduce CREDO—Conformalized Risk Estimation for Decision Optimization, a framework that quantifies, for any candidate decision, a distribution-free lower bound on the probability that it remains (near-)optimal. The proposed approach integrates inverse optimization (Chan et al. 2025) with distribution-free uncertainty quantification via conformal prediction (CP) (Angelopoulos and Bates 2021) (see Figure 1 for an illustration). The core challenge is to evaluate the likelihood that an uncertain scenario falls within the decision's inverse feasible region, namely the set of scenarios under which that decision remains optimal or near-optimal, and directly characterizing this region is generally intractable due to its implicit and complex geometry. CREDO circumvents this challenge by constructing inner approximations of the inverse feasible region using multiple CP balls generated from a conditional generative model. Each ball constitutes a valid prediction region for the outcome with guaranteed coverage. By shrinking the radius so that the ball lies entirely within the inverse feasible region, we obtain a certified lower bound on the decision's optimality probability given by the corresponding coverage level. Averaging

Figure 1 Our proposed framework (CREDO) compared against conventional predict-then-optimize (PTO).



Note. For a given decision z in feasible region $\mathcal{Z}$, CREDO first finds the inverse feasible region $\pi^{-1}(z)$ within scenario space $\mathcal{Y}$ and then estimates the associated decision risk by quantifying the probability that the realized y lies within this region using conformal prediction (CP). In contrast, the conventional predict-then-optimize takes an estimated scenario $\hat{y}$ as input to policy π and outputs a single prescribed decision without a reliability assessment.

these calibrated bounds yields a distribution-free, finite-sample valid estimator of decision risk. CREDO is compatible with modern generative forecasting models, supports multiple conformal radius constructions, and applies broadly to convex decision problems. We establish theoretical validity for the estimator, develop efficient computational procedures, and demonstrate through extensive numerical experiments that CREDO provides accurate, interpretable, and reliable assessments of decision reliability across various optimization settings. The key contributions of our work are:

- *Problem Formulation*: We formulate the decision risk assessment as estimating the probability that a candidate decision remains (near-)optimal under uncertainty. By expressing this probability through inverse feasible regions, we convert an intractable evaluation problem into a tractable probability estimation task.
- *CREDO Framework*: We introduce CREDO, a distribution-free framework based on CP that delivers valid and finite-sample lower bounds on decision optimality, and establish marginal conservativeness, asymptotic consistency, and accuracy guarantees.
- *Computational Algorithm*: We derive a closed-form estimator for linear programs and develop an efficient computational procedure applicable to general convex optimization problems.
- *Empirical Validation*: Through controlled experiments featuring various optimization paradigms such as linear, quadratic, and second-order conic programs, as well as a real-world power grid planning application, we verify CREDO's theoretical guarantees and demonstrate its practical advantages. Our method achieves 100% empirical validity under conservative radius choices, superior true positive rates (up to 78.75% improvement over baselines), consistently selects decisions with higher empirical confidence rankings than existing prescriptive methods, and is superior compared to various ablation variants.

The remainder of this paper is organized as follows. Section 2 reviews the related literature and positions our contributions. Section 3 formalizes the decision risk assessment problem and provides motivating examples. In Section 4, we present the CREDO methodology, combining inverse optimization with generative CP.

Section 5 establishes theoretical guarantees including validity, consistency, and true positive rate analysis. Section 6 details the proposed computational implementation of CREDO, while Section 7 demonstrates its effectiveness through three numerical experiments. Section 8 concludes with discussions of implications and future work. Technical proofs and additional experiment details are included in the Appendix.

2. Related Work

Our work draws from and contributes to four distinct research streams: (*i*) decision-making under uncertainty in operations research, (*ii*) distribution-free uncertainty quantification through CP, (*iii*) inverse optimization for understanding decision structures, and (*iv*) integration of CP and human-in-the-loop in robust decision making. In what follows, we position the CREDO approach relative to these lines of research.

Classical approaches to optimization under uncertainty can be categorized by their treatment of unknown parameters. Stochastic optimization seeks decisions that minimize the expected value of the objective function across probabilistic scenarios, typically implemented through sample average approximation over historical data (Shapiro et al. 2021, Kleywegt et al. 2002, Lan et al. 2025). The predict-then-optimize (PTO) paradigm offers an alternative by first estimating unknown parameters via predictive models, then solving the resulting deterministic optimization problem (Bertsimas and Kallus 2020, Elmachtoub and Grigas 2022). Acknowledging that probability distributions themselves may be misspecified, robust optimization (RO) is employed to hedge against worst-case realizations within specified uncertainty sets (Bertsimas and Thiele 2006, Ben-Tal and Nemirovski 2002). Distributionally robust optimization (DRO) generalizes this concept further, considering worst-case probability distributions rather than point realizations, thereby balancing conservativeness with statistical plausibility (Delage and Ye 2010, Rahimian and Mehrotra 2022). Recognizing that prediction errors can compound when models are trained separately from their downstream use, decision-focused learning (DFL) trains predictive models end-to-end by differentiating through the optimization layer, directly minimizing downstream decision costs rather than prediction errors (Amos and Kolter 2017, Mandi et al. 2024). While these prescriptive paradigms effectively recommend decisions, our work adopts a parallel goal: rather than *prescribing* a decision, we focus on *auditing* decisions by providing a quantitative, data-driven assessment of the risk associated with taking a given decision. This perspective complements existing decision-prescription methods and even human judgment by adding an additional layer of transparency and accountability, thereby supporting more reliable and defensible decision-making.

Methodologically, our work builds upon the conformal prediction (CP) literature, a distribution-free approach for uncertainty quantification through calibrated prediction sets (Shafer and Vovk 2008, Vovk et al. 2005). Multiple strands of recent research are relevant to our work. First, generative conformal methods draw multiple samples from generative models to construct both the calibration sets and the resulting CP regions, thus significantly improving efficiency (*i.e.*, tightness of the prediction sets) especially when the data distribution is highly complex or dispersed (Zheng and Zhu 2024, Zhou et al. 2025). Second, inverse CP

estimates miscoverage rates for fixed prediction sets by identifying the smallest miscoverage level at which conformal regions are contained within the target set (Prinster et al. 2023, Singh et al. 2024). Advances in e-value CP strengthen this approach by enabling post-hoc selection of the miscoverage rate while preserving finite-sample validity at the theoretical level (Vovk 2025, Gauthier et al. 2025b). Our algorithm synthesizes both advances by leveraging generative models to enhance decision-risk estimation accuracy and subsequently employing generalized inverse CP to assess the decision risk. However, we emphasize that our problem setting is fundamentally different and cannot be addressed by direct application of these methods. Substantive reformulation is required to adapt their underlying ideas to the decision-auditing context.

Our work also connects to inverse optimization (IO), which seeks to infer the unknown parameters of an optimization problem from observed solutions (Ahuja and Orlin 2001, Chan et al. 2025, Aswani et al. 2018, Zattoni Scroccaro et al. 2025, Besbes et al. 2025). Recent work on Conformal-IO (Lin et al. 2024, Chan et al. 2024) combines IO with CP to prescribe robust decisions by first estimating decision-makers' preference parameters from observed choices and then applying robust forward optimization using CP to construct an uncertainty set over these parameters. While we similarly integrate IO and CP, our learning setting fundamentally differs. We assume access to samples of the unknown parameters rather than the actual decisions. This leads us to employ inverse CP to assess decision reliability (Singh et al. 2024) rather than forward CP to prescribe decisions. Our focus on evaluation rather than prescription distinguishes our approach from Conformal-IO's goal of generating policy recommendations. Meanwhile, though efficient IO algorithms exist for structured problems like linear, integer, and convex programs (Tavaslıoğlu et al. 2018, Schaefer 2009, Iyengar and Kang 2005), the lack of a unified framework for general IO leads to the adoption of structured formulations for tractability (Chan et al. 2024). This motivates our focus on linear models as part of our theoretical analysis, allowing us to derive an efficient closed-form solution for the proposed implementation algorithm.

Finally, our research complements a growing line of work on the integration of CP to robust decision-making (Kiyani et al. 2025, Patel et al. 2024, Cortes-Gomez et al. 2024, Yeh et al. 2024, Hullman et al. 2025, Cresswell et al. 2024, Zhao et al. 2025). Unlike prior work that prescribes decisions, we leverage CP techniques to audit and assess the risk of candidate decisions, supporting decision-making through evaluation rather than recommendation. Additionally, our work contributes to the growing literature on human-algorithm collaboration through the use of generative model recommendations (Grand-Clément and Pauphilet 2024, Ibrahim et al. 2021, Orfanoudaki et al. 2022). By producing a diverse set of plausible decisions, generative algorithms can act as advisory tools, thus promoting exploration and helping users develop more effective decision strategies (Ajay et al. 2022, Li and Zhu 2024).

Aligned with this perspective, our approach encourages decision-makers to evaluate alternatives beyond standard prescriptive recommendations, revealing risks and improvements that the nominal optimal decision alone may obscure. This emphasis on expanding the space of considered decisions is a central conceptual contribution of our work.

3. Problem Setup

We consider a general parametric decision-making model in which the decision is determined as the (near-)optimal solution to a constrained optimization problem under an observed scenario $y \in \mathcal{Y}$ (*e.g.*, demand, price, or system load). Formally, we define the ϵ-*optimal decision rule* as

$$\pi_\epsilon(y) \coloneqq \arg\min_{\epsilon,\, z\in\mathcal{Z}} f(z; y), \tag{1}$$

where f denotes the objective function (*e.g.*, cost, delay, or loss), and $\mathcal{Z}$ denotes the feasible region. We assume that $\mathcal{Z}$ is nonempty and compact, and that $f(\cdot; y)$ is continuous and convex in z for each y, ensuring the set $\pi_\epsilon(y)$ is non-empty for every $y \in \mathcal{Y}$. The operator $\arg\min_\epsilon$ generalizes the standard $\arg\min$ to allow a tolerance level $\epsilon \geq 0$, defined as:

$$\arg\min_\epsilon f(z) \;\coloneqq\; \left\{ z \;\middle|\; f(z) \leq \inf_{z'} f(z') + \epsilon \right\}.$$

Note that setting $\epsilon = 0$ recovers the *exact optimal decision rule*. Throughout the manuscript, we suppress the subscript ϵ when no ambiguity arises, writing $\pi(y)$ for $\pi_\epsilon(y)$.

In many operational contexts, the realized scenario y is unobserved at the time of decision-making. Let $z \in \mathcal{Z}$ denote a prescribed (implemented) decision—potentially produced by a human operator, heuristic policy, or black-box model—made without access to the true realization of $Y \sim \mathcal{P}_Y$. This setup captures a common class of practical decision environments where actions are based on implicit or subjective beliefs about the underlying uncertainty rather than a fully specified probabilistic model.

Our objective is to assess how well a prescribed decision z aligns with the (unknown) optimal response under the true realization of Y. Formally, this quantity is expressed as the probability that z belongs to the (near-)optimal decision set $\pi(Y)$, *i.e.*, $\mathbb{P}\{z \in \pi(Y)\}$, where the probability is taken with respect to $Y \sim \mathcal{P}_Y$. This captures how frequently the prescribed decision coincides with an (approximately) optimal choice across possible realizations. Since the distribution $\mathcal{P}_Y$ is typically unknown and only limited samples $\mathcal{D} = \{Y_i\}_{i=1}^n$ are available, this probability cannot be estimated precisely. We therefore focus on its data-driven lower bound $1 - \alpha(z)$ defined as

$$\mathbb{P}\big\{z \in \pi(Y)\big\} \geq 1 - \alpha(z). \tag{2}$$

The quantity $1 - \alpha(z)$ serves as a conservative optimality certificate, providing a guaranteed confidence level that the prescribed decision z is (near-)optimal under the true, unknown distribution $\mathcal{P}_Y$.

In the general setting, the quantities in (2) may depend on additional contextual covariates $X \in \mathcal{X}$, such as spatial, temporal, or environmental factors that influence the realization of Y. Accordingly, both the decision set $\pi(Y)$ and the confidence function $\alpha(z)$ can be defined conditionally on X, reflecting the dependence of the outcome distribution $\mathcal{P}_{Y|X}$ on its context. Also, the available data take the form $\mathcal{D} = \{(Y_i, X_i)\}_{i=1}^n$ in this case, capturing paired observations of outcomes and their associated covariates. For clarity of exposition, in the remainder of the paper, we omit the explicit dependence on X and focus on the marginal representation.

3.1. Motivating Examples

We provide two illustrative examples to demonstrate how the proposed problem setup applies to different decision-making contexts, ranging from simple binary decisions to complex resource allocation problems.

Example 1 (Stylized Umbrella-Carrying Decision). A human decision-maker decides whether to carry an umbrella without checking the weather. Let $z \in \{0, 1\}$ denote the decision ($z = 1$: carry; $z = 0$: not carry) and $y \in \{0, 1\}$ denote the realized weather ($y = 1$: rain; $y = 0$: no rain). The feasible set is $\mathcal{Z} = \{0, 1\}$, and the objective function is a hinge-type misclassification cost:

$$f(z; y) = \begin{cases} 1, & \text{if } z \neq y \text{ (mismatch between decision and weather),} \\ 0, & \text{otherwise,} \end{cases} \quad \text{or simply } f(z; y) = \mathbb{1}\{z \neq y\}.$$

In this example, the decision-maker acts based on an implicit perception of local weather patterns Y. The chosen decision z thus reflects the decision-maker's subjective belief about the likelihood of rain—carrying an umbrella if rain is perceived as more probable, and abstaining otherwise. Our goal is to assess whether this private knowledge aligns with the true local weather distribution by evaluating, through empirical data, how frequently the prescribed decision coincides with the optimal decision under the observed realizations of Y.

Example 2 (Clinical Triage). A clinician must allocate limited treatment resources under time pressure without reviewing complete patient records. Let there be d patients, and define the decision $\mathbf{z} = (z_1, \ldots, z_d) \in \{0, 1\}^d$, where $z_j = 1$ if patient j receives treatment, subject to a capacity constraint $\sum_{j=1}^{d} w_j z_j \leq B$ (*e.g.*, limited ICU beds, clinician time, or medication supply), with $w_j > 0$ denoting resource use of patient j and $B > 0$ the available budget. Let $\mathbf{x} = (x_1, \ldots, x_d)$ represent observable clinical covariates, and $\mathbf{y} = (y_1, \ldots, y_d)$ with $y_j \geq 0$ the realized benefit of treating patient j given x_j. The feasible region and objective function are defined as

$$\mathcal{Z} = \left\{ \mathbf{z} \in \{0, 1\}^d : \sum_{j=1}^{d} w_j z_j \leq B \right\}, \quad f(\mathbf{z}; \mathbf{y}) = -\sum_{j=1}^{d} y_j z_j.$$

In this more realistic setting, the clinician makes treatment decisions based on incomplete information or heuristic prioritization, such as triage scores or observable symptoms (Parenti et al. 2014). The prescribed decision vector $\mathbf{z}$ encodes the clinician's implicit belief about which patients are most likely to benefit given the resource constraint. We aim to evaluate whether such implicit prioritization aligns with the optimal clinical decisions suggested by data, by comparing the prescribed allocation with the optimal allocation derived from empirical patient records.

4. The CREDO Framework

We now present our method, which adapts the distribution-free uncertainty quantification framework of CP (Vovk et al. 2005) to estimate decision risk through calibrated inverse feasible regions.

4.1. Preliminary: Conformal Prediction

CP is a model-agnostic framework for uncertainty quantification that delivers finite-sample valid prediction regions under the exchangeability assumption:

Assumption 1 **(Exchangeability)**. *The calibration data* $\{(X_i, Y_i)\}_{i=1}^n$ *and the test point* (X, Y) *are exchangeable. That is, the joint distribution of* $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ *is invariant under any permutation of these* $(n+1)$ *pairs.*

Let the calibration dataset be $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, drawn from an arbitrary joint distribution over $\mathcal{X} \times \mathcal{Y}$. For a user-specified miscoverage rate α, the objective is to construct a prediction set $C(\cdot; \alpha)$ such that

$$\mathbb{P}\{Y \in C(X; \alpha)\} \geq 1 - \alpha, \tag{3}$$

where the probability is taken jointly over the randomness in the test point (X, Y) and the calibration data $\mathcal{D}$. This requirement is referred to as *marginal validity*.

Split CP (Papadopoulos et al. 2002) provides a practical and computationally efficient method for achieving the guarantee in (3). The procedure begins by training a prediction model $g : \mathcal{X} \to \mathcal{Y}$ using data independent of the calibration set. We then specify a nonconformity score function $l : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which quantifies how atypical a candidate label is relative to the model's prediction; for instance, one may take $l(x, y) = \|y - g(x)\|_2$. Given the calibration dataset, we compute the corresponding nonconformity scores

$$L_i = l(X_i, Y_i), \quad \forall i = 1, \ldots, n.$$

These scores determine how large a deviation from the model is acceptable in order to guarantee the desired coverage. For any $\alpha \in [1/(n+1), 1)$, we define the adjusted empirical quantile as

$$\hat{Q}(\alpha) = \inf\left\{ l \in \mathbb{R} : \hat{F}_n(l) \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\}, \quad \text{where} \quad \hat{F}_n(l) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{L_i \leq l\}.$$

This quantile defines the prediction region:

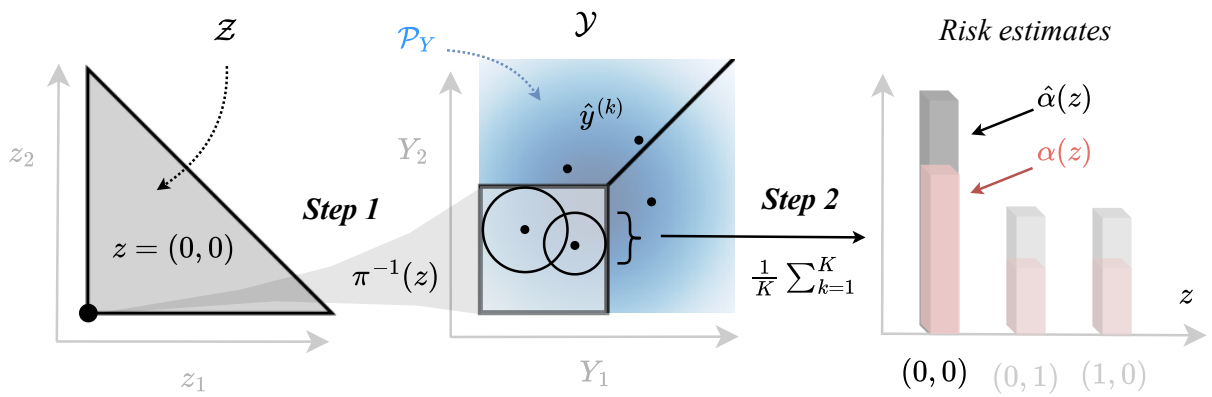
$$C(x; \alpha) = \left\{ y \in \mathcal{Y} : l(x, y) \leq \hat{Q}(\alpha) \right\}.$$

See Angelopoulos and Bates (2021) for a comprehensive survey of CP.

4.2. Conformalized Decision Risk Assessment

We now develop our framework for estimating the decision risk $\alpha(z)$ defined in Section 3. We accomplish this by reformulating the optimality probability in (2) as an *inverse* CP problem, where we seek to quantify how often uncertain parameters fall within regions that render a decision optimal. The key idea is to project the decision's optimality condition onto an *inverse feasible region* (Chan et al. 2025, Tavaslıoğlu et al. 2018) in the space of $\mathcal{Y}$, and to construct inner geometric approximations of this region using multiple generated conformal balls (Shafer and Vovk 2008, Wang et al. 2023). The resulting average miscoverage rate across balls then serves as a data-driven lower-bound of decision risk. We detail this procedure in two steps, with an example illustrated in Figure 2.

Figure 2 **Example of the CREDO procedure applied to a linear programming problem.**



Note. There are two main steps: (*i*) Map the given candidate decision z (*e.g.*, $(0,0)$) to its inverse feasible region $\pi^{-1}(z)$; (*ii*) Assess the risk by constructing inner approximation of $\pi^{-1}(z)$ using generative conformal prediction, and averaging their miscoverage levels.

4.2.1. Reformulation with Inverse Feasible Region Our first step reformulates (2) as an inverse optimization problem. For any given realization y, a decision z is (near-)optimal if and only if it attains the (near-)minimal objective value among all feasible decisions. Accordingly, we define *inverse feasible region* as the set of scenarios y under which z remains the (near-)optimal decision. Formally,

$$\pi_\epsilon^{-1}(z) := \left\{ y \in \mathcal{Y} \;\middle|\; f(z;y) \le \min_{z' \in \mathcal{Z}} f(z';y) + \epsilon \right\}. \tag{4}$$

This definition can be viewed as a generalization of its classical notion defined by Chan et al. (2025), providing a relaxed definition that accommodates near-optimal decisions. For notation simplicity, we use π^{-1} to denote π_ϵ^{-1} when clear from context. The reformulation is formally stated as follows.

Lemma 1 **(Reformulation)**. *The probability in* (2) *can be equivalently expressed as:*

$$\mathbb{P}\{z \in \pi(Y)\} \equiv \mathbb{P}\left\{Y \in \pi^{-1}(z)\right\}. \tag{5}$$

Its proof follows immediately by the definition of $\pi(Y)$ and $\pi^{-1}(z)$. This equivalence has important computational implications: the original formulation in (2) requires computing the probability that a fixed decision z lies in the random (near-)optimal set $\pi(Y)$. This task remains computationally intractable because it entails solving the underlying optimization problem for every possible realization of Y. By contrast, the reformulation in Lemma 1 reverses this viewpoint by evaluating the probability that the *random* variable Y lies within the *deterministic* region $\pi^{-1}(z)$. This shift yields a tractable probability estimation problem.

4.2.2. Risk Estimation via Generative Conformal Prediction The second step estimates the probability $\mathbb{P}\{Y \in \pi^{-1}(z)\}$. Direct estimation is generally intractable, as it requires computing the probability mass over $\pi^{-1}(z)$—a region with complex and implicitly defined geometry. Such estimation typically demands either strong parametric assumptions on the distribution of Y or efficient sampling access to it, neither of which is

often feasible in practice. A common workaround is to train a conditional generative model that samples from $\mathcal{P}_{Y|X}$ and then approximate the desired probability by the fraction of these samples falling within $\pi^{-1}(z)$. However, this approach lacks any finite-sample guarantee for the validity requirement in (2).

To address this, we develop a generative conformal approach that constructs statistically valid inner approximations of $\pi^{-1}(z)$ using calibration data. Given an input X, we generate a collection of CP balls $C(X;\alpha)$ such that (i) each region lies entirely within $\pi^{-1}(z)$ and (ii) its coverage probability is lower-bounded with statistical guarantees obtained from calibration. This yields the following bound on decision risk:

$$\mathbb{P}\left\{Y \in \pi^{-1}(z)\right\} \overset{(i)}{\geq} \mathbb{P}\left\{Y \in C(X;\alpha)\right\} \overset{(ii)}{\geq} 1-\alpha.$$

The right-hand guarantee follows from the finite-sample validity of CP. To satisfy the left-hand inequality, we identify the smallest α for which the corresponding ball $C(X;\alpha)$ remains entirely within $\pi^{-1}(z)$. The resulting average value of α across generated balls provides a data-driven estimator of the decision risk, denoted by $\hat{\alpha}(z)$.

4.2.3. Proposed Algorithm We begin with training a (conditional) generative model $g : \mathcal{X} \to \mathcal{Y}$ on a training dataset to approximate the conditional distribution $\mathcal{P}_{Y|X}$. Then, for a test input x, we draw a prediction $\hat{y} \sim g(x)$ and construct a prediction set as an ℓ_2 ball centered at $\hat{y}$:

$$C(x;\alpha) = \{y \in \mathcal{Y} \mid \|y - \hat{y}\|_2 < R(\alpha)\}, \tag{6}$$

where the radius $R(\alpha)$ is obtained by calibrating the nonconformity scores $\{L_i\}_{i=1}^n$ on $\mathcal{D}$. Of note, $R(\cdot)$ must be a decreasing function, meaning that a smaller α corresponds to a larger $R(\alpha)$ and higher empirical coverage. We consider three specifications of the radius function $R(\alpha) : [0,1] \to \mathbb{R}_+ \cup \{\infty\}$, each by default satisfies the boundary conditions $R(\alpha) = \infty$ for $\alpha \in [0, 1/(n+1))$ and $R(1) = 0$. For $\alpha \in [1/(n+1), 1)$, define:

$$R_p(\alpha) = \hat{Q}(\alpha); \qquad (p\text{-value radius}) \tag{7}$$

$$R_e(\alpha) = \frac{\sum_{i=1}^n L_i}{\alpha(n+1) - 1}; \qquad (e\text{-value radius}) \tag{8}$$

$$R_\infty(\alpha) = \infty. \qquad (\text{Monte Carlo radius}) \tag{9}$$

We note that R_p corresponds to the classical conformal radius based on the empirical quantiles of calibration residuals (Vovk et al. 2005, Singh et al. 2024); R_e is an e-value–based variant (Grünwald et al. 2024), offering stronger post-hoc validity guarantees (Vovk 2025, Balinsky and Balinsky 2024, Gauthier et al. 2025b); and R_∞ is used in a Monte Carlo-based estimation. We will discuss their respective properties in Section 5 and Section 7, showing that R_e provides the strongest robustness, R_∞ the highest accuracy, and R_p a balanced trade-off between the two.

For all variants, the coverage level of each generated prediction set is determined by solving

$$\tilde{\alpha}(z) = \min_{\alpha \in [0,1]} \alpha \quad \text{s.t.} \quad C(x;\alpha) \subseteq \pi^{-1}(z). \tag{10}$$

Algorithm 1 Conformalized Decision Risk Assessment (CREDO)

Require: Fitted generative model g; Calibration dataset $\{(x_i, y_i)\}_{i=1}^{n}$; Sample size K; Decision z; Test covariate x;

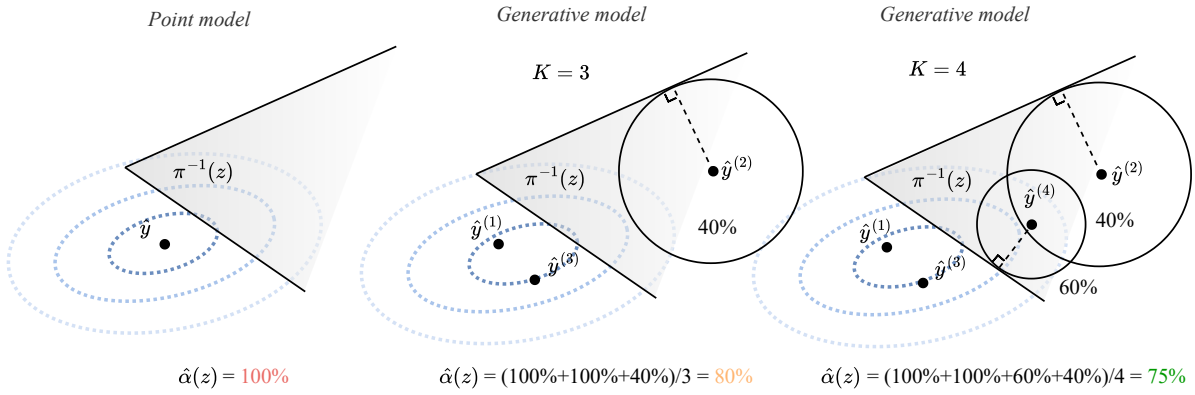
Initialize nonconformity score set $\mathcal{L} \leftarrow \emptyset$;
for $i \in \{1, \ldots, n\}$ **do**
  $\hat{y}_i \sim g(x_i)$; $L_i \leftarrow \|y_i - \hat{y}_i\|_2$; $\mathcal{L} \leftarrow \mathcal{L} \cup \{L_i\}$;
end for
for $k = 1, \ldots, K$ **do**
  $\hat{y}^{(k)} \sim g(x)$ generate prediction using x;
  $C^{(k)}(x; \alpha) \leftarrow$ construct conformal set given $\mathcal{L}$ and $\hat{y}^{(k)}$ via (6);
  $\tilde{\alpha}^{(k)}(z) \leftarrow$ solve for the k-th decision risk via (10);
end for
$\hat{\alpha}(z) \leftarrow 1/K \cdot \sum_{k=1}^{K} \tilde{\alpha}^{(k)}(z)$;
return Risk estimator $\hat{\alpha}(z)$.

The entire procedure above is repeated across K sets to obtain the collection of estimates $\{\tilde{\alpha}^{(k)}(z)\}_{k=1}^{K}$, which are then averaged to yield the final risk estimator:

$$\hat{\alpha}(z) = \left(\tilde{\alpha}^{(1)}(z) + \ldots + \tilde{\alpha}^{(K)}(z)\right)/K. \tag{11}$$

We highlight the necessity of using generative models in estimating (11) by contrasting them with traditional point prediction models commonly used in standard CP. As illustrated in Figure 3, point prediction models may produce outputs that lie near the boundary of, or even outside, the inverse feasible region $\pi^{-1}(z)$, leading to overly conservative risk estimates (*i.e.*, extremely small or even zero values when predictions fall outside

Figure 3 **Comparison of risk estimation using a point model and generative models ($K = 3$ and $K = 4$) in CREDO.**



Note. The gray shaded area represents the inverse feasible region $\pi^{-1}(z)$. The blue dotted ellipsoids represent the conditional distribution of $Y \mid X$, and the black dots indicate model predictions $\hat{y}^{(k)}$. The balls indicate calibrated prediction sets $C^{(k)}(x; \hat{\alpha}(z))$.

the region). In contrast, generative models enable multiple stochastic draws; increasing K improves the likelihood that at least one sample lies within $\pi^{-1}(z)$, yielding more accurate and less conservative risk estimates. This insight will be formalized through the notion of the true positive rate in our theoretical analysis (Section 5). The complete algorithmic procedure is summarized in Algorithm 1.

5. Theoretical Analysis

In this section, we now establish the statistical properties of our risk estimator $\hat{\alpha}(z)$, providing three main theoretical results. First, we show that our estimator is *conservative* (Theorem 1 and Corollary 1), ensuring it provides a valid upper bound on the true decision risk. Second, we prove that the estimator is asymptotically consistent (Proposition 1), admitting a tight characterization for the decision risk in large samples. Finally, we highlight the pivotal role of the generative model in our framework by analyzing the estimator's true positive rate (Proposition 2), which directly influences the quality of downstream decisions. All proofs are deferred to Appendix A.

5.1. Validity

A fundamental requirement for any risk assessment tool is that it provides reliable guarantees. We therefore first establish that CREDO's risk estimates are conservative for both the e- and p-value radii, ensuring its trustworthiness to the practitioners.

THEOREM 1 **(Conservativeness under e-value)**. *Under Assumption 1 and $R(\alpha) = R_e(\alpha)$, the estimator $\hat{\alpha}(z)$ defined in* (11) *satisfies*

$$\mathbb{P}\{z \in \pi(Y)\} \geq 1 - \mathbb{E}[\hat{\alpha}(z)].$$

Thus, $\hat{\alpha}(z)$ provides an expectation-wise upper bound on the true decision risk, without requiring correctness of the generative model. This highlights the robustness of the e-value construction. On the other hand, the p-value radius has the following guarantee.

COROLLARY 1 **(Conservativeness under p-value)**. *Under Assumption 1 and $R(\alpha) = R_p(\alpha)$, the estimator $\hat{\alpha}(z)$ defined in* (11) *satisfies*

$$\mathbb{P}\{z \in \pi(Y)\} \geq 1 - \mathbb{E}[\hat{\alpha}(z)] - \frac{1}{n+1} \sum_{i=1}^{n} d_{\mathrm{TV}}^{(i)}(\hat{\alpha}(z)),$$

where $d_{\mathrm{TV}}^{(i)}(\hat{\alpha}(z))$ is the conditional total variation (TV) distance between the data vector and the same vector with entries i and $(n+1)$ swapped.

Here, the additional TV term reflects the deviation from full exchangeability caused by post-hoc selection of $\hat{\alpha}(z)$ (Barber et al. 2023, Gauthier et al. 2025a), which disrupts the conservativeness of p-value radius's risk estimate but was circumvented by e-value radius. However, we show in our experiments (Section 7.1) that conservativeness for p-value radius does hold in practice, and its accuracy is also superior compared to e-value radius. This implies that the additive term in Corollary 1 is usually small in practice, making p-value radius a more effective choice when a theoretical guarantee is not the top priority.

5.2. Consistency

We next study whether $\hat{\alpha}(z)$ converges to the true risk under mild assumptions on the generative model, measured through total variation distance $d_{\mathrm{TV}}(\cdot,\cdot)$. The result is summarized as follows.

Proposition 1 **(Asymptotic conditional consistency)**. *Let $\hat{\mathcal{P}}_{Y|X}$ be the conditional distribution learned by the generative model g. If for some $\delta \geq 0$,*

$$d_{\mathrm{TV}}\left(\hat{\mathcal{P}}_{Y|X=x}, \mathcal{P}_{Y|X=x}\right) \leq \delta \quad \textit{for all } x \in \mathcal{X},$$

then with $R(\alpha) = R_\infty(\alpha)$, the estimator (11) *satisfies*

$$|\hat{\alpha}(z) - \mathbb{P}\{z \notin \pi(Y) \mid X = x\}| \leq O_p(K^{-1/2}) + \delta.$$

Proposition 1 guarantees marginal consistency as $K \to \infty$, provided that the generative model approximates the conditional distribution within TV distance δ. Thus, with a well-trained generative model and sufficiently many generated samples, $\hat{\alpha}(z)$ converges at a parametric rate to the true decision risk, complementing the finite-sample conservativeness of Theorem 1.

5.3. True Positive Rate

Finally, we analyze how often the estimator correctly retains feasible decisions. Define the true positive rate as

$$\mathrm{TPR}(K) := \frac{\mathbb{E}\left[\#\{z \in \mathcal{Z} : \alpha(z) < 1 \text{ and } \hat{\alpha}(z) < 1\}\right]}{\#\{z \in \mathcal{Z} : \alpha(z) < 1\}}, \tag{12}$$

where $\alpha(z) = \mathbb{P}\{z \notin \pi(Y)\}$ is the true risk. TPR measures the fraction of genuinely feasible decisions that are not mistakenly discarded by the estimator. From a practical perspective, TPR quantifies the ability of CREDO to correctly identify decisions that have non-trivial optimality probability while maintaining conservative guarantees. A low TPR corresponds to more "false-positive" exclusions, which may eliminate profitable or optimal decisions; thus, a higher TPR indicates better decision quality. We have the following result.

Proposition 2 **(True positive rate)**. *The TPR(K) increases monotonically with K.*

Proposition 2 highlights the importance of using generative models and of sampling diversity: small K (*e.g.*, $K = 1$) or deterministic predictors increase the likelihood of false exclusions, lowering TPR. Increasing K mitigates this issue and improves downstream decision quality. The result holds uniformly across all calibrated radii $R(\alpha)$.

Together, these results demonstrate that CREDO provides conservative risk estimates (validity), converges to the true risk under reasonable conditions (consistency), and benefits from generative sampling to avoid excessive conservativeness (high TPR). These properties ensure both theoretical rigor and practical utility in decision support applications.

6. Computational Implementation

The theoretical analysis in the previous section establishes that CREDO provides conservative risk estimates with strong finite-sample guarantees. However, realizing these guarantees in practice requires efficiently solving (10) for each decision z and generated sample $\hat{y}$. This presents a fundamental computational challenge as we must verify whether a calibrated conformal ball $C(x;\alpha)$ is entirely contained within the inverse feasible region $\pi^{-1}(z)$, which is NP-hard in general.

This section develops practical algorithms to address this challenge. We begin by reformulating the set containment constraint into a standard optimization problem (Section 6.1). Subsequently, this allows us to derive efficient solution strategies for two important problem classes (Section 6.2): (*i*) linear programs (LPs), where we derive closed-form expressions, and (*ii*) general convex problems, where we employ gradient-based approximations. These computational insights enable practitioners to deploy CREDO for real-world decision support while maintaining the theoretical guarantees established in Section 5. The proofs in this section are deferred to Section A.

6.1. Reformulating the Set-Containment Constraint

We begin by noticing that the set containment condition $C(x;\alpha) \subseteq \pi^{-1}(z)$ is equivalent to requiring that the calibrated conformal ball does *not* intersect the complement of the inverse feasible region:

$$C(x;\alpha) \subseteq \pi^{-1}(z) \quad \Longleftrightarrow \quad C(x;\alpha) \cap \left(\pi^{-1}(z)\right)^c = \emptyset.$$

Since $C(x;\alpha)$ is an ℓ_2 ball centered at $\hat{y}$, the right hand side is further equivalent to

$$\underbrace{\text{for all } y \in \left(\pi^{-1}(z)\right)^c,\ \|y-\hat{y}\|_2 \geq R(\alpha)}_{(i)} \qquad \text{if } \underbrace{\hat{y} \in \pi^{-1}(z)}_{(ii)}.$$

Here, condition (*i*) requires that every scenario y violating the optimality of decision z must lie at least with distance $R(\alpha)$ away from $\hat{y}$; condition (*ii*) states that the former condition is enforced only when $\hat{y}$ lies within of $\pi^{-1}(z)$. Otherwise, then no radius can guarantee containment, and the α is trivially mapped to one by design. This reformulation allows us to circumvent the need for validating the original set containment relation defined in Equation (10), and derive a more principled computational representation, summarized as follows.

Proposition 3 **(Computation)**. *Let $\hat{y}$ be a generated prediction and $\tilde{\alpha}$ be its estimated risk. Then*

$$\tilde{\alpha}(z) = 1, \qquad \text{if } z \notin \pi_\epsilon(\hat{y}), \tag{13a}$$

$$\tilde{\alpha}(z) = \sup_{y \in \mathcal{Y}} R^{-1}\left(\|y-\hat{y}\|_2\right) \quad \text{s.t.} \quad f(z;y) > \min_{z' \in \mathcal{Z}} f(z';y) + \epsilon, \qquad \text{if } z \in \pi_\epsilon(\hat{y}), \tag{13b}$$

where we expand the notation π to π_ϵ to avoid ambiguity, and $R^{-1}(\cdot)$ is the inverse radius function:

$$R^{-1}(l) = \min_{\alpha \in [0,1]} \left\{\alpha : R(\alpha) \leq l\right\}.$$

Furthermore, the inverse radius functions for the three calibrated radii in Equations (7)–(9) of Proposition 3 admit closed-form expressions, which makes solving for (13b) a more explicit task.

Lemma 2. *For the radii defined in Equations* (7)–(9)*, the corresponding inverse functions are:*

$$R_p^{-1}(l) = \left(1 - \frac{1}{n+1}\left\lfloor \sum_{i=1}^{n} \mathbb{1}\{L_i \le l\} \right\rfloor\right)^+, \tag{14}$$

$$R_e^{-1}(l) = \left(\frac{\sum_{i=1}^{n} L_i + l}{(n+1)l}\right)^+, \tag{15}$$

$$R_\infty^{-1}(l) = 0, \tag{16}$$

where $(\cdot)^+ = \max\{0, \cdot\}$ *denotes the rectified linear unit (ReLU) operator.*

Additionally, we note that these results also shed light on the namesake of the Monte-Carlo radius R_∞, which is summarized in the following remark.

Remark 1. Combining Equation (16) with Proposition 3, one can verify that the risk estimate computed with the Monte Carlo radius degenerates to the form:

$$\hat{\alpha}(z) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{z \notin \pi(\hat{y}^{(k)})\}.$$

Namely, the risk estimator is essentially a Monte Carlo estimator of whether the generated samples from the generative model $g(x)$ fall inside $\pi(\hat{y}^{(k)})$.

6.2. Computational Strategies for Specific Problem Classes

Leveraging Proposition 3, the computation of each individual risk estimate $\tilde{\alpha}(z)$ can be executed as two steps: (*i*) Given a generated sample $\hat{y}$, we solve the original decision-making problem (1) once to check if $z \in \pi(\hat{y})$. If not, we immediately set $\tilde{\alpha}(z)$ conservatively to one. Otherwise, we compute $\tilde{\alpha}(z)$ via solving (13b), which admits different strategies depending on the form of the objective $f(z; y)$ and the geometry of $\mathcal{Z}$. In what follows, we analyze by specific problem classes, where we first provide a closed-form estimator for linear problems, and then introduce a heuristic approach applicable to general convex decision models.

6.2.1. Linear Problems We first examine the special case in which the decision problem is a linear program (LP) of the form

$$f(z; y) \coloneqq \langle y, z \rangle \quad \text{s.t.} \quad Az \le b, \tag{17}$$

where $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ define a nonempty polyhedral feasible region. We use $\mathcal{V}$ to denote the set of extreme points of this polytope, which can be solved from A and b using procedures such as the double description method (Motzkin et al. 1953, Fukuda and Prodon 1995).

In this setting, we can derive a closed-form solution for the optimization problem (13b) by leveraging $\mathcal{V}$ and the property that the optimal solution to an LP always occurs at an extreme point (Bertsimas and

Tsitsiklis 1997). Specifically, by the fact that optimal solutions for LPs are always incurred at extreme points (Krein and Milman 1940), the constraint in (13b) can be equivalently expressed as

$$f(z;y) > \min_{z' \in \mathcal{Z}} f(z';y) + \epsilon \quad \Longleftrightarrow \quad \text{For all } v \in \mathcal{V} \setminus \{z\}, \quad \langle y, v - z \rangle - \epsilon < 0,$$

Consequently, its feasible region is the union of a finite number of halfspaces, each associated with a direction vector $z - v$ and offset $-\epsilon$. Therefore, the nested optimization problem (13b) reduces to evaluating a finite set of closed-form expressions, yielding an explicit formula for the risk estimate.

Corollary 2 **(Closed-Form Risk Estimation for Linear Programs)**. *Under* (17)*, the risk estimate* $\tilde{\alpha}(z)$ *for a single generated sample* $\hat{y}$ *admits the closed-form expression for* (13b)*:*

$$\tilde{\alpha}(z) = \max_{v \in \mathcal{V} \setminus \{z\}} R^{-1}\left(\frac{\left|\langle \hat{y}, z - v \rangle - \epsilon\right|}{\|z - v\|_2}\right), \quad \text{if } z \in \pi_\epsilon(\hat{y}),$$

where we also expand the notation π *to* π_ϵ *to avoid ambiguity.*

6.2.2. Convex Problems In the linear setting, the success in deriving a closed-form estimate critically relies on the polyhedral description of the ϵ-suboptimality region. However, for general convex decision problems, where we assume $\mathcal{Z}$ is convex and $f(z;y)$ is convex in each argument marginally, such a representation is unavailable. Additionally, this also renders the constraint $f(z;y) - \min_{z' \in \mathcal{Z}} f(z';y) > \epsilon$ to be potentially nonconvex in (z, y), and cannot be directly handled by off-the-shelf convex solvers.

We adopt a heuristic algorithm (Algorithm 2) that combines ideas from the difference-of-convex algorithm (DCA) (Tao and An 1997) and coordinate descent (Wright 2015) to solve for Problem (13b). By the monotonically decreasing property of $R^{-1}(\cdot)$, we begin by noticing that the original optimization problem can be reformulated as first solving the following single-level minimization problem

$$\min_{y, z'} \|y - \hat{y}\|_2 \quad \text{s.t.} \quad f(z;y) - f(z';y) > \epsilon \ \text{ and } \ z' \in \mathcal{Z}. \tag{18}$$

and then applying the inverse radius function $R^{-1}(\cdot)$ to the optimal objective value. To solve (18), which jointly minimizes over (y, z'), we adopt an alternating scheme that iteratively fixes one variable and solves the resulting marginal convex subproblem in the other. The algorithm starts by initialize y at the generated prediction $y^{(0)} = \hat{y}$. For each iteration $t = 1, \ldots, T$, the procedure: (i) Solve for $z^{(t)}$, which is computed as the optimal solution to the original decision-making problem with $y = y^{(t-1)}$. It represents the most feasible solution $z \in \mathcal{Z}$ at the current iteration, as one can verify that it maximizes the slack of the constraint in Problem (18). (ii) Solve for $y^{(t)}$, where we adopt a convex surrogate to the original constraint by approximating $f(z;y)$ with its first-order Taylor expansion, which can then be solved with standard solvers. After T iterations, the final iterate $y^{(T)}$ is used to form the approximate risk estimate

$$\tilde{\alpha}(z) \coloneqq R^{-1}\left(\|y^{(T)} - \hat{y}\|_2\right), \quad \text{if } z \in \pi(\hat{y}).$$

In Section 7.3, we empirically validate this adopted approach by showing it is reliable for producing high-quality risk estimates, providing a computationally tractable solution to decision risk assessment tasks in broader convex settings.

Algorithm 2 Alternating Optimization for Solving (13b) in General Convex Settings

Require: Prediction $\hat{y}$; Decision z; Tolerance level ϵ.

$y^{(0)} \leftarrow \hat{y}$;

for $t \in \{1, \ldots, T\}$ **do**

  $z^{(t)} \leftarrow \arg\min_{z' \in \mathcal{Z}} f(z'; y^{(t-1)})$;

  $\tilde{f}^{(t)}(z; y) \leftarrow f(z; y^{(t-1)}) + \left\langle \nabla_y f(z; y)\big|_{y=y^{(t-1)}} \, , \, y - y^{(t-1)} \right\rangle$;

  $y^{(t)} \leftarrow \arg\min_{y \in \mathcal{Y}} \|y - \hat{y}\|_2 \quad \text{s.t.} \quad \tilde{f}^{(t)}(z; y) - f(z^{(t)}; y) > \epsilon$;

end for

return $R^{-1}(\|y^{(T)} - \hat{y}\|_2)$.

7. Experiments

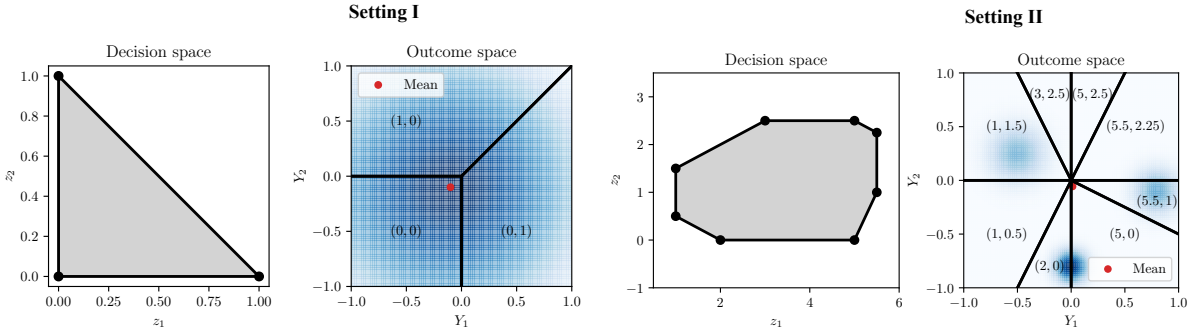
In this section, we evaluate the performance of CREDO through extensive numerical experiments. Our results demonstrate that: (*i*) High-quality risk estimation (Section 7.1): CREDO produces risk estimates that can be tuned toward either conservativeness or accuracy through the choice of radius, with both aspects outperforming baseline models. This flexibility underscores its suitability for decision risk-assessment tasks. (*ii*) Risk-aware decision prescriptions (Section 7.2): The risk estimates produced by CREDO can be directly used to prescribe decisions that achieve consistent low risk, demonstrating CREDO's reliability as a practical tool for risk-aware decision making. (*iii*) Effective modular design (Section 7.3): Each component of CREDO exhibits superior performance on its respective subtask compared with alternative ablation variants, highlighting the advantages of our modeling choices. Additional details of the experiments are provided in Section B.

We test CREDO across a diverse set of optimization problems, including linear programming (LP), quadratic programming (QP), second-order conic programming (SOCP), integer programming (IP), and a penalized knapsack problem that uses semi-synthetic data from a real-world grid infrastructure investment problem (Real Data). Specifically, we consider two stylized setups for the LPs, referred to as Setting I and Setting II, as illustrated in Figure 4. Setting I features a triangular feasible region with three vertices and a highly stochastic distribution of Y. Setting II introduces a more complex octagonal feasible region with five vertices and a more dispersed, multimodal distribution of Y. These two settings serve as running examples throughout our experiments due to their clarity and simplicity. The variance of the synthetic data distribution is controlled by a hyperparameter σ. Tabular results report the mean ± standard deviation across all trials. Boldface indicates the best-performing method, and underlining indicates the second-best within each setting.

7.1. Risk Estimation Evaluation

In this subsection, we evaluate CREDO's ability to produce risk estimates that flexibly trade off conservativeness and accuracy depending on the choice of radius, demonstrating that it outperforms baseline models along both dimensions. To this end, we benchmark CREDO's risk estimation performance against

Figure 4 Illustration of settings I and II.



Note. The gray region represents the feasible region in the decision space, and the cones to the right are the corresponding inverse feasible regions in the outcome space. The blue shade denotes the density mass of Y.

generic probability estimation baselines. Note that the quantity of interest, $\mathbb{P}\{z \in \pi(Y)\}$, can be interpreted as predicting, for a given test instance (x, y), the probability that a prescribed decision remains ϵ-optimal under y. This perspective naturally allows us to adapt standard regression and classification models as baselines for the decision risk-assessment task.

Specifically, we consider five representative baseline models: (*i*) SA: sample average of indicator functions computed from historic data; (*ii*) LR: logistic regression; (*iii*) NN: a neural network-based (multi-layer perceptron) classifier; (*iv*) QE: quantile estimator; (*v*) CP (Papadopoulos et al. 2002, Vovk et al. 2005): standard conformal prediction model using the upper prediction interval as its output. These baseline models capture the types of prediction approaches that decision-makers are likely to adopt as main candidates for estimating decision risk. We adopt two evaluation metrics for the decision risk estimate: (*i*) Validity: the ratio of trials where the estimated risk successfully upper bounds the true risk; (*ii*) MAE: the mean absolute error of the risk estimate compared to the true risk. These two metrics quantify conservativeness and accuracy, which are the two crucial dimensions of risk estimate evaluation. Additional details of the experimental settings are included in Section B.

Table 1 summarizes the results. For the first three baselines (SA, LR, NN), we observe low estimation error but poor validity. This is expected: these models are trained to maximize predictive accuracy, not conservativeness, and thus systematically underestimate risk. In contrast, the last two baselines (QE and CP) exhibit higher validity but substantially larger errors. Although both approaches can enforce conservativeness by adjusting the quantile level q, the choice of q is continuous in $[0, 1]$ and lacks any principled mechanism to guarantee validity across problems or distributional shifts. By comparison, CREDO achieves the best of both worlds. With p-value and e-value radii, CREDO attains 100% validity across all settings; with the Monte Carlo radius, it yields highly accurate risk estimates with (near) minimal error. Together, these results highlight the versatility of CREDO: it can be tuned to deliver either guaranteed conservativeness or high-precision estimation, offering a principled and robust framework for decision risk assessment.

Table 1 Evaluated metrics for different risk estimation methods across different optimization settings.

	LP Setting I		LP Setting II		QP ($\epsilon = 0.1$)		SOCP ($\epsilon = 0.2$)		IP ($\epsilon = 0.3$)	
	Validity (↑)	MAE (↓)	Validity (↑)	MAE (↓)	Validity (↑)	MAE (↓)	Validity (↑)	MAE (↓)	Validity (↑)	MAE (↓)
SA	0.53 ± 0.50	$\mathbf{0.04 \pm 0.03}$	0.56 ± 0.44	$\mathbf{0.03 \pm 0.02}$	0.43 ± 0.48	$\mathbf{0.04 \pm 0.03}$	0.47 ± 0.48	$\mathbf{0.03 \pm 0.03}$	0.41 ± 0.46	$\mathbf{0.04 \pm 0.03}$
LR	0.50 ± 0.48	0.06 ± 0.04	0.59 ± 0.39	0.03 ± 0.02	0.47 ± 0.50	0.05 ± 0.04	0.50 ± 0.45	0.06 ± 0.05	0.45 ± 0.46	0.07 ± 0.04
NN	0.50 ± 0.45	0.10 ± 0.09	0.39 ± 0.38	0.05 ± 0.03	0.63 ± 0.46	0.08 ± 0.05	0.40 ± 0.48	0.08 ± 0.06	0.48 ± 0.49	0.09 ± 0.06
QE	0.00 ± 0.00	0.62 ± 0.13	0.89 ± 0.16	0.15 ± 0.06	0.03 ± 0.10	0.60 ± 0.10	0.00 ± 0.00	0.56 ± 0.00	0.14 ± 0.04	0.53 ± 0.06
CP	$\mathbf{1.00 \pm 0.00}$	0.31 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.12 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.37 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.43 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.31 ± 0.00
CREDO (p)	$\mathbf{1.00 \pm 0.00}$	0.27 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.11 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.16 ± 0.03	$\mathbf{1.00 \pm 0.00}$	0.38 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.25 ± 0.02
CREDO (e)	$\mathbf{1.00 \pm 0.00}$	0.31 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.12 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.18 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.44 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.31 ± 0.00
CREDO (∞)	0.50 ± 0.50	$\underline{0.05 \pm 0.03}$	0.53 ± 0.43	$\underline{0.03 \pm 0.02}$	0.60 ± 0.48	$\underline{0.05 \pm 0.03}$	0.53 ± 0.49	$\underline{0.05 \pm 0.03}$	0.47 ± 0.47	$\underline{0.05 \pm 0.04}$

7.2. Decision Prescription Evaluation

In this part, we examine CREDO's performance when its outputted risk estimate is combined with the risk minimization criterion to prescribe decisions, where the candidate decision with the lowest risk is selected as the output. we show that CREDO can be used to select decisions with consistently high confidence, demonstrating its effectiveness in guiding practical decision-making.

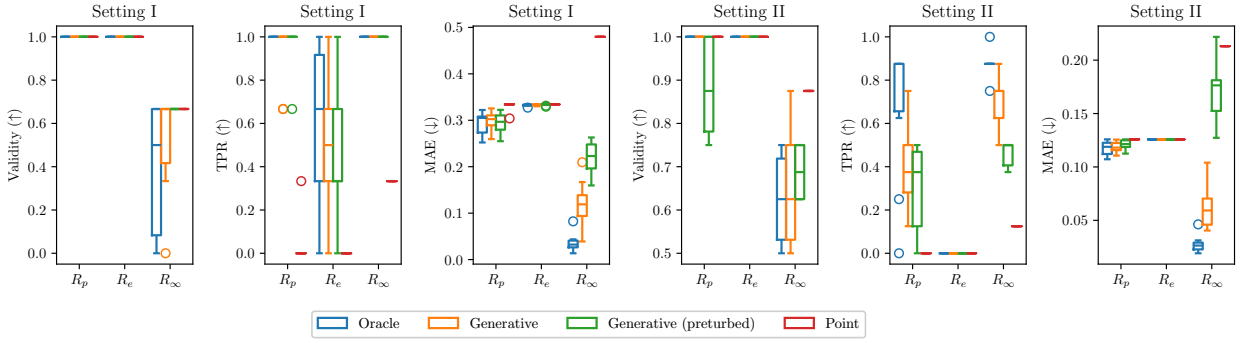
We benchmark CREDO against four baselines: predict-then-optimize (PTO) (Bertsimas and Kallus 2020), robust optimization (RO) (Bertsimas and Thiele 2006), smart predict-then-optimize (SPO+) (Elmachtoub and Grigas 2022), and decision-focused learning (DFL) (Amos and Kolter 2017). These methods are chosen for comparison as they represent widely used, risk-averse, data-driven approaches. For CREDO's generative model design, we adopt four different variants: the Gaussian Mixture model (GMM) with one, three, and five components, as well as the variational autoencoder (VAE) (Kingma and Welling 2013). We use *empirical confidence ranking* as our primary evaluation metric: given a decision policy π and a test dataset $\{(x_i, y_i)\}_{i=1}^m$, we apply π to each input x_i to generate predicted decisions $\{z_i\}_{i=1}^m$. We then compute the score $\sum_{i=1}^m \phi(z_i)$, where $\phi : \mathcal{Z} \to \mathbb{Z}_+$ is a mapping that maps each prediction to a discrete rank based on its frequency among the ground-truth optimal decisions $\{z_i^*\}_{i=1}^m$ in the test set. This metric is designed to capture a method's tendency to select decisions that are most likely to be optimal—a direct representation of risk in decision-making contexts.

Table 2 presents the comparison results. It can be seen that CREDO achieves the smallest ranking metric value across most datasets, on average selecting the top two most likely decisions across all datasets. Though it might seem concerning that the in Setting I ($\sigma = 0.1$), PTO, RO, and SPO+ all achieve better performance than CREDO, this is because when σ is small, the data becomes highly concentrated around the mean, rendering the problem nearly deterministic and can be best dealt with point-prediction baselines. Across different generative model designs in CREDO, the 1-GMM performs significantly worse than the other generative model specifications. Among the remaining models, VAE and 5-GMM generally achieve slightly better performance than 3-GMM across most trials. This trend supports the intuition that more expressive generative models within CREOD lead to improved risk estimation, which further improves its capability to guide risk-aware decision making, especially under highly uncertain environments.

Table 2 **Evaluated empirical confidence ranking (↓) for different methods across different optimization settings.**

Method	Setting I			Setting II			Real Data
	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 10$	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 10$	
PTO	$\mathbf{1.00 \pm 0.00}$	2.76 ± 0.59	2.24 ± 0.79	3.55 ± 0.50	3.36 ± 0.48	2.04 ± 1.65	1.75 ± 1.69
RO	$\mathbf{1.00 \pm 0.00}$	2.98 ± 0.14	3.00 ± 0.00	4.99 ± 0.10	6.00 ± 0.00	3.98 ± 0.80	3.00 ± 1.29
SPO+	$\mathbf{1.00 \pm 0.00}$	2.68 ± 0.65	2.02 ± 0.82	3.95 ± 1.20	4.67 ± 1.56	3.56 ± 1.50	2.67 ± 1.43
DFL	2.44 ± 0.64	1.83 ± 0.81	2.06 ± 0.79	3.60 ± 1.52	3.96 ± 2.07	3.66 ± 2.48	1.92 ± 1.04
CREDO (1-GMM)	1.94 ± 0.87	1.56 ± 0.54	1.49 ± 0.50	3.74 ± 0.98	3.94 ± 1.37	2.02 ± 1.41	1.92 ± 1.04
CREDO (3-GMM)	1.75 ± 0.77	$\mathbf{1.61 \pm 0.56}$	$\mathbf{1.48 \pm 0.52}$	1.05 ± 0.22	$\mathbf{1.00 \pm 0.00}$	2.03 ± 0.96	$\mathbf{1.75 \pm 0.92}$
CREDO (5-GMM)	1.89 ± 0.87	1.65 ± 0.62	1.54 ± 0.52	1.03 ± 0.17	$\mathbf{1.00 \pm 0.00}$	1.92 ± 0.89	1.92 ± 1.04
CREDO (VAE)	1.01 ± 0.10	1.61 ± 0.58	1.77 ± 0.71	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.06 \pm 0.24}$	1.92 ± 1.04

Figure 5 **Comparison of CREDO performance for three different calibrated radii.**

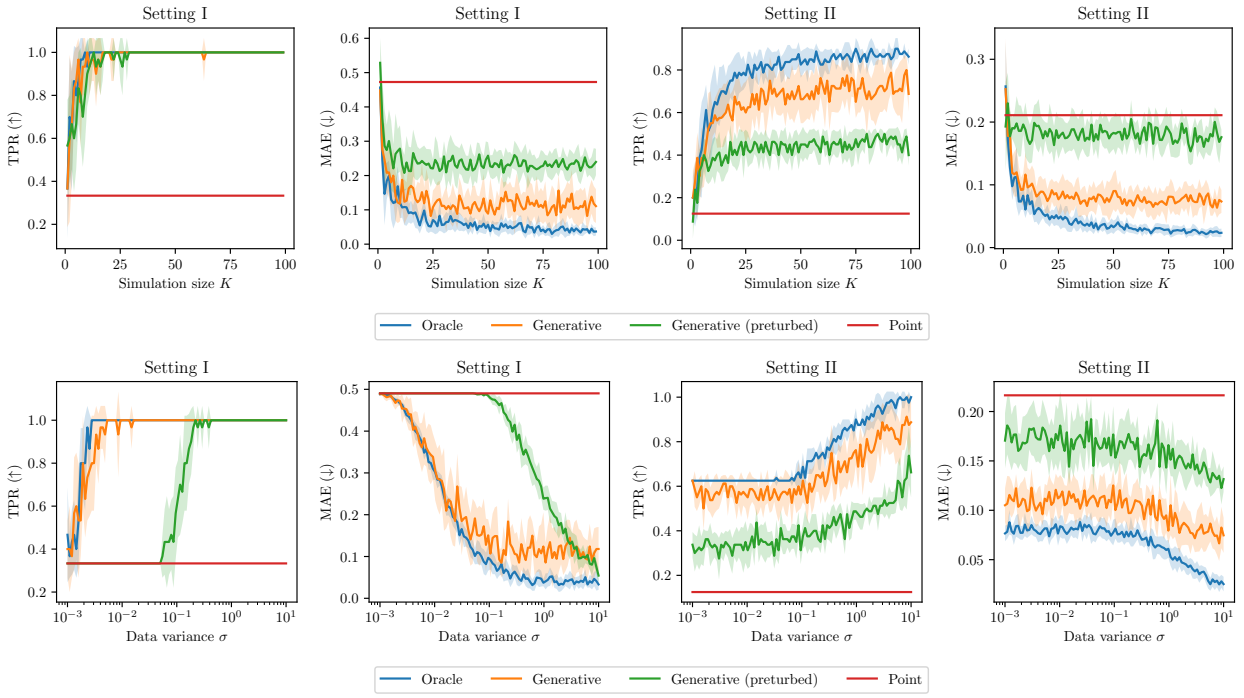


7.3. Ablation Studies

In this part, we examine ablation variants of CREDO through three controlled experiments that isolate the effects of radius design, generative modeling, and optimization procedures. These studies provide insight into CREDO's sensitivity while also demonstrating the effectiveness of our design choices.

Figure 5 presents the evaluation results comparing different types of calibrated radii (R_p, R_e and R_∞) under different settings and generative model designs. It can be seen that both R_p and R_e achieve a 100% conservativeness rate across nearly all settings, while R_∞ only achieves around 50%. However, in terms of TPR and MAE, R_p attains substantially better performance than both R_e and R_∞. This pattern was also observed in Section 7.1, and here we further dig deeper to analyze the underlying insights: (*i*) Although for R_p, a formal theoretical proof of conservativeness is not available, it empirically demonstrates comparably strong conservativeness as R_e. (*ii*) Even though R_∞ fails to ensure conservativeness, it achieves superior accuracy compared to both R_p and R_e by fully exploiting the generative model's fidelity. These insights highlight an essential trade-off when deploying CREDO: decision-makers must choose the calibrated radius according to their priorities. If conservativeness is paramount, R_p and R_e should be preferred; if accuracy is more critical, R_∞ is the better choice. Additionally, a consistent pattern can be observed from Figure 5: for both the TPR and accuracy metrics, R_p achieves substantially higher values than R_e across all settings. Since

Figure 6 Comparison of CREDO performance for four different generative model configurations.



Note. *Point* shows constant trend across different specifications of K and σ, as they do not affect the fitting of the point model.

their conservativeness rates are nearly identical, this indicates that R_p may serve as a more practical and effective choice than R_e, despite lacking a formal theoretical guarantee. Consequently, R_p can be interpreted as a balanced compromise between R_e and R_∞, suitable for scenarios where conservativeness and accuracy are of comparable importance.

For the second set of ablation studies, we implement CREDO with four different configurations of the underlying prediction (generative) model: (*i*) *Oracle*: using the ground-truth data distribution as the generative model. (*ii*) *Generative*: a three-component Gaussian Mixture model, fitted using the EM algorithm for 1×10^2 epochs. (*iii*) *Generative (perturbed)*: same as *Generative* except that the fitted mean of the components is perturbed with a small noise, which follows a $U(0.5, 1)$ distribution. (*iv*) *Point*: point prediction model that captures the mean of the marginal distribution of Y. The metrics used in this experiment follow those in the risk-estimation evaluation (Section 7.1), with additional inclusion of the true positive rate (TPR) in (12).

Figure 5 presents the evaluation results comparing different types of calibrated radii. We observe that R_p and R_e achieve a 100% conservativeness rate across nearly all settings, while R_∞ maintains only about 50%. However, in terms of TPR and accuracy, R_p attains substantially higher values than both R_e and R_∞. Two key insights emerge: (*i*) Although we lack a formal theoretical proof of conservativeness for R_p as established for R_e in Theorem 1, R_p empirically demonstrates comparably strong empirical performance on conservativeness. (*ii*) Despite its failure to ensure conservativeness, R_∞ achieves superior accuracy, as it behaves as an Monte Carlo estimator that fully exploits the generative model's fidelity. Overall, these findings

Table 3 Evaluated metrics for different optimization procedures solving (13b) across different settings.

	LP Setting I			LP Setting II			QP			SOCP			IP		
	Obj	Vio	Err	Obj	Vio	Err	Obj	Vio	Err	Obj	Vio	Err	Obj	Vio	Err
GD	0.44	0.40	0.45	0.06	0.07	0.05	0.60	0.23	0.50	1.53	0.13	1.31	0.22	**0.00**	0.19
BF	0.66	**0.00**	0.43	0.11	**0.00**	0.10	0.52	**0.00**	0.34	0.66	**0.00**	0.44	0.11	**0.00**	0.09
RS	0.89	**0.00**	0.66	0.12	**0.00**	0.11	0.70	**0.00**	0.52	0.91	**0.00**	0.69	0.12	**0.00**	0.10
RG	5.21	0.20	4.98	1.35	0.17	1.34	4.81	0.17	4.63	5.21	0.17	4.99	1.35	**0.00**	1.32
CREDO	**0.23**	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.10**	0.10	**0.08**	**0.14**	0.03	**0.08**	**0.00**	**0.00**	**0.02**

highlight an essential trade-off when deploying CREDO: decision-makers must choose the calibrated radius according to their priorities. If conservativeness is paramount, R_e or R_p should be preferred; if accuracy is more critical, R_∞ is the better choice. Additionally, a consistent pattern can be observed from Figure 5: for both the TPR and accuracy metrics, R_p achieves substantially higher values than R_e across all settings. Since their conservativeness rates are nearly identical, this indicates that R_p may serve as a more practical and effective choice than R_e, despite lacking a formal theoretical guarantee. Consequently, R_p can be interpreted as a balanced compromise between R_e and R_∞, suitable for scenarios where conservativeness and accuracy are of comparable importance.

Figure 6 presents the comparison results of CREDO risk estimation using different generative models. Across all settings, it can be observed that as both K and σ increase, the three generative-based methods (*Oracle*, *Generative*, and *Generative (perturbed)*) exhibit a clear improving trend, whereas the *Point* model remains constant. This pattern indicates that incorporating generative models within CREDO enhances the accuracy of risk estimation, particularly when the underlying data exhibit strong stochasticity (*i.e.*, larger variance). Moreover, the trends in Figure 6 generally follow the order *Oracle* > *Generative* > *Generative (perturbed)* > *Point*. Since both *Generative (perturbed)* and *Point* can be regarded as instances of misspecified models, this observation highlights that a well-trained generative model plays a crucial role in achieving accurate risk estimation.

In the third set of experiments, we evaluate the performance of different optimization procedures solving Problem (13b). We use three general metrics: (*i*) Obj: the solution's objective value averaged across trials; (*ii*) Vio: the average ratio of solutions that violate the optimization constraint. and (*iii*) Err: the average distance between the solution and the ground-truth solution. The ground truth solution is computed from the closed-form solution under the linear setting, and approximated using a highly granular brute-force enumeration strategy in nonlinear settings. We include five baselines: (*i*) GD: a generic gradient descent algorithm solving the optimization by penalizing the objective function with the constraint; (*ii*) BF: brute force enumeration, where its discretization resolution is set to be the same as the number of iterations T for other methods; (*iii*) RS: random search algorithm, which is similar to BF except that the enumeration is done via sampling from a standard Gaussian distribution; (*iv*) RG: random guess, which is equivalent to a one-step

RS. These baselines are chosen as they are to the best of our knowledge, the few tractable procedures that can effectively handle Problem (13b). To ensure fair comparison, all the above optimization baselines are tuned with the number of iterations and/or epochs such that they take approximately the same time to run as our proposed method.

Table 3 reports the comparison of optimization procedures. Across all settings, CREDO consistently achieves low error. In the linear setting, CREDO attains zero error because the corresponding closed-form estimator recovers the exact solution. For nonlinear convex settings, CREDO continues to exhibit low overall error, attributable to its ability to balance achieving a small objective value (Obj) while maintaining a low violation ratio (Vio). Together, these results demonstrate that CREDO's optimization procedure effectively handles both linear and nonlinear problems, supporting its efficacy and suitability for decision risk–assessment tasks and yielding the strong risk-estimation performance observed in Section 7.1.

8. Discussion and Conclusions

We introduced CREDO, a distribution-free framework for decision risk assessment that fundamentally shifts the paradigm from prescribing decisions to evaluating their reliability under uncertainty. By combining inverse optimization with conformal prediction, CREDO provides calibrated estimates of the probability that any candidate decision may be suboptimal, enabling practitioners to audit both human-proposed and algorithm-generated decisions with statistical guarantees.

Our theoretical analysis establishes that CREDO achieves conservative risk estimates, providing expectation-wise validity even under model misspecification. The framework's consistency properties ensure that these bounds tighten asymptotically as more data becomes available, while the true positive rate analysis demonstrates that generative sampling effectively reduces excessive conservativeness. These properties position CREDO as both theoretically rigorous and practically useful, addressing the long-standing challenge of quantifying decision stability without distributional assumptions.

The computational strategies we developed make these theoretical guarantees accessible in practice. For linear programs, our closed-form solution leverages the polytope structure to evaluate risk efficiently through vertex enumeration. For general convex problems, our gradient-based approach using differentiable optimization layers provides a practical approximation scheme. While computational complexity increases with problem scale and the number of generative samples, the modular design allows decision makers to balance computational cost against the desired level of risk assessment fidelity.

Our empirical evaluation reveals important trade-offs that emerge when deploying CREDO. The choice of calibrated radius fundamentally determines the balance between conservativeness and accuracy. The e-value approach offers rigorous post-hoc validity guarantees, but alternatives such as the p-value variant provide tighter risk estimates at the cost of weakened validity. Selecting an appropriate radius thus requires balancing the trade-off between validity and informativeness, ensuring that decision support remains both safe and

actionable. This flexibility allows practitioners to select the variant that best matches their risk tolerance and application requirements. Moreover, our experiments demonstrate that CREDO's effectiveness scales with the quality of the underlying generative model. While validity is maintained even under misspecification, the informativeness of risk estimates depends critically on how well the generative model captures the true parameter distribution.

The framework's applicability extends to operational domains where parameter uncertainty significantly impacts decision quality. For example, for inventory management systems, CREDO can evaluate the robustness of reorder point policies and safety stock levels against demand variability and lead time uncertainty, providing quantitative assessments of when established inventory rules may fail to minimize total costs. Rather than prescribing singular optimal solutions, our framework provides decision-makers with risk profiles that characterize solution stability across the parameter space, enabling more informed trade-offs between expected performance and robustness. This paradigm shift from deterministic optimization to risk-aware assessment is particularly valuable in regulated industries where decision justification and risk documentation are required, as CREDO's calibrated guarantees provide auditable evidence of decision quality under uncertainty.

Several limitations of the current framework suggest directions for future research. (*i*) Overconservativeness with high-dimensional $\mathcal{Y}$: CREDO's risk estimates become increasingly loose as the dimension of $\mathcal{Y}$ grows due to the curse of dimensionality—conformal sets expand in volume with dimension even at fixed coverage. While this preserves validity, it can reduce practical usefulness in decision-making tasks with many uncertain components. Future work may explore more flexible prediction-set geometries to improve efficiency (Izbicki et al. 2022, Zheng and Zhu 2024). (*ii*) Lack of convergence guarantees for the alternating algorithm: Our alternating scheme in Algorithm 2 works well empirically for solving (18), but a formal convergence analysis remains open. The problem structure suggests potential for adapting contraction-type or block coordinate descent results, offering a promising avenue for theoretical study. (*iii*) Human-in-the-loop evaluation: Because CREDO is intended to support human–algorithm collaboration, controlled user studies are useful to examine how real-world practitioners interact with its conservative risk assessments and whether these assessments improve decision quality relative to purely algorithmic recommendations.

To conclude, CREDO represents a step toward more transparent and accountable decision support systems. As machine learning increasingly influences critical decisions across all industry and public policy domains, the ability to quantify decision risk becomes essential for responsible deployment. The framework's distribution-free guarantees are particularly valuable in high-stakes settings where distributional assumptions are difficult to verify or where model misspecification could have severe consequences. By enabling algorithms to express uncertainty about their recommendations, CREDO facilitates a more nuanced collaboration between human expertise and machine intelligence, where each can contribute their respective strengths.

Acknowledgments

Wenbin Zhou and Shixiang Zhu 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, Center for Intelligent Business. Agni Orfanoudaki acknowledges support from the AI^2 Partnership Grant (from the UKRI and the AXA Insurance company), which facilitated this work.

References

Agrawal A, Amos B, Barratt S, Boyd S, Diamond S, Kolter Z (2019) Differentiable convex optimization layers. *Advances in Neural Information Processing Systems*.

Ahuja RK, Orlin JB (2001) Inverse optimization. *Operations research* 49(5):771–783.

Ajay A, Du Y, Gupta A, Tenenbaum J, Jaakkola T, Agrawal P (2022) Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657* .

Amos B, Kolter JZ (2017) Optnet: Differentiable optimization as a layer in neural networks. *International conference on machine learning*, 136–145 (PMLR).

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

Aswani A, Shen ZJ, Siddiq A (2018) Inverse optimization with noisy data. *Operations Research* 66(3):870–892.

Balinsky AA, Balinsky AD (2024) Enhancing conformal prediction using e-test statistics. *The 13th Symposium on Conformal and Probabilistic Prediction with Applications*, 65–72 (PMLR).

Barber RF, Candes EJ, Ramdas A, Tibshirani RJ (2023) Conformal prediction beyond exchangeability. *The Annals of Statistics* 51(2):816–845.

Ben-Tal A, Nemirovski A (2002) Robust optimization–methodology and applications. *Mathematical programming* 92(3):453–480.

Bertsimas D, Boussioux L, Cory-Wright R, Delarue A, Digalakis V, Jacquillat A, Kitane DL, Lukin G, Li M, Mingardi L, et al. (2021) From predictions to prescriptions: A data-driven response to covid-19. *Health care management science* 24:253–272.

Bertsimas D, Gupta V, Kallus N (2018) Robust sample average approximation. *Mathematical Programming* 171(1):217–282.

Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* 66(3):1025–1044.

Bertsimas D, Sim M (2004) The price of robustness. *Operations research* 52(1):35–53.

Bertsimas D, Thiele A (2006) Robust and data-driven optimization: modern decision making under uncertainty. *Models, methods, and applications for innovative decision making*, 95–122 (INFORMS).

Bertsimas D, Tsitsiklis JN (1997) *Introduction to linear optimization*, volume 6 (Athena scientific Belmont, MA).

Besbes O, Fonseca Y, Lobel I (2025) Contextual inverse optimization: Offline and online learning. *Operations Research* 73(1):424–443.

Chan T, Delage E, Lin B (2024) Conformal inverse optimization for adherence-aware prescriptive analytics. *Available at SSRN* .

Chan TC, Mahmood R, Zhu IY (2025) Inverse optimization: Theory and applications. *Operations Research* 73(2):1046–1074.

Chen S, Fioretto F, Qiu F, Zhu S (2025) Global-decision-focused neural odes for proactive grid resilience management. *arXiv preprint arXiv:2502.18321* .

Cortes-Gomez S, Patiño C, Byun Y, Wu S, Horvitz E, Wilder B (2024) Decision-focused uncertainty quantification. *arXiv preprint arXiv:2410.01767* .

Cresswell JC, Sui Y, Kumar B, Vouitsis N (2024) Conformal prediction sets improve human decision making. *Forty-first International Conference on Machine Learning*.

DeCarolis JF (2011) Using modeling to generate alternatives (mga) to expand our thinking on energy futures. *Energy Economics* 33(2):145–152.

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

Delarue A, Lian Z, Martin S (2025) Algorithmic precision and human decision: A study of interactive optimization for school schedules. *Management Science* .

Diamond S, Boyd S (2016) CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17(83):1–5.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science* 64(3):1155–1170.

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

Fukuda K (1997) cdd/cdd+ reference manual. *Institute for Operations Research, ETH-Zentrum* 91–111.

Fukuda K, Prodon A (1995) Double description method revisited. *Franco-Japanese and Franco-Chinese conference on combinatorics and computer science*, 91–111 (Springer).

Gauthier E, Bach F, Jordan MI (2025a) Backward conformal prediction. *arXiv preprint arXiv:2505.13732* .

Gauthier E, Bach F, Jordan MI (2025b) E-values expand the scope of conformal prediction. *arXiv preprint arXiv:2503.13050* .

Grand-Clément J, Pauphilet J (2024) The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science* .

Grünwald P, de Heide R, Koolen W (2024) Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86(5):1091–1128.

Hullman J, Wu Y, Xie D, Guo Z, Gelman A (2025) Conformal prediction and human decision making. *arXiv preprint arXiv:2503.11709* .

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Iyengar G, Kang W (2005) Inverse conic programming with applications. *Operations Research Letters* 33(3):319–330.

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

Kim K, Mehrotra S (2015) A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research* 63(6):1431–1451.

Kingma DP, Welling M (2013) Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Kiyani S, Pappas G, Roth A, Hassani H (2025) Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. *arXiv preprint arXiv:2502.02561* .

Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization* 12(2):479–502.

Kochenderfer MJ (2015) *Decision making under uncertainty: theory and application* (MIT press).

Krein M, Milman D (1940) On extreme points of regular convex sets. *Studia Mathematica* 9:133–138.

Lan H, Liao L, Elmachtoub AN, Kroer C, Lam H, Zhang H (2025) The bias-variance tradeoff in data-driven optimization: A local misspecification perspective. *arXiv preprint arXiv:2510.18215* .

Li ML, Zhu S (2024) Balancing optimality and diversity: Human-centered decision making through generative curation. *arXiv preprint arXiv:2409.11535* .

Lin B, Delage E, Chan T (2024) Conformal inverse optimization. *Advances in Neural Information Processing Systems* 37:63534–63564.

Lombardi F, van Greevenbroek K, Grochowicz A, Lau M, Neumann F, Patankar N, Vågerö O (2025) Near-optimal energy planning strategies with modeling to generate alternatives to flexibly explore practically desirable options. *Joule* .

Mandi J, Kotary J, Berden S, Mulamba M, Bucarey V, Guns T, Fioretto F (2024) Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research* 80:1623–1701.

Motzkin TS, Raiffa H, Thompson GL, Thrall RM (1953) The double description method. *Contributions to the Theory of Games* 2(28):51–73.

Orfanoudaki A, Saghafian S, Song K, Chakkera HA, Cook C (2022) Algorithm, human, or the centaur: How to enhance clinical care? .

Palmintier B (2014) Flexibility in generation planning: Identifying key operating constraints. *2014 power systems computation conference*, 1–7 (IEEE).

Papadopoulos H, Proedrou K, Vovk V, Gammerman A (2002) Inductive confidence machines for regression. *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, 345–356 (Springer).

Parenti N, Reggiani MLB, Iannone P, Percudani D, Dowding D (2014) A systematic review on the validity and reliability of an emergency department triage scale, the manchester triage system. *International Journal of Nursing Studies* 51(7):1062–1069, ISSN 0020-7489, URL http://dx.doi.org/https://doi.org/10.1016/j.ijnurstu.2014.01.013.

Patel YP, Rayan S, Tewari A (2024) Conformal contextual robust optimization. *International Conference on Artificial Intelligence and Statistics*, 2485–2493 (PMLR).

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in python. *the Journal of machine Learning research* 12:2825–2830.

Prinster D, Saria S, Liu A (2023) Jaws-x: addressing efficiency bottlenecks of conformal prediction under standard and feedback covariate shift. *International Conference on Machine Learning*, 28167–28190 (PMLR).

Rahimian H, Mehrotra S (2022) Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization* 3:1–85.

Schaefer AJ (2009) Inverse integer programming. *Optimization Letters* 3:483–489.

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

Shapiro A, Dentcheva D, Ruszczynski A (2021) *Lectures on stochastic programming: modeling and theory* (SIAM).

Singh S, Sarna N, Li Y, Lin Y, Orfanoudaki A, Berger M (2024) Distribution-free risk assessment of regression-based machine learning algorithms. *The 13th Symposium on Conformal and Probabilistic Prediction with Applications*, 44–64 (PMLR).

Tao PD, An LH (1997) Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica* 22(1):289–355.

Tavaslıoğlu O, Lee T, Valeva S, Schaefer AJ (2018) On the structure of the inverse-feasible region of a linear program. *Operations Research Letters* 46(1):147–152.

Tian X, Yan R, Liu Y, Wang S (2023) A smart predict-then-optimize method for targeted and cost-effective maritime transportation. *Transportation Research Part B: Methodological* 172:32–52.

Topkis DM (1998) *Supermodularity and complementarity* (Princeton university press).

Vovk V (2025) Conformal e-prediction. *Pattern Recognition* 111674.

Vovk V, Gammerman A, Shafer G (2005) *Algorithmic learning in a random world*, volume 29 (Springer).

Wang Z, Gao R, Yin M, Zhou M, Blei D (2023) Probabilistic conformal prediction using conditional random samples. *International Conference on Artificial Intelligence and Statistics*, 8814–8836 (PMLR).

Wright SJ (2015) Coordinate descent algorithms. *Mathematical programming* 151(1):3–34.

Yeh C, Christianson N, Wu A, Wierman A, Yue Y (2024) End-to-end conformal calibration for optimization under uncertainty. *arXiv preprint arXiv:2409.20534* .

Zattoni Scroccaro P, Atasoy B, Mohajerin Esfahani P (2025) Learning in inverse optimization: Incenter cost, augmented suboptimality loss, and algorithms. *Operations Research* 73(5):2661–2679.

Zhang Z, Shi P, Ward A (2025) Admission decisions under imperfect classification: An application in criminal justice. *Available at SSRN 5214197* .

Zhao L, Jiang H, Qi W (2025) Conformal robust optimization and satisficing for prescriptive analytics. *Available at SSRN 5338354* .

Zheng M, Zhu S (2024) Generative conformal prediction with vectorized non-conformity scores. *arXiv preprint arXiv:2410.13735* .

Zhou W, Zhu S, Qiu F, Wu X (2025) Hierarchical spatio-temporal uncertainty quantification for distributed energy adoption. *2025 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5 (IEEE).

Zhu S, Wang H, Xie Y (2022) Data-driven optimization for atlanta police-zone design. *INFORMS Journal on Applied Analytics* 52(5):412–432.

Appendix A: Proofs

A.1. Proof of Theorem 1

For notation clarity, we will denote the k-th prediction set constructed by $\hat{y}^{(k)}$ as $C^k(\cdot;\cdot)$ in this and the following sections. We begin by proving *E-value post-hoc validity*, which is a key lemma to proving Theorem 1.

LEMMA 3 **(E-value post-hoc validity)**. *Denote the prediction set constructed by $\hat{y}^{(k)}$ as C^k. Then under Assumption 1,*

$$\mathbb{P}\left\{Y \in C^{(k)}(X;\hat{\alpha})\right\} \geq 1 - \mathbb{E}[\hat{\alpha}], \quad \forall k = 1,\ldots,K.$$

where $\hat{\alpha}$ is some arbitrary function of $\{(X_i,Y_i)\}_{i=1}^{n+1}$.

Proof of Lemma 3 When $\hat{\alpha} < 1/(n+1)$ or $\hat{\alpha} = 1$, by the definition of all three variants of $R(\alpha)$, the statement holds trivially. So we only need to consider the case when $1 > \hat{\alpha} \geq 1/(n+1)$. Recall that $\hat{y}_i$ denotes a single prediction generated from the calibration data $g(x_i)$. For any $k = 1,\ldots,K$, we begin by expanding the left-hand side:

$$\begin{aligned}
\mathbb{P}\left\{Y \notin C^{(k)}(X;\hat{\alpha})\right\} &= \mathbb{P}\left\{\|\hat{y}^{(k)} - Y\|_2 > \frac{\sum_{i=1}^n \|\hat{y}_i - Y_i\|_2}{\hat{\alpha}(n+1) - 1}\right\} \\
&= \mathbb{P}\left\{\hat{\alpha}(n+1)\|Y - \hat{y}^{(k)}\|_2 > \sum_{i=1}^n \|\hat{y}_i - Y_i\|_2 + \|\hat{y}^{(k)} - Y\|_2\right\} \\
&= \mathbb{P}\left\{\hat{\alpha} > \frac{\sum_{i=1}^n \|\hat{y}_i - Y_i\|_2 + \|\hat{y}^{(k)} - Y\|_2}{(n+1)\|\hat{y}^{(k)} - Y\|_2}\right\} \\
&= \mathbb{P}\left\{\frac{(n+1)\|\hat{y}^{(k)} - Y\|_2}{\sum_{i=1}^n \|\hat{y}_i - Y_i\|_2 + \|\hat{y}^{(k)} - Y\|_2} > \frac{1}{\hat{\alpha}}\right\}.
\end{aligned}$$

Denote the following random variables,

$$\begin{aligned}
F_i &= \frac{(n+1)\|\hat{y}_i - Y_i\|_2}{\sum_{i=1}^n \|\hat{y}_i - Y_i\|_2 + \|\hat{y}^{(k)} - Y\|_2}, \quad \forall i = 1,\ldots,n, \\
F_{n+1} &= \frac{(n+1)\|\hat{y}^{(k)} - Y\|_2}{\sum_{i=1}^n \|\hat{y}_i - Y_i\|_2 + \|\hat{y}^{(k)} - Y\|_2}.
\end{aligned}$$

One can prove that the following two conditions hold:

$$\begin{aligned}
(a): &\quad \mathbb{E}[F_1 + \ldots + F_n + F_{n+1}] = n+1, \\
(b): &\quad \mathbb{E}[F_1] = \ldots = \mathbb{E}[F_n] = \mathbb{E}[F_{n+1}],
\end{aligned}$$

where (b) holds by exchangeability (Assumption 1). Therefore, there is

$$\mathbb{E}[F_{n+1}] = 1.$$

Using this result, it can be derived that

$$\sup_{\tilde{\alpha}} \mathbb{E}\left[\frac{\mathbb{P}(F_{n+1} > 1/\tilde{\alpha})}{\tilde{\alpha}}\right] \leq \sup_{\tilde{\alpha}} \mathbb{E}\left[\frac{\tilde{\alpha}\cdot\mathbb{E}[F_{n+1}]}{\tilde{\alpha}}\right] = \mathbb{E}[F_{n+1}] = 1,$$

where the inequality follows from Markov's inequality. Therefore, for any $\hat{\alpha}$ which may depend on the data $\{X_i,Y_i\}_{i=1}^{n+1}$, there is:

$$\mathbb{E}\left[\frac{\mathbb{P}(F_{n+1} > 1/\hat{\alpha} \mid \hat{\alpha})}{\hat{\alpha}}\right] \leq \sup_{\tilde{\alpha}} \mathbb{E}\left[\frac{\mathbb{P}(F_{n+1} > 1/\tilde{\alpha})}{\tilde{\alpha}}\right] \leq 1$$

Using a first-order Taylor expansion, the left-hand side of the first inequality is:

$$\mathbb{E}\left[\frac{\mathbb{P}(F_{n+1} > 1/\hat{\alpha} \mid \hat{\alpha})}{\hat{\alpha}}\right] \approx \frac{\mathbb{E}\left[\mathbb{P}(F_{n+1} > 1/\hat{\alpha} \mid \hat{\alpha})\right]}{\mathbb{E}[\hat{\alpha}]} \tag{19}$$

Assuming that this approximation is exact (see discussion in Remark 2), and by combining the two equations above, there is

$$\mathbb{E}\left[\mathbb{P}(F_{n+1} > 1/\hat{\alpha} \mid \hat{\alpha})\right] \leq \mathbb{E}[\hat{\alpha}] \Longleftrightarrow \mathbb{P}\left\{Y \in C^{(k)}(X; \hat{\alpha})\right\} \geq 1 - \mathbb{E}[\hat{\alpha}],$$

where $\hat{\alpha}$ can probabilistically depend on $\{(X_i, Y_i)\}_{i=1}^{n+1}$ in arbitrary ways. □

Proof of Theorem 1 We begin by noticing the following decomposition

$$\mathbb{P}\left\{z \in \pi(Y)\right\} = \mathbb{P}\left\{Y \in \pi^{-1}(z)\right\} \geq \frac{1}{K}\sum_{k=1}^{K} \mathbb{P}\left\{Y \in C^{(k)}\left(X; \tilde{\alpha}^{(k)}(z)\right)\right\}. \tag{20}$$

The first equality follows from our problem reformulation (Lemma 1). The second inequality holds due to the definition of $\tilde{\alpha}^{(k)}(z)$, which guarantees that the k-th generated CP region is always contained in $\pi^{-1}(z)$. By Lemma 3, there is:

$$\mathbb{P}\left\{Y \in C^{(k)}\left(X; \tilde{\alpha}^{(k)}(z)\right)\right\} \geq 1 - \mathbb{E}\left[\tilde{\alpha}^{(k)}(z)\right].$$

Therefore, combining this with (20) we obtain:

$$\mathbb{P}\left\{z \in \pi(Y)\right\} \geq 1 - \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \tilde{\alpha}^{(k)}(z)\right] = 1 - \mathbb{E}\left[\hat{\alpha}(z)\right].$$

We conclude the proof for Theorem 1. □

Remark 2 (Discussion on the Taylor approximation). We comment on the rationality of the first-order Taylor approximation used in (19). To begin with, the approximation trick has been adopted in prior works (Gauthier et al. 2025b), and argued that its error is small when the estimator $\hat{\alpha}$ is well concentrated around its mean. Such a condition is usually satisfied in our setting. For example, when CREDO is deployed in a human-algorithm collaboration setting, the candidate decisions provided from the decision maker would be expected to be near optimal and should already enjoy a relatively small ground truth risk. This makes $\hat{\alpha}$ have a relatively small variance and well concentrated around its mean, which allows the Taylor approximation to be relatively tight.

Even when this condition does not hold, we can resort to an alternative way to account for the approximation error during risk estimator construction—deriving the approximation error and then manually offsetting the error in our risk estimator to achieve an exact conservativeness guarantee. The derivation goes as follows: let $h(\hat{\alpha}) \coloneqq \mathbb{E}[\mathbb{P}(F_{n+1} > 1/\hat{\alpha} \mid \hat{\alpha})]$, there is:

$$\left|\mathbb{E}\left[\frac{h(\hat{\alpha})}{\hat{\alpha}}\right] - \frac{\mathbb{E}[h(\hat{\alpha})]}{\mathbb{E}[\hat{\alpha}]}\right| \leq \mathbb{E}[h(\hat{\alpha})]\left(\mathbb{E}\left[\frac{1}{\hat{\alpha}}\right] - \frac{1}{\mathbb{E}[\hat{\alpha}]}\right) + \sqrt{\mathrm{Var}(h(\hat{\alpha})) \cdot \mathrm{Var}\left(\frac{1}{\hat{\alpha}}\right)}.$$

The first term is the Jensen gap, and the second term is by the Cauchy-Schwarz inequality. Assuming $\hat{\alpha} \in [\delta, 1]$ almost surely, then

$$\mathbb{E}\left[\frac{1}{\hat{\alpha}}\right] - \frac{1}{\mathbb{E}[\hat{\alpha}]} \leq \frac{1}{\delta^3}\mathrm{Var}(\hat{\alpha}), \quad \text{and} \quad \mathrm{Var}\left(\frac{1}{\hat{\alpha}}\right) \leq \frac{1}{\delta^3}\mathrm{Var}\left(\hat{\alpha}\right).$$

Plugging them into the first equation, we get

$$\left|\mathbb{E}\left[\frac{h(\hat{\alpha})}{\hat{\alpha}}\right] - \frac{\mathbb{E}[h(\hat{\alpha})]}{\mathbb{E}[\hat{\alpha}]}\right| \leq \frac{1}{\delta^3}\mathrm{Var}\left(\hat{\alpha}\right) + \frac{1}{2\delta^2}\sqrt{\mathrm{Var}(\hat{\alpha})}.$$

Since for random variables bounded within $[\delta, 1]$, there is the following trivial upper bound:

$$\mathrm{Var}(\hat{\alpha}) \leq \frac{1}{K^2} \sum_{k,k'} \mathrm{Cov}\left(I_k, I_k'\right) \leq \frac{1}{4},$$

therefore, we can conclude that:

$$\left| \mathbb{E}\left[\frac{\mathbb{P}(F_{n+1} > 1/\hat{\alpha} \mid \hat{\alpha})}{\hat{\alpha}} \right] - \frac{\mathbb{E}\left[\mathbb{P}(F_{n+1} > 1/\hat{\alpha} \mid \hat{\alpha})\right]}{\mathbb{E}[\hat{\alpha}]} \right| \leq \frac{1}{4\delta^3} + \frac{1}{4\delta^2}.$$

Plugging this result back into the proof of Theorem 1, we get

$$\mathbb{P}\left\{z \in \pi(Y)\right\} \geq 1 - \mathbb{E}[\hat{\alpha}(z)] - 1/4(\delta^{-3} + \delta^{-2}).$$

Therefore, one can take the final estimator as

$$\min\{\hat{\alpha}(z) + 1/4(\delta^{-3} + \delta^{-2}), 1\} \tag{21}$$

so that exact conservativeness is achieved. A trivial value that the user can take for δ is $1/(n+1)$, which is guaranteed by the design of the CREDO algorithm. One can also manually tune the value of δ by modifying the calibrated radius as

$$R'(\alpha) = \begin{cases} +\infty, & \text{if } \alpha \in [0, \delta), \\ R(\alpha) & \text{if } \alpha \in [\delta, 1), \\ 0 & \text{if } \alpha = 1, \end{cases}$$

to achieve a tighter bound (*i.e.*, smaller offset). One can prove that as long as δ is chosen such that $\delta > 1/(n+1)$, all theorems presented in the main text remain valid. So one can safely adopt $R'(\alpha)$ as the conformalized radius and take (21) as the final estimator. In the meantime, this procedure can be equivalently viewed as truncating the lower part of $\hat{\alpha}(z)$ at δ, *i.e.*, setting $\max\{\hat{\alpha}(z), \delta\}$ as the risk estimator, and then taking (21) as the final estimator.

A.2. Proof of Corollary 1

Proof Following the proof of Theorem 1 in Barber et al. (2023), with the exception that all steps are conditioned on the σ-field generated by $\tilde{\alpha}^{(k)}(z)$, there is:

$$\mathbb{P}\left[Y \in C^{(k)}(X; \tilde{\alpha}^{(k)}(z)) \middle| \tilde{\alpha}^{(k)}(z) \right] \geq 1 - \alpha - \frac{\sum_{i=1}^{n} d_{\mathrm{TV}}^{(i)}(\hat{\alpha}(z))}{n+1}.$$

Here we implicitly utilized the fact that the σ-field generated by $\tilde{\alpha}^{(k)}(z)$ is the same across all k, as they are just different simulation trials based on the same input and generative model, therefore they are equal to the σ-field of $\hat{\alpha}(z)$. Plugging this equation back in (20), we get

$$\mathbb{P}\left\{z \in \pi(Y)\right\} \geq 1 - \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\tilde{\alpha}^{(k)}(z) + \frac{\sum_{i=1}^{n} \mathbb{E}\left[d_{\mathrm{TV}}^{(i)}(\hat{\alpha}(z)) \right]}{n+1} \right] = 1 - \mathbb{E}\left[\tilde{\alpha}(z)\right] - \frac{\sum_{i=1}^{n} d_{\mathrm{TV}}^{(i)}(\hat{\alpha}(z))}{n+1}.$$

The expectations are taken with respect to all sources of randomness of $\tilde{\alpha}^{(k)}(z)$. □

A.3. Proof of Proposition 1

Proof By Remark 1, when using R_∞, the risk estimator is a Monte Carlo estimator:

$$\hat{\alpha}(z) = 1 - \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\left\{\hat{y}^{(k)} \in \pi^{-1}(z)\right\}.$$

Therefore, for any $x \in \mathcal{X}$, conditioning on $X = x$ and given $f(x) = \hat{\mathcal{P}}_{Y|X=x}$, there is

$$\begin{aligned}
|\hat{\alpha}(z) - \mathbb{P}\{z \in \pi(Y) \mid X = x\}| &= 1 - \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{1}\left\{\hat{y}^{(k)} \in \pi^{-1}(z)\right\} - \mathbb{P}\{z \in \pi(Y) \mid X = x\}\right) - 1 \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{1}\left\{\hat{y}^{(k)} \in \pi^{-1}(z)\right\} - \mathbb{P}\{z \in \pi(Y) \mid Y \sim \mathcal{P}_{Y|X=x}\}\right) \\
&= \frac{1}{K}\sum_{k=1}^{K}\Bigg(\underbrace{\mathbb{1}\left\{\hat{y}^{(k)} \in \pi^{-1}(z)\right\} - \mathbb{P}\{z \in \pi(Y) \mid Y \sim \hat{\mathcal{P}}_{Y|X=x}\}}_{A_k}\Bigg) + \\
&\quad \frac{1}{K}\sum_{k=1}^{K}\Bigg(\underbrace{\mathbb{P}\{z \in \pi(Y) \mid Y \sim \hat{\mathcal{P}}_{Y|X=x}\} - \mathbb{P}\{z \in \pi(Y) \mid Y \sim \mathcal{P}_{Y|X=x}\}}_{B_k}\Bigg) \\
&= \frac{1}{K}\sum_{k=1}^{K} A_k + \frac{1}{K}\sum_{k=1}^{K} B_k.
\end{aligned}$$

For the first term, since: (*i*) A_k are i.i.d. random variables, as $\hat{y}^{(k)}$ are independently generated from the same generative model given x; (*ii*) $\mathbb{E}[A_k \mid X = x] = 0$, (*iii*) A_k has finite variance, therefore by the central limit theorem, there is $\frac{1}{K}\sum_{k=1}^{K} A_k = O_p(K^{-1/2})$. For the second term, note that by the assumption

$$T_k^{(2)} = \mathbb{P}\{Y \in \pi^{-1}(z) \mid Y \sim \hat{\mathcal{P}}_{Y|X=x}\} - \mathbb{P}\{Y \in \pi^{-1}(z) \mid Y \sim \mathcal{P}_{Y|X=x}\} \le d_{\mathrm{TV}}(\hat{\mathcal{P}}_{Y|X=x}, \mathcal{P}_{Y|X=x}) \le \delta,$$

where the first inequality follows from the definition of total variation distance. Therefore, we conclude that

$$|\hat{\alpha}(z) - \mathbb{P}\{z \in \pi(Y) \mid X = x\}| = O_p(K^{-1/2}) + \delta.$$

This completes the proof. □

A.4. Proof of Proposition 2

Proof For notation simplicity, denote the sets in the numerator and the denominator of TPR defined in (12) as:

$$\begin{aligned}
A &= \{z \in \mathcal{Z} \mid \alpha(z) < 1 \text{ and } \hat{\alpha}(z) < 1\}, \\
B &= \{z \in \mathcal{Z} \mid \alpha(z) < 1\}.
\end{aligned}$$

For simplicity, we assume that B (therefore A) is a finite set, *i.e.*, there is only a finite set of decisions that have ground-truth risk smaller than one (though note that Proposition 2 and its proof naturally extend to the infinite case by replacing the counting measure "#" with continuous measures, such as Lebesgue measure defined within the decision space $\mathcal{Z}$). Since B is irrelevant to K, we begin by expanding the following expression:

$$\begin{aligned}
\mathbb{E}[\#A] &= \mathbb{E}\left[\sum_{z \in \mathcal{Z}} \mathbb{1}\{\hat{\alpha}(z) < 1 \text{ and } \alpha(z) < 1\}\right] \\
&= \mathbb{E}\left[\sum_{z \in B} \mathbb{1}\{\hat{\alpha}(z) < 1\}\right] \\
&= \mathbb{E}\left[\sum_{z \in B}\left(1 - \prod_{k=1}^{K}\left(\mathbb{1}\left\{\hat{y}^{(k)} \notin \pi^{-1}(z)\right\}\right)\right)\right]
\end{aligned}$$

The third equality results from the observation that the risk estimation is not one when at least one $\hat{y}^{(k)}$ falls within $\pi^{-1}(z)$. By the last expression, it can be seen that $\mathbb{E}[\#A]$ monotonically increases with K. Since $\mathrm{TPR} = \mathbb{E}[\#A]/\mathbb{E}[\#B]$, it also monotonically increases with K. □

A.5. Proof of Proposition 3

Proof The first part of the proof resembles the main text. We begin by noticing that the set containment condition $C(x;\alpha) \subseteq \pi^{-1}(z)$ is equivalent to requiring that the calibrated conformal ball does *not* intersect the complement of the inverse feasible region:

$$C(x;\alpha) \subseteq \pi^{-1}(z) \quad \Longleftrightarrow \quad C(x;\alpha) \cap \left(\pi^{-1}(z)\right)^c = \emptyset.$$

Since $C(x;\alpha)$ is an ℓ_2 ball centered at $\hat{y}$, the right hand side is further equivalent to

$$\underbrace{\text{for all } y \in \left(\pi^{-1}(z)\right)^c,\ \|y-\hat{y}\|_2 \geq R(\alpha)}_{(i)} \qquad \text{if } \underbrace{\hat{y} \in \pi^{-1}(z)}_{(ii)}. \tag{22}$$

Here, condition (i) requires that every scenario y violating the optimality of decision z must lie at least with distance $R(\alpha)$ away from $\hat{y}$; condition (ii) states that the former condition is enforced only when $\hat{y}$ lies within of $\pi^{-1}(z)$. Otherwise, then no radius can guarantee containment, and the α is trivially mapped to one by design. Therefore under (ii), (10) can be reformulated as

$$\min_{\alpha} \max_{y} \alpha \quad \text{s.t.} \quad \|y-\hat{y}\|_2 \geq R(\alpha) \quad \text{and} \quad y \in \left(\pi^{-1}(z)\right)^c. \tag{23}$$

Since by the monotonicity of R, there is

$$\|y-\hat{y}\|_2 \geq R(\alpha) \quad \Longleftrightarrow \quad \alpha \geq R^{-1}\left(\|y-\hat{y}\|_2\right),$$

so the first constraint in (23) can be replaced, and the optimization can be rewritten as

$$\min_{\alpha} \max_{y} \alpha \quad \text{s.t.} \quad \alpha \geq R^{-1}\left(\|y-\hat{y}\|_2\right) \quad \text{and} \quad y \in \left(\pi^{-1}(z)\right)^c.$$

In this problem. α can be viewed as a slack variable, and can be dropped so that the constraint becomes the objective:

$$\max_{y} R^{-1}\left(\|y-\hat{y}\|_2\right) \quad \text{s.t.} \quad y \in \left(\pi^{-1}(z)\right)^c.$$

This turns the original minmax problem into a single-level maximization problem. Finally, by definition of π^{-1}, the constraint can be expanded as

$$y \in \left(\pi^{-1}(z)\right)^c \quad \Longleftrightarrow \quad f(z;y) > \min_{z' \in \mathcal{Z}} f(z';y) + \epsilon.$$

Plugging this into the previously derived optimization yields the desired statement. □

A.6. Proof of Corollary 2

Proof Using the linear assumption, we begin the derivation by expanding the constraint:

$$f(z;y) > \min_{z' \in \mathcal{Z}} f(z';y) + \epsilon \iff \min_{z' \in \mathcal{Z}} \langle y, z'-z \rangle + \epsilon < 0$$

Since $\mathcal{Z}$ is a compact set, by the Krein–Milman theorem (Krein and Milman 1940), there is

$$\min_{z' \in \mathcal{Z}} \langle y, z'-z \rangle = \min_{v \in \mathcal{V}} \langle y, v-z \rangle.$$

Plugging this into (13b) derived in Proposition 3, and using similar reformulation arguement as (18), we only need to solve the following optimization problem:

$$\min_{y} \|y - \hat{y}\|_2 \quad \text{s.t.} \quad \exists v \in \mathcal{V} \setminus \{z\},\ \langle y, v - z\rangle + \epsilon < 0$$

This optimization finds the closest distance from a point $\hat{y}$ to *any* halfspaces defined by norm vectors $z - v$ and offset ϵ. By using the well-known point to halfspace distance formula, the optimal objective value of this optimization problem is:

$$\min_{v \in \mathcal{V}\setminus\{z\}} \frac{|\langle \hat{y}, z - v\rangle - \epsilon|}{\|z - v\|_2}.$$

Therefore, the expression for the risk estimate is

$$\tilde{\alpha}(z) = R^{-1}\left(\min_{v \in \mathcal{V}\setminus\{z\}} \frac{|\langle \hat{y}, z - v\rangle - \epsilon|}{\|z - v\|_2}\right) = \max_{v \in \mathcal{V}\setminus\{z\}} R^{-1}\left(\frac{|\langle \hat{y}, z - v\rangle - \epsilon|}{\|z - v\|_2}\right), \quad \text{if } z \in \pi_\epsilon(\hat{y}),$$

where the second equality follows from the monotonic decreasing property of R^{-1}. This concludes the proof. □

Remark 3 (Indicator term). We note that under the linear case, the indicator term also admits the following closed-form:

$$\mathbb{1}\left\{\hat{y}^{(k)} \in \pi^{-1}(z)\right\} = \prod_{v \in \mathcal{V}} \mathbb{1}\left\{\langle \hat{y}^{(k)}, z - v\rangle \le 0\right\}. \tag{24}$$

To see why, we begin with the following expansion:

$$\mathbb{1}\left\{y \in \pi^{-1}(z)\right\} = \mathbb{1}\left\{f(z; y) \le \min_{z' \in \mathcal{Z}} f(z'; y) + \epsilon\right\} = \mathbb{1}\left\{\max_{z' \in \mathcal{Z}} \langle y, z - z'\rangle \le \epsilon\right\}.$$

Using the Krein–Milman theorem from the proof above, we can derive that the right-hand side is equal to:

$$\mathbb{1}\left\{\max_{z' \in \mathcal{Z}} \langle y, z - z'\rangle \le \epsilon\right\} = \prod_{v \in \mathcal{V}} \mathbb{1}\left\{\langle y, z - v\rangle \le \epsilon\right\}.$$

Finally, setting $y = \hat{y}^{(k)}$ yields (24), which concludes the proof.

Appendix B: Additional Experiment Details

This section presents additional details of the numerical experiments that extend what has been described in the main text, and is organized as follows: Section B.1 shows the resources and dependencies; Section B.2 presents the detailed configurations for the optimization problems; Section B.3 describes the detailed configurations of the used baselines; Section B.4 and Section B.5 describes the metrics and additional experiment results.

B.1. Implementation Environments

All experiments were conducted on a machine running Windows 11, equipped with a 13th-generation Intel Core i7 CPU with 16 cores and 16 GB of RAM. No GPU acceleration was used in any of the experiments. The code is implemented in Python, and we list some key external packages: `Scikit-learn` (Pedregosa et al. 2011) is used for implementing Gaussian mixture models and some other statistical models that were used as baselines; `CDD` (Fukuda 1997) is used for computing the vertices of polytopes to compute the closed-form solutions under the linear assumption; `cvxpy` and `cvxpylayers` (Diamond and Boyd 2016, Agrawal et al. 2019) are used to solve convex optimization problems and implement differentiable optimization layers; `PyTorch` is used to build gradient-based baseline models. Unless otherwise specified, we use the default parameter for all models in the package. The complete code and dependencies are available in our codebase.

B.2. Optimization problems

We describe the details of the six optimization problems that are featured in our experiments. Throughout this section, we denote σ as the component variance scale ($\sigma = 1$ by default), and $\mathbf{I}_d$ as the d-dimensional identity matrix.

LP (Setting I) A linear programming problem with a triangular feasible region, defined as:

$$\min_{z \in \mathbb{R}^2} Y_1 z_1 + Y_2 z_2 \quad \text{s.t.} \quad z_1 + z_2 \leq 1, z_1 \geq 0, z_2 \geq 0,$$

where $(Y_1, Y_2)^\top$ is a Gaussian random vector with mean $(-1, -1)^\top$ and covaraince matrix $\sigma \cdot \mathbf{I}_2$. This optimization problem can be interpreted as a profit maximization task, where a manufacturer chooses the optimal production quantities z_1 and z_2 under a budget constraint, and Y_1 and Y_2 represent a risky market scenario that could still yield profit under favorable conditions where the expected revenue is negative. Its feasible region can be more compactly represented in matrix form $\mathbf{A}z \leq \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

LP (Setting II) A linear programming problem that employs a more complex octagonal feasible region with five vertices and multimodal objective uncertainty, defined as

$$\min_{z \in \mathbb{R}^2} Y_1 z_1 + Y_2 z_2 \quad \text{s.t.} \quad \mathbf{A}z \leq \mathbf{b},$$

where the constraint matrix $\mathbf{A}$ and constraint vector $\mathbf{b}$ are defined as

$$\mathbf{A} = \begin{pmatrix} -0.5 & 0 & -0.5 & 0.5 & 2 & 1 & 0 & -1 \\ -1 & -1 & 1 & 1 & -1 & 0 & 1 & 0 \end{pmatrix}^\top, \quad \mathbf{b} = \begin{pmatrix} -1 & 0 & 1 & 5 & 10 & 5.5 & 2.5 & -1 \end{pmatrix}^\top.$$

The random vector $Y \in \mathbb{R}^2$ is drawn from a three-component Gaussian mixture distribution

$$p(x) = \sum_{k=1}^{3} w_k \, \mathcal{N}\left(x \mid \mu_k, \, \sigma \cdot \sigma_k^2 \mathbf{I}_2\right),$$

where the mixture weights are $\boldsymbol{w} = (0.3, 0.4, 0.3)$, the component means $\boldsymbol{\mu}$ are defined as

$$\mu_1 = \left(0.0, -0.8\right)^\top, \quad \mu_2 = \left(-0.5, 0.25\right)^\top, \quad \mu_3 = \left(0.8, -0.1\right)^\top,$$

and the component variances $\boldsymbol{\sigma}$ are defined as

$$\sigma_1^2 = (0.01)^2, \quad \sigma_2^2 = (0.03)^2, \quad \sigma_3^2 = (0.02)^2.$$

Compared to the first setting, this setting allows us to assess risk assessment performance in more complex scenarios with multiple potentially optimal decisions.

QP A quadratic programming problem defined as

$$\min_{z \in \mathbb{R}^2} \frac{1}{2} z^\top Q z + \langle z, y \rangle \quad \text{s.t.} \quad \mathbf{A}z \leq \mathbf{b},$$

where the parameters are defined similar to *LP (Setting I)*:

$$Q = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

SOCP A second-order conic programming problem, defined as

$$\min_{z\in\mathbb{R}^2} \langle z, y\rangle \quad \text{s.t.} \quad 0.1\cdot\|z\|_2 \le \mathbf{A}z - \mathbf{b}.$$

where the parameters are defined similar to *LP (Setting I)*:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

IP An integer programming problem, defined as

$$\min_{z\in\mathbb{Z}^2} \langle z, y\rangle \quad \text{s.t.} \quad \mathbf{A}z - \mathbf{b} \le 0.$$

where the parameters are defined similar to *LP (Setting II)*:

$$\mathbf{A} = \begin{pmatrix} -0.5 & 0 & -0.5 & 0.5 & 2 & 1 & 0 & -1 \\ -1 & -1 & 1 & 1 & -1 & 0 & 1 & 0 \end{pmatrix}^\top, \quad \mathbf{b} = \begin{pmatrix} -1 & 0 & 1 & 5 & 10 & 5.5 & 2.5 & -1 \end{pmatrix}^\top.$$

Real Data This setting considers a real-world power grid investment decision-making problem motivated by (Zhou et al. 2025). A utility company based in Indianapolis, Indiana, has compiled detailed records of over 1,700 solar panel installations between 2010 and 2024, including the installation dates and affiliated grid components. With the renewable energy sector now at full scale, the management team anticipates a steady and significant monthly increase in solar adoption in the downtown area. In preparation for the incoming demand, they are planning targeted upgrades to grid-level inverters at $d = 4$ selected substations (we refer to them as Substation A to D) to minimize loss for grid failures subject to a limited budget. This decision-making problem can be formulated as a penalized knapsack problem, defined as

$$\min_{\mathbf{a}\in\{0,1\}^d} \sum_{i=1}^{d} \mathbb{1}(Y_i \ge \tau_i)(1-a_i)l_i \quad \text{s.t.} \quad \sum_{i=1}^{d} a_i c_i \le b,$$

where $a_i \in 0, 1$ is a binary decision variable indicating whether substation i is upgraded ($a_i = 1$) or not ($a_i = 0$); Y_i is a random variable representing the monthly increase in solar panel installations connected to substation i; c_i denotes the cost associated with upgrading substation i; τ_i is the capacity threshold, corresponding to the maximum allowable number of solar connections at substation i; l_i represents the penalty incurred when the realized installations exceed τ_i without an upgrade; and b denotes the total available budget.

The optimization problem above can be further relaxed into an LP problem:

$$\max_{z\in\mathbb{R}^d} \langle Y, z\rangle \quad \text{s.t.} \quad \mathbf{A}z \le \mathbf{b},$$

where the parameters are defined as

$$Y = \begin{pmatrix} l_1\cdot(1+e^{-\beta(Y_1-\tau_1)})^{-1} \\ \cdots \\ l_d\cdot(1+e^{-\beta(Y_d-\tau_d)})^{-1} \end{pmatrix}, \quad z = \begin{pmatrix} 1-a_1 \\ \cdots \\ 1-a_d \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} -\mathbf{c}^\top \\ \mathbf{I}_d \\ -\mathbf{I}_d \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b - \mathbf{1}_d^\top \mathbf{c} \\ \mathbf{1}_d \\ \mathbf{0}_d \end{pmatrix}.$$

They are manually configured as follows: The cost c_i and loss l_i for all substations equals one unit; The capacity threshold τ_i is set to be the historical average solar panel monthly increment; The budget b is set to half the cost of upgrading all substations, allowing at most two out of four substations to be upgraded; The smoothing parameter β is set to 0.5, which controls the sharpness of approximation to the thresholding function $\mathbb{1}(Y_i \ge \tau_i)$. This relaxation is used in our experiment so that the derived closed-form solution could be applied.

Table 4 Hyperparameter settings for CREDO.

Hyperparameter	Description	Default value
K	Simulation size	1×10^2
σ	Component variance scale	1
n	Calibration dataset size	1×10^1
m	Training dataset size	1×10^1

Table 5 Summary table of selected metrics used in the experiments.

Metric Name	Description	Mathematical Formula
Validity	The percentage of decisions where the risk estimate satisfies the conservativeness guarantee in (2).	$\frac{1}{\lvert\mathcal{V}\rvert} \sum_{z \in \mathcal{V}} \mathbb{1}\{\hat{\alpha}(z) \leq \alpha(z)\}$
MAE	The mean absolute error of the estimated compared to the true risks.	$\frac{1}{\lvert\mathcal{V}\rvert} \sum_{z \in \mathcal{V}} \lvert\hat{\alpha}(z) - \alpha(z)\rvert$

B.3. Baselines

In this part, we describe the baselines used in the experiments. To control for variables, the baselines share the same set of hyperparameters as CREDO, summarized in Table 4. Learning rates and their maximum epochs are set to 1×10^{-2} and 1×10^2. All optimization procedures are run for 10 epochs.

B.3.1. Baselines in Section 7.1

(*i*) *SA (sample average)*: A naive sample average using observed historic data,

$$\hat{\mathbb{P}}\{z \notin \pi(Y)\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{z \notin \pi(y_i)\}$$

Note that this estimator does not have any predictive power as it does not take x as input, and does not guarantee conservativeness.

(*ii*) *LR (logistic regression)*: Models the probability of z falling within $\pi(Y)$ via the logistic regression model:

$$\hat{\mathbb{P}}\{z \notin \pi(Y)\} = \frac{1}{1 + \exp\big(-(\theta^\top x + \theta_0)\big)},$$

where x denotes the feature vector and (θ, θ_0) are learned parameters.

(*iii*) *NN (neural network)*: Models the probability of z falling within $\pi(Y)$ via a neural network classifier: $\hat{\mathbb{P}}\{z \notin \pi(Y)\} = \sigma\big(f_\theta(x)\big)$, where $f_\theta(\cdot)$ denotes the neural network mapping parameterized by θ, and $\sigma(t) = (1 + \exp(-t))^{-1}$ is the logistic sigmoid function.

(*iv*) *QE (quantile estimator)*: The 75% empirical quantile of $\{\mathbb{1}\{z \notin \pi(y_i)\}\}_{i=1}^n$. This is equivalent to quantile regression methods when X is omitted.

(*v*) *CP (conformal prediction)*: The naive conformal prediction method (Vovk et al. 2005) applied to $\{(x_i, \mathbb{1}\{z \in \pi(y_i)\})\}_{i=1}^n$, taking the upper bound of the 25% prediction interval as the final risk estimate:

$$\hat{\mathbb{P}}\{z \notin \pi(Y)\} = \hat{\mathbb{E}}[\mathbb{1}\{z \notin \pi(Y)\} \mid X = x] + \hat{Q}_e\left(\frac{\lfloor (1 - 25\%)(n+1) \rfloor}{n}\right),$$

where $\hat{Q}_e(\cdot)$ denotes the empirical quantile of the nonconformity scores (*i.e.*, regression prediction residuals). The regression model is taken as SA (the sample average estimator).

Algorithm 3 GD (Gradient Descent) for Solving (13b) in Generla Convex Settings

Require: Prediction $\hat{y}$; Decision z; Penalty parameter λ; Learning rate η; Tolerance level ϵ.

$y^{(0)} \leftarrow \hat{y}$ initialization;
for $t \in \{1, \ldots, T\}$ **do**
  $\psi_1^{(t-1)} \leftarrow \nabla_y \min_{z' \in \mathcal{Z}} f(z'; y)\big|_{y=y^{(t-1)}}$ compute via differentiable optimization;
  $\psi_2^{(t-1)} \leftarrow \nabla_y f(z; y)\big|_{y=y^{(t-1)}}$;
  $\psi^{(t-1)} \leftarrow \frac{y^{(t-1)}-\hat{y}}{\|y^{(t-1)}-\hat{y}\|_2} + \lambda \cdot (\psi_1 - \psi_2 + \epsilon)$ if constraint is positive else $\frac{y^{(t-1)}-\hat{y}}{\|y^{(t-1)}-\hat{y}\|_2}$;
  $y^{(t)} \leftarrow y^{(t-1)} - \eta \cdot \psi^{(t-1)}$;
end for
return $\tilde{\alpha}(z) = R^{-1}(\|y^{(T)} - \hat{y}\|_2)$, if $z \in \pi_\epsilon(\hat{y})$.

B.3.2. Baselines in Section 7.2

(*i*) *PTO (predict-then-optimize)*: the standard two-stage predict-then-optimize approach (Bertsimas and Kallus 2020), which first predicts parameters and then solves the resulting optimization problem $\min_{z \in \mathcal{Z}} f(z; \hat{Y})$, where $\hat{Y}$ is a point estimate of $\mathbb{E}[Y|X]$. Note that `PTO` is equivalent to stochastic optimization, since X is omitted in our setting. $\hat{Y}$ is estimated using the sample average estimator from historic data.

(*ii*) *RO (robust optimization)*: a minmax optimization problem $\min_{z \in \mathcal{Z}} \max_{y \in \mathcal{U}} f(z; y)$, where the uncertainty set $\mathcal{U}$ is constructed using naive conformal prediction under the ℓ_∞ norm;

(*iii*) *SPO+ (smart predict-then-optimize)*: a predict-then-optimize method where the prediction model is trained using a surrogate loss that is convex and explicit (Elmachtoub and Grigas 2022);

(*iv*) *DFL (decision-focused learning)*: prediction models are trained by directly optimizing the downstream objective through end-to-end differentiation of the optimization layer (Agrawal et al. 2019, Amos and Kolter 2017).

The baselines are selected based on the following criteria: (*i*) The ability to produce a decision that maximizes a (linear) objective function; and (*ii*) The capacity to handle randomness in the underlying optimization problem.

B.3.3. Baselines in Section 7.3

We describe the baselines used for comparing the optimization procedures:

(*i*) *GD (gradient descent)*: A gradient descent-based algorithm that approximates (18) by minimizing its Lagrangian function, where its gradient is:

$$\nabla_y \mathcal{L}(y, \lambda) = \frac{y - \hat{y}}{\|y - \hat{y}\|_2} + \lambda \cdot \nabla_y \left[\min_{z' \in \mathcal{Z}} f(z'; y) - f(z; y) + \epsilon \right]^+ .$$

Here, λ is the regularization hyperparameter, which is set to 1×10^3. The term $\nabla_y \min_{z' \in \mathcal{Z}} f(z'; y)$ is computed through differentiable optimization layers (Agrawal et al. 2019, Amos and Kolter 2017). The pseudocode of GD is provided in Algorithm 3.

(*ii*) *BF (brute force)*: A heuristic algorithm that first approximates the constraint space and decision space with gridded sets:

$$\tilde{\mathcal{Y}} \approx \{y \in \mathcal{Y} \mid f(z; y) - f(z'; y) > \epsilon\} \quad \text{and} \quad \tilde{\mathcal{Z}} \approx \mathcal{Z},$$

then solves the finite optimization problem $\min_{y \in \tilde{\mathcal{Y}}} \min_{z' \in \tilde{\mathcal{Z}}} \|y - \hat{y}\|_2$.

We note that *RS (random search)* is similar to BF except that the gridded sets are constructed by drawing T samples from a standard Gaussian distribution over the corresponding spaces.

Algorithm 4 Empirical Confidence Ranking

Require: True policy $\pi^* : \mathcal{Y} \to \mathcal{Z}$; Candidate policy $\pi : \mathcal{X} \to \mathcal{Z}$; Test dataset $\{(x_i, y_i)\}_{i=1}^{\ell}$.

$\mathcal{Z}^* = \emptyset$
for $i = \{1, \ldots, \ell\}$ **do**
  $z_i^* \leftarrow \pi^*(y_i)$.
  $\mathcal{Z}^* \leftarrow \mathcal{Z}^* \cup \{z_i^*\}$.
end for
for $z \in \mathcal{V}(\theta)$ **do**
  $h(z) \leftarrow$ the total occurence of decision $| \{i \mid z_i^* = z\} |$.
end for
for $i = \{1, \ldots, \ell\}$ **do**
  $\hat{z}_i \leftarrow \pi(x_i)$.
  $\text{rank}_i \leftarrow$ ranking of estimated decision $| \{z' \in \mathcal{V}(\theta) \mid h(z') \geq h(\hat{z}_i)\} |$
end for
return Empirical confidence ranking $1/\ell \cdot \sum_{i=1}^{\ell} \text{rank}_i$.

B.4. Metrics

Table 5 summarizes three evaluation metrics used in our experiments. The reporting standards for the metrics are their means ± standard deviations across $T = 100$ repeated trials.

Algorithm 4 outlines the generic procedure for computing empirical confidence rankings. In the synthetic setting, this procedure is repeated over $T = 1 \times 10^2$ independent trials with $\ell = 1 \times 10^3$ test data points; In the real-world setting, the evaluation is performed rolling over $T = 12$ periods, spanning from 2010 to 2022. Each trial uses a two-year window (24 months) of data, which is sequentially split into training, calibration, and testing sets in an $8 : 8 : 8$ ratio.

B.5. Additional Experiment Results

We augment the third set of experiments in Section 7.3 with Table 7, which is the full version of Table 3 with standard deviation included. The interpretation of the table is the same as its main-text version.

Table 6 Evaluated metrics for different optimization procedures solving (13b)**.**

	LP Setting I			LP Setting II			QP		
	Obj	Vio	Err	Obj	Vio	Err	Obj	Vio	Err
GD	0.44 ± 0.64	0.40 ± 0.20	0.45 ± 0.51	0.06 ± 0.12	0.07 ± 0.06	0.05 ± 0.11	0.60 ± 0.43	0.23 ± 0.26	0.50 ± 0.33
BF	0.66 ± 0.42	$\mathbf{0.00 \pm 0.00}$	0.43 ± 0.37	0.11 ± 0.14	$\mathbf{0.00 \pm 0.00}$	0.10 ± 0.12	0.52 ± 0.30	$\mathbf{0.00 \pm 0.00}$	0.34 ± 0.20
RS	0.89 ± 0.77	$\mathbf{0.00 \pm 0.00}$	0.66 ± 0.72	0.12 ± 0.15	$\mathbf{0.00 \pm 0.00}$	0.11 ± 0.14	0.70 ± 0.50	$\mathbf{0.00 \pm 0.00}$	0.52 ± 0.42
RG	5.21 ± 2.97	0.20 ± 0.22	4.98 ± 2.89	1.35 ± 1.61	0.17 ± 0.15	1.34 ± 1.60	4.81 ± 3.04	0.17 ± 0.17	4.63 ± 2.96
CREDO	$\mathbf{0.23 \pm 0.14}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.01 \pm 0.01}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.10 \pm 0.09}$	0.10 ± 0.15	$\mathbf{0.08 \pm 0.05}$

	SOCP			IP		
	Obj	Vio	Err	Obj	Vio	Err
GD	1.53 ± 0.92	0.13 ± 0.16	1.31 ± 0.81	0.22 ± 0.27	$\mathbf{0.00 \pm 0.00}$	0.19 ± 0.24
BF	0.66 ± 0.42	$\mathbf{0.00 \pm 0.00}$	0.44 ± 0.29	0.11 ± 0.14	$\mathbf{0.00 \pm 0.00}$	0.09 ± 0.10
RS	0.91 ± 0.77	$\mathbf{0.00 \pm 0.00}$	0.69 ± 0.64	0.12 ± 0.15	$\mathbf{0.00 \pm 0.00}$	0.10 ± 0.12
RG	5.21 ± 2.97	0.17 ± 0.17	4.99 ± 2.88	1.35 ± 1.61	$\mathbf{0.00 \pm 0.00}$	1.32 ± 1.57
CREDO	$\mathbf{0.14 \pm 0.12}$	0.03 ± 0.10	$\mathbf{0.08 \pm 0.05}$	$\mathbf{0.00 \pm 0.01}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.02 \pm 0.03}$