Carnegie Mellon University

Link to paper

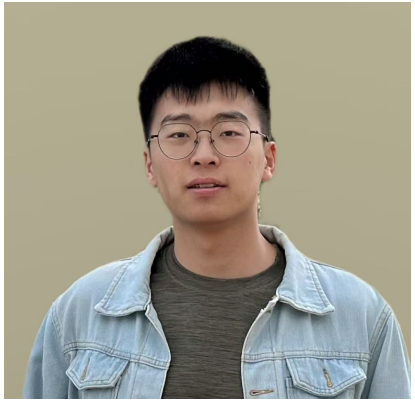
Conformalized Decision Risk Assessment

Presenter: Wenbin Zhou

Caltech RSRG/FALCON Weekly Seminar

Shorter version: ICLR 2026; Long version: Under review

Co-authors

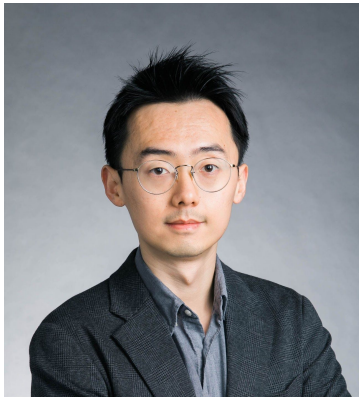


Wenbin Zhou

CMU

Agni Orfanoudaki

University of Oxford



Shixiang (Woody) Zhu

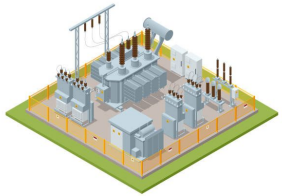
CMU

Carnegie Mellon University

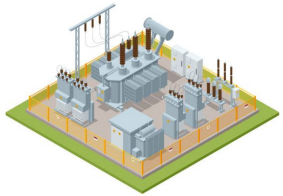
Introduction: predict-then-optimize

Ex. power grid infrastructure upgrade:

Incoming future demand →



Substation 1



Substation 2

How should the budget be allocated between the two substations?

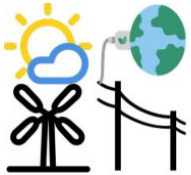
Operator

Carnegie Mellon University

Introduction: predict-then-optimize

Ex. power grid infrastructure upgrade:

Prediction model $g : \mathcal{X} \rightarrow \mathcal{Y}$



Features

(e.g., climate, load)

Scenario

(e.g., substation-level demand ~ return per unit-budget allocated)

Carnegie Mellon University

Introduction: predict-then-optimize

Ex. power grid infrastructure upgrade:

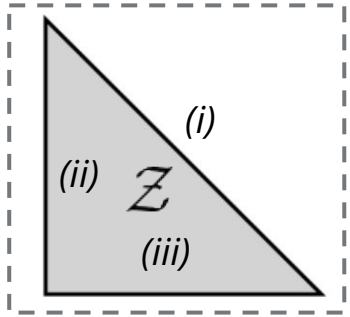
Maximize total return

$$\max_{z \in \mathcal{Z}} (\hat{y}_1 z_1 + \hat{y}_2 z_2)$$

feasible region

Budget constraint

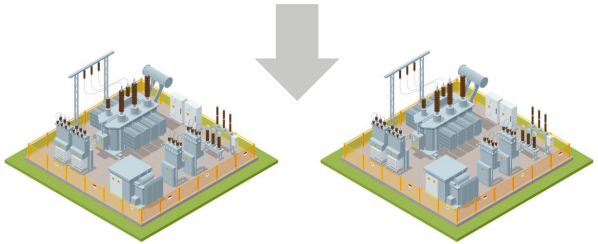
$$\mathcal{Z} := \begin{cases} z_1 + z_2 \leq 1 & (i) \\ z_1 \geq 0 & (ii) \\ z_2 \geq 0 & (iii) \end{cases}$$



Introduction: predict-then-optimize

Ex. power grid infrastructure upgrade:

Budget allocation plan $z = (z_1, z_2)$



Substation 1 *Substation 2*

Operator

Carnegie Mellon University

Introduction: decision-making under uncertainty

- In the literature, lots of **prescriptive** frameworks have been proposed (predict-then-optimize, robust optimization, decision-focused learning, etc.)
- Yet, in many cases, human usually don't want to just be **passive decision-executors!**

Operator

I'd prefer having the final decision made by **me**, not the **algorithm**.

→ *Algorithm distrust*; *Liability issue*; *...* → *"Algorithm aversion"*

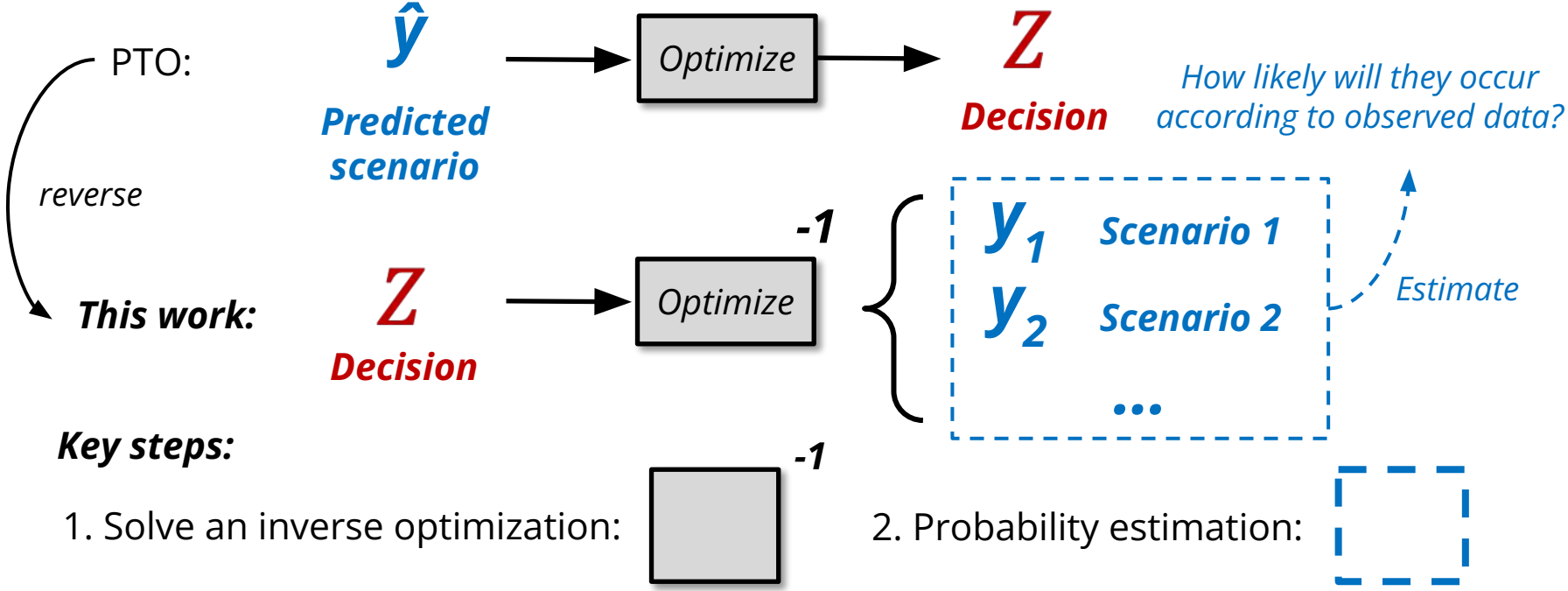
...I'd appreciate if algorithms can help with **assessing** my decisions.

→ *No direct intervention*; *Communicate **uncertainty*** ✓

***Our setup**:* Let human decide first, then assess how likely this decision is optimal.

Overview of the proposed framework

***Our setup**:* Let human decide first, then assess how likely this decision is optimal.



Problem setup

Definition: decision policy

Define **decision policy π** as a constrained optimization problem:

$$\pi(y) := \arg\min_{z \in \mathcal{Z}} f(z; y)$$

Variables

z : decision

y : scenario

x : features

Three technical notes:

- For simplicity, we assume the feasible region ***Z*** does not dependent on ***y***
- Objective function ***f*** can be convex, linear, …
- Also naturally generalize to ***ε***-optimal policy

*z is optimal up to some **ε**-tolerance*

$$\pi_\epsilon(y) := \left\{ z \in \mathcal{Z} : f(z; y) \leq \min_{z' \in \mathcal{Z}} f(z'; y) + \epsilon \right\}$$

Carnegie Mellon University

Problem setup

Objective: decision risk assessment

Given ***z***, estimate the following probability:

$$\mathbb{P}\{z \in \pi(Y)\}$$

Randomness is from the scenario random variable Y

*"How likely is **z** the optimal decision given random realization of **Y**?"*

Ex. power grid infrastructure upgrade:

PTO → Budget allocation plan ***z*** → *Operator*

Operator → Budget allocation plan ***z*** → How likely is this the best plan?

Decision Risk Assessment

Presented to

(Complementary)

Problem setup

Objective: decision risk assessment

Given ***z***, estimate the following probability:

$$\mathbb{P}\{z \in \pi(Y)\}$$

*Can we trust the estimate with **limited data**?*

- Overestimate ✗
- Underestimate ✓

→ *"With at least **XX**% probability my decision is optimal"*

Objective: decision risk assessment

Given z, estimate a certified **decision risk *α(z)*** *from data, such that:*

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

← *Target of estimation*

Algorithm (step 1): inverse optimization

Definition: inverse feasible region (Chan et. al, 2021)

Given a **decision *z***, its inverse feasible region w.r.t. **policy *π*** is defined as:

$$\pi^{-1}(z) = \left\{ y \in \mathcal{Y} : f(z;y) = \min_{z' \in \mathcal{Z}} f(z';y) \right\}$$

*All scenarios **y** that would make **z** optimal.*

Note: the inverse feasible region w.r.t. ***ε***-optimal policy's can be similarly defined:

$$\pi_\epsilon^{-1}(y) := \left\{ z \in \mathcal{Z} : f(z;y) \leq \min_{z' \in \mathcal{Z}} f(z';y) + \epsilon \right\}$$

Y → π → Z (Scenario → Decision)

Z → π^{-1} → Y (Decision → Scenario)

Carnegie Mellon University

Algorithm (step 1): inverse optimization

Lemma: objective reformulation

$$\mathbb{P}\{z \in \pi(Y)\} \equiv \mathbb{P}\{Y \in \pi^{-1}(z)\}$$

How likely is ***z*** *optimal given* ***π(Y)****?* *How likely would* ***Y*** *fall in* $\boldsymbol{\pi^{-1}(Y)}$*?*

Insight: **Deterministic** variable ∈ **random** set. ⇒ **Random** variable ∈ **deterministic** set.

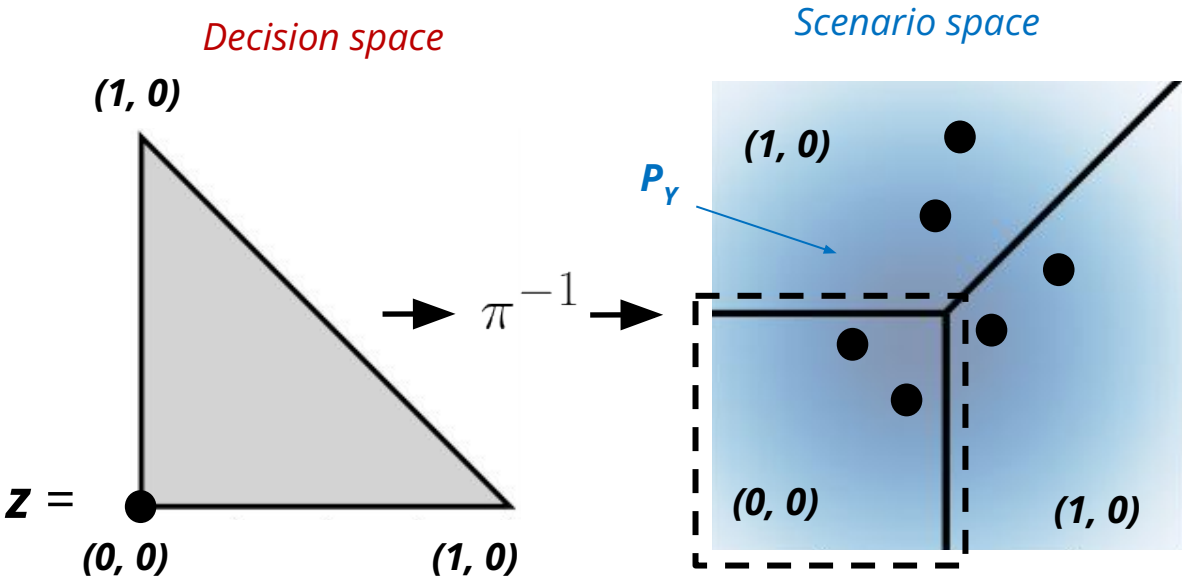
We have many ML & statistics tools for the latter

Let's setup a running example (linear program), which will be used throughout the presentation

Carnegie Mellon University

Algorithm (step 1): inverse optimization

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



Monte-Carlo estimation?

= 3/6 ≈ 33% ✗

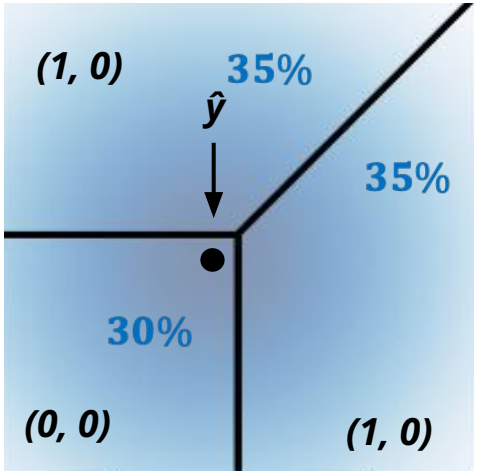
- No access to ground truth P_Y
- Doesn't satisfy **lower bound**

Step 2 (In the next few slides)

Algorithm (step 1): inverse optimization

Revisiting PTO: Suppose we use the prediction model $g(x) = \hat{\mathbb{E}}[Y|X = x]$

Scenario space



However:

PTO: ⟷ Find ***z*** such that $\hat{y} \in \pi^{-1}(z)$

↓

z = (0, 0) is optimal

Decision risk assessment (minimize risk): → ***z = (1, 0)*** *and* ***(0, 1)*** are equally optimal.

Note*: similar counterexamples can be constructed for SA, RO, …*

Insight: PTO may *fail* to identify decision that are most likely to be optimal

Algorithm (step 2)

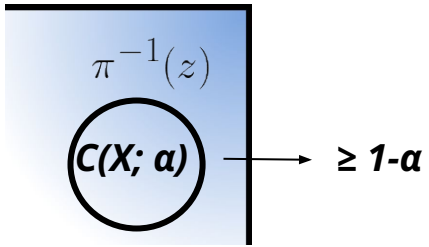
Idea: To satisfy the lower bound guarantee, we use a **surrogate set *C(X; α)* :**

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

where:

(i) holds if ***C(X; α)*** is contained by $\boldsymbol{\pi^{-1}(z)}$

(ii) holds if ***C(X; α)*** is some valid ***1-α*** prediction set of ***Y***



Note: Miscoverage rate ***α*** is dependent on both ***z*** and distribution of ***Y*** due to the two conditions.

***Q**: How to construct **C(X; α)**? How to find **α**?*

Carnegie Mellon University

Algorithm (step 2): conformalization

High-level idea:

- Construct **conformal prediction set** for ***Y***.
- Tune its miscoverage level ***α*** to make it just big enough to be contained within $\boldsymbol{\pi^{-1}(z)}$.
- Take ***α*** as the desired estimate.

Assumption: exchangeability (Vovk 2005)

We assume access to a set of *n* calibration data $\{(X_i, Y_i)\}_{i=1}^{n}$, such that the joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ is invariant under permutation.

Note: *Standard assumption in CP, weaker than i.i.d. → less restrictive*

Carnegie Mellon University

Algorithm (step 2): conformalization

Conformal prediction procedure

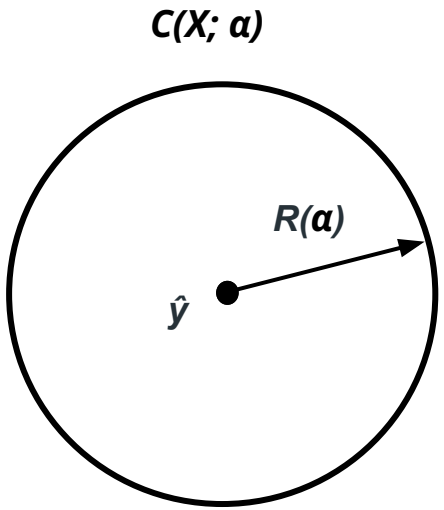
Given pre-trained prediction model ***g: X → Y***.

Using $\{(X_i, Y_i)\}_{i=1}^{n}$, define nonconformity scores**:**

$$L_i = \|Y_i - \hat{Y}_i\|_2 \qquad \forall i = 1, \ldots, n.$$

Given some ***α > 0***, define the conformalized radius ***R(α)*** the (adjusted) ***1-α***-empirical quantile of L_i. defined prediction set:

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



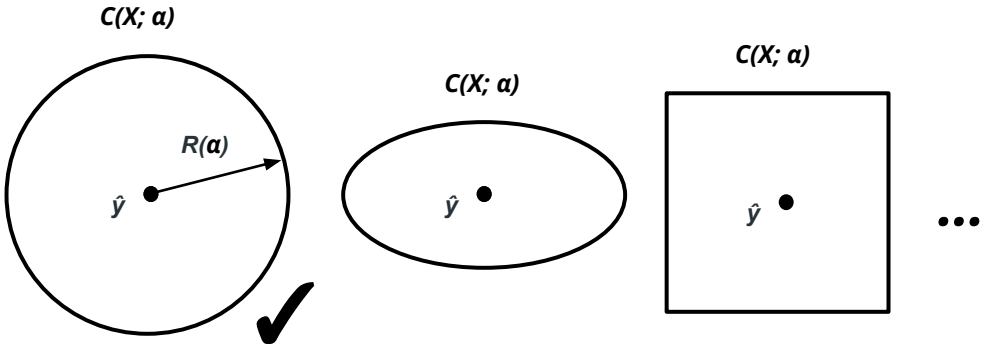
Takeaway: For *any given* miscoverage rate ***α***, we can use the calibration data to construct a ***1-α*** *valid prediction set on* ***Y.*** → *Inequality (ii)*

Carnegie Mellon University

Algorithm (step 2): conformalization

Two additional clarifications:

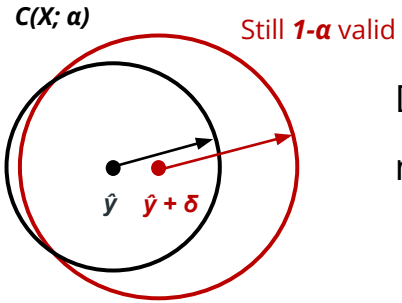
First,



We only consider balls:

- Simple geometry
- Easier computation (later slides)
- Shape choice is less concerning with using gen. models (later slides)

Second, the prediction model ***g*** does not have to be "accurate" for validity of prediction set to hold



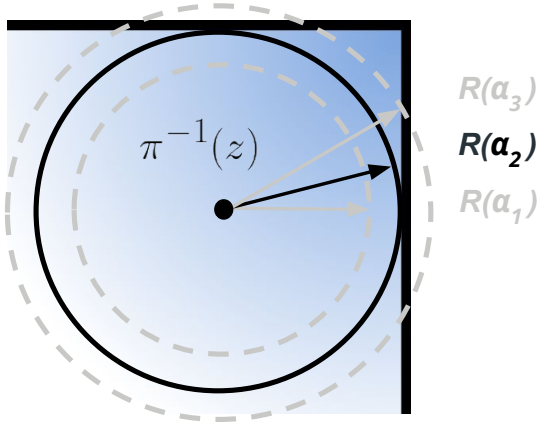
Distribution-free (adaptive) nature of conformal prediction

Carnegie Mellon University

Algorithm (step 2): conformalization

The miscoverage rate ***α*** is selected *post-hoc* such that ***C(X; α):***

- Achieves maximum possible size
- Subject to the fact that it must be contained within $\pi^{-1}(z)$.



Optimization problem: miscoverage rate selection

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

Then, according to the chain equality, use ***E[α̂(z)]*** as the certified decision risk!

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

But, is that the end of the story?

Carnegie Mellon University

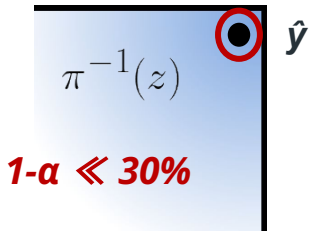
Issue 1: over-conservatism

Generally, we hope that $C(X; \alpha)$ could cover as much space within $\pi^{-1}(z)$ as possible

However, this might be unrealistic in some cases, since prediction $\hat{y}$ is independent of decision z

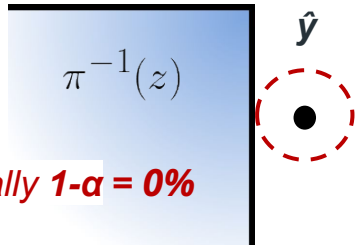
(Boundary-case)

(1) if $\hat{y}$ falls near the boundary of $\pi^{-1}(z)$



(Human-algorithm misalignment)

(2) if $\hat{y} \notin \pi^{-1}(z)$



In both cases, the algorithm outputs very conservative risk estimates!

Can we tighten this?

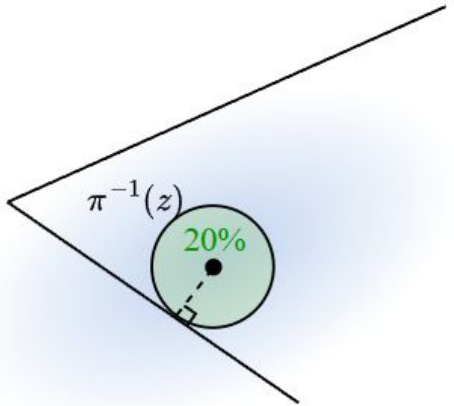
Carnegie Mellon University

Fix 1: generative model

***Idea**:* Create "diverse, repeated trials" of **ŷ**, so the percentage of unfavorable scenarios is diluted.

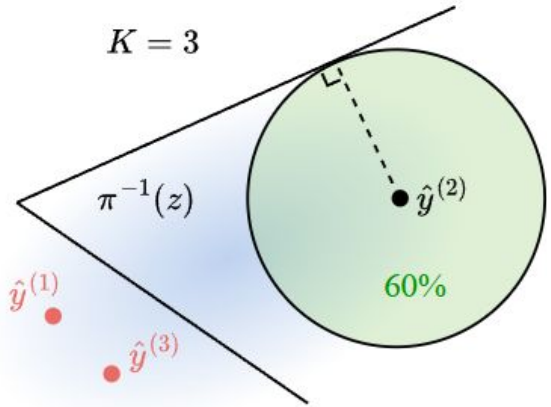
- Specify ***g: X → Y*** as a (conditional) generative model instead of a point prediction model.
- Average over ***K*** risk estimates as the final risk estimation

Point prediction model

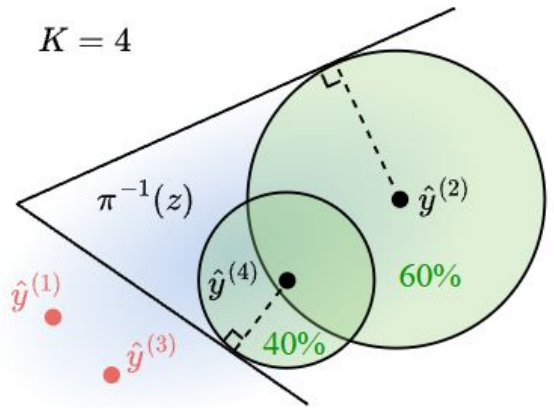


20%

Generative model



$(0\% + 0\% + 60\%) / 3 = 20\%$

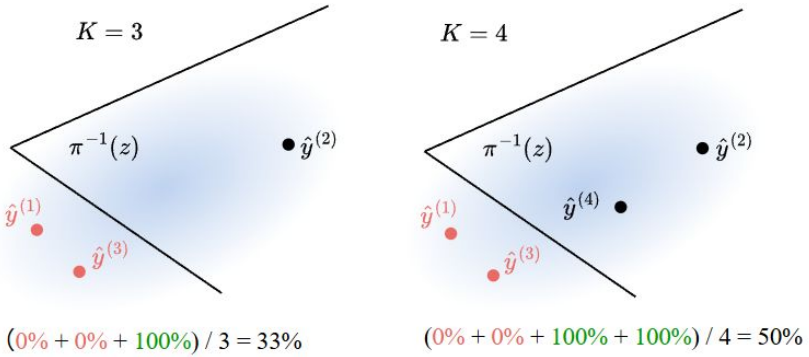


$(0\% + 0\% + 60\% + 40\%) / 4 = 25\%\uparrow$

Fix 1: alternative interpretation

Considering estimating the probability $\mathbb{P}\{Y \in \pi^{-1}(z)\}$

Monte-Carlo



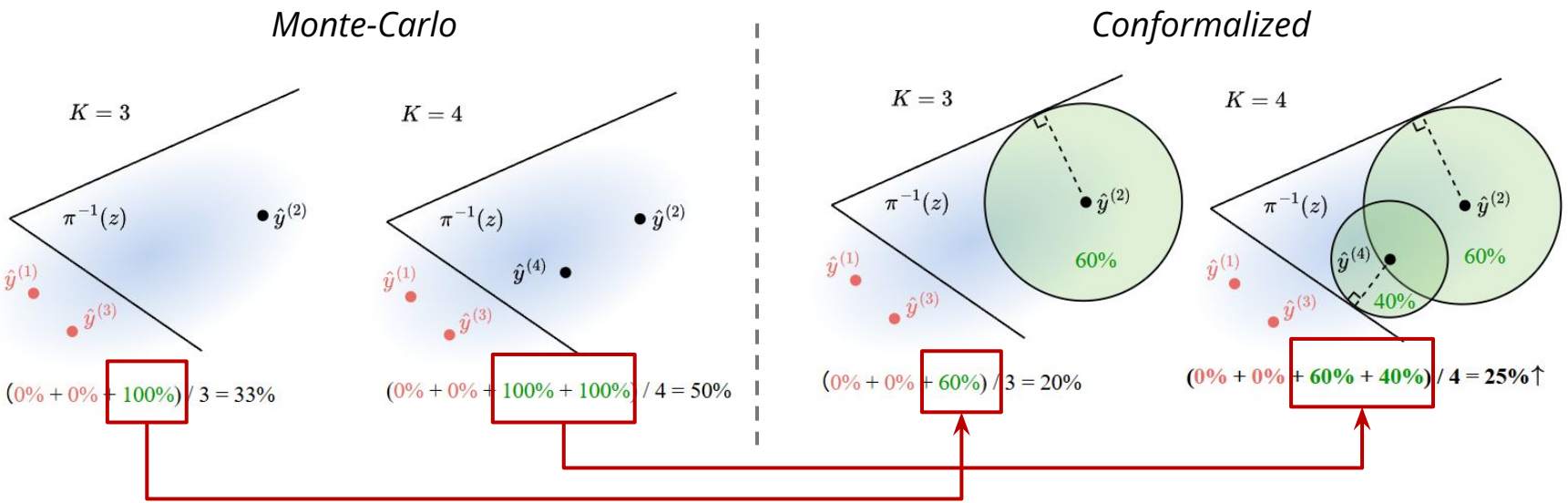
Procedure:

- Draw ***K*** repeated samples of ***Y*** from the generative model
- Compute the percentage of trials that the samples fell in $\boldsymbol{\pi^{-1}(z)}$

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

Carnegie Mellon University

Fix 1: alternative interpretation



***Pattern**: Monte Carlo discounted by the conformalized coverage rate*

Carnegie Mellon University

Fix 1: alternative interpretation

Lemma: weighted Monte Carlo estimator form

The estimator ***α̂(z)*** averaged over ***K*** generative model samples can be equivalently written as

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

Conformalized weight ***∈ [0, 1]*** *that is dependent on the calibration data*

Insight:

- The generative version of ***α̂(z)*** can be interpreted as a **weighted Monte Carlo estimator** for probability evaluation.
- The weights discount the estimate to satisfy conservatism

Footnote: $w^{(k)}(z) = 1-\min_{\alpha}\{\alpha \in [0,1] : C^{(k)}(x;\alpha) \subseteq \pi^{-1}(z)\}$

Carnegie Mellon University

Issue 2: degraded post-hoc validity

Corollary

There exists an error term **Δ(P) > 0**, such that

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

The error term **Δ(P)** arises because ***α̂(z)*** is dependent on the calibration data, which disrupts the **exchangeability** assumption → over-confidence estimate

Optimization problem: miscoverage rate selection

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

"Constructing hypothesis with data, and testing it with the same data again"

Footnote: $\Delta(P) = \frac{\sum_{i=1}^{n} d_{\mathrm{TV}}^{(i)}(\hat{\alpha}(z))}{n+1}$

Carnegie Mellon University

Fix 2: conformal-*e*-prediction

Replace (standard) conformal prediction with **conformal *e*-value prediction (Vovk 2020)**

$$R_p(\alpha) = \hat{Q}\left(\frac{\lfloor (n+1)(1-\alpha) \rfloor}{n}\right) \Rightarrow R_e(\alpha) = \frac{\sum_{i=1}^{n} L_i}{\alpha(n+1) - 1}$$

"Convert testing to betting, probability → expectation"

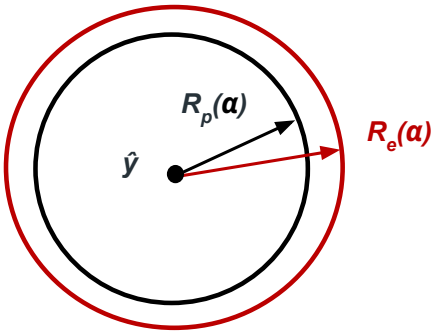
Theorem: exact post-hoc validity with e-value

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

Practical note:

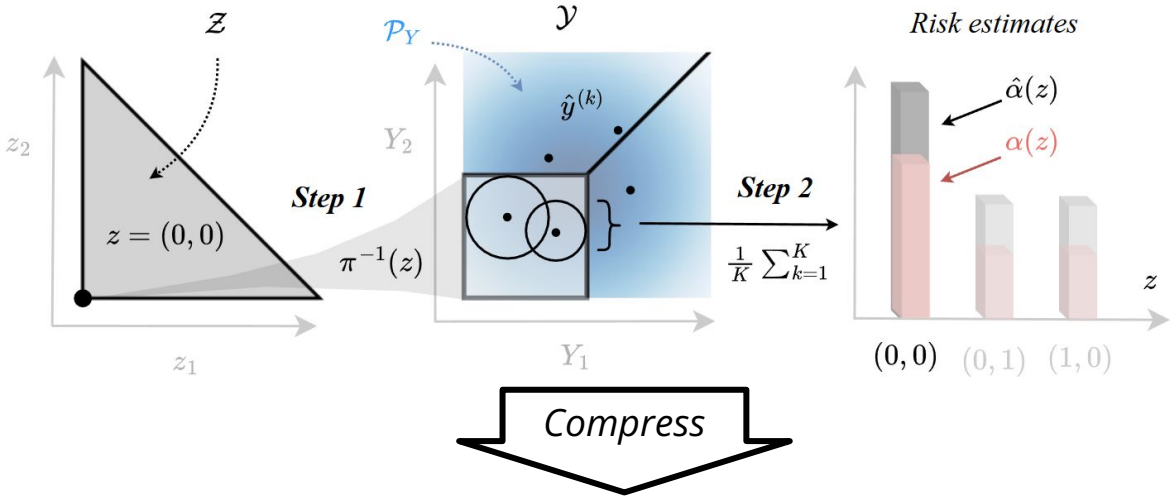
- **Validity**: E-value > P-value
- **Tightness**: P-value > E-value

Should be carefully traded-off in practice!



Computation

Geometrically:



Computationally:

Theorem: the end-to-end optimization problem

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

Insight: Don't have to explicitly build $\boldsymbol{\pi^{-1}(z)}$ to check containment → avoid the most inefficient part!

Computation

Specifically, for ***LP*** problems, we derived its **closed-form solution:**

Corollary: closed-form estimator for linear programs

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

Computation complexity

$$O(K \cdot n \cdot |\mathcal{V}|)$$

***Notes*:**

- ***V*** is the set of extreme points (i.e., vertices) of the feasible region, which is **finite** for LP problems
- ***max*** only searches for maximum value over a discrete set.

For ***convex*** problems, where we assume:

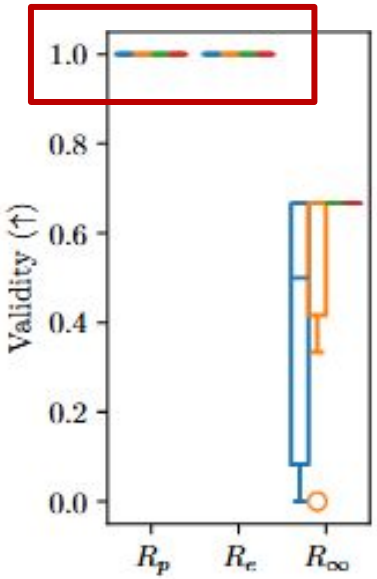
(i) ***f(z; y)*** is a bi-convex function; *(ii)* feasible region ***Z*** is convex

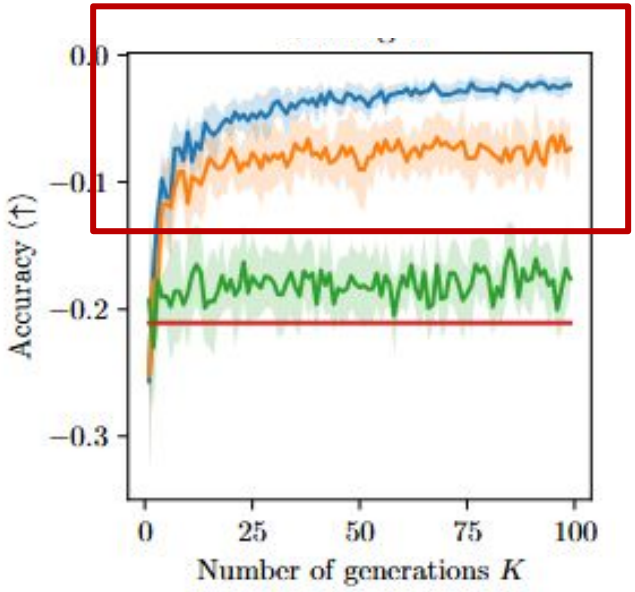
We developed an *alternating optimization* algorithm to compute ***â(z).***

Carnegie Mellon University

Experiment: risk estimation quality ablation

Achieves ~100% validity → ***conservative***





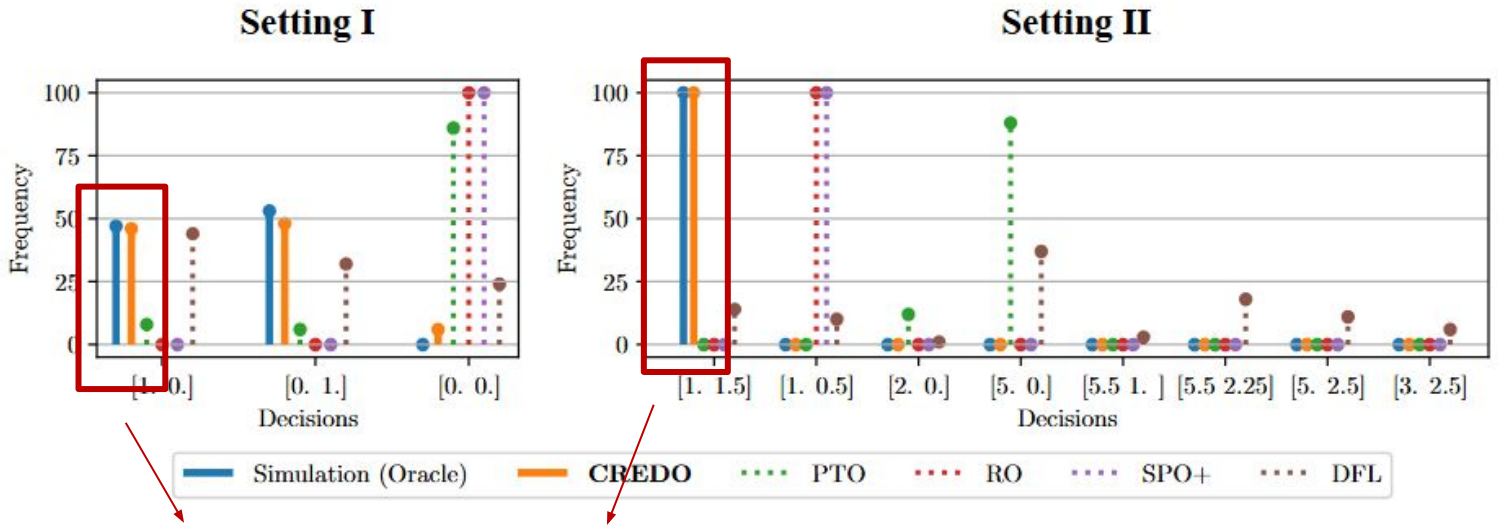
Approximation error converges to 1 with more generations → ***consistent***

Figure: Evaluation results. Different colors represent different ablation variants of the method

Takeaways: (1) the proposed framework produce risk estimate that is both **valid** and **accurate**; (2) generative design helps model reduce **over-conservatism**

Carnegie Mellon University

Experiment: decision prescription analysis



The oracle model (simulated optimal decisions) and our method align

Takeaways: Our method can also be used for decision prescription for choosing the decisions with smallest risk.

Carnegie Mellon University

Conclusion

- Motivated by challenges decision-making under certainty, we study the novel problem of decision risk assessment
- Proposed a conformalized framework for providing a certified risk estimate for the decision risk
- Developed statistical theories and conducted numerical experiments to validate the approach.

E-mail: *wenbinz2@andrew.cmu.edu*

Link to this paper

Link to my homepage

Thank you!

Carnegie Mellon University

Appendix

Experiments: optimization algorithm evaluation

Optimization evaluation metrics

Other optimization methods

able 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**

Takeaway: Replacing point prediction model with generative model helps improve reduce over conservatism

Carnegie Mellon University

Experiments: risk estimation evaluation

Optimization settings

Metrics

Prob. estimation baselines

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	0.05 ± 0.03	0.53 ± 0.43	0.03 ± 0.02	0.60 ± 0.48	0.05 ± 0.03	0.53 ± 0.49	0.05 ± 0.03	0.47 ± 0.47	0.05 ± 0.04

Takeaway: CREDO consistently produce risk estimates that satisfies conservativeness (high validity) while not over conservative (low MAE).

Carnegie Mellon University

Experiments: decision prescription evaluation

Metric: empirical confidence ranking

Decision-prescription baselines

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	**1.00 ± 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	**1.00 ± 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+	**1.00 ± 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	**1.61 ± 0.56**	**1.48 ± 0.52**	1.05 ± 0.22	**1.00 ± 0.00**	2.03 ± 0.96	**1.75 ± 0.92**
CREDO (5-GMM)	1.89 ± 0.87	1.65 ± 0.62	1.54 ± 0.52	1.03 ± 0.17	**1.00 ± 0.00**	1.92 ± 0.89	1.92 ± 1.04
CREDO (VAE)	1.01 ± 0.10	1.61 ± 0.58	1.77 ± 0.71	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.06 ± 0.24**	1.92 ± 1.04

Takeaway: CREDO consistently prescribe decisions that have higher confidence ranking (i.e., more likely to be optimal) compared to baselines

Carnegie Mellon University