Distributionally Robust Reinforcement Learning and Optimization

Siqi Yao

SCHOOL OF DATA SCIENCE

October 20, 2025

Table of Contents

Table of Contents

Existing DR-RL Algorithms

- Model deployment environment may differ from data collection environment ⇒ **Distributionally Robust Reinforcement Learning.**
- Trade-off: Still need provable guarantees on sample complexity.
- Table 1 shows the worst-case sample complexity of model-based distributionally RL algorithms, with ambiguity set based on KL divergence with ambiguity size δ.

Algorithm	Sample Complexity	Origin
DRVI	$\widetilde{O}(\lvert\mathbf{S}\rvert^2\lvert\mathbf{A}\rvert e^{O(1-\gamma)^{-1}}(1-\gamma)^{-4}\varepsilon^{-2}\delta^{-2})$	[Zhou et al., 2021]
REVI/DRVI	$\widetilde{O}(\lvert\mathbf{S}\rvert^2\lvert\mathbf{A}\rvert e^{O(1-\gamma)^{-1}}(1-\gamma)^{-4}\varepsilon^{-2}\delta^{-2})$	[Panaganti and Kalathil, 2022]
DRVI	$\widetilde{O}(\lvert\mathbf{S}\rvert^2\lvert\mathbf{A}\rvert(1-\gamma)^{-4}\varepsilon^{-2}\mathfrak{p}_\wedge^{-2}\delta^{-2})$	[Yang et al., 2022]
DRVI-LCB	$\widetilde{O}(\lvert\mathbf{S}\rvert\lvert\mathbf{A}\rvert(1-\gamma)^{-4}\varepsilon^{-2}\mathfrak{p}_\wedge^{-1}\delta^{-2})$	[Shi and Chi, 2024]

Table 1: Sample complexity upper bounds for finding an ε-optimal robust policy in model-based distributionally robust RL.

Drawbacks and Objectives

- $\widetilde{O}(\delta^{-2})$ dependence: When $\delta \downarrow 0$, uncertainty vanishes but $\delta^{-2} \downarrow \infty$. The robust value function should converge to the non-robust optimal cumulative reward as $\delta \downarrow 0$.
- Model-based methods are computationally intensive, require more memory to store MDP models, and often do not generalize well to non-tabular RL settings.
- Q-Learning is model-free but not robust.
- **Objective:** Design a variant of Q-Learning that is robust & the sample complexity scales correctly with δ.

Table of Contents

Recap on RL

- For policy $\pi \in \Pi$, the value function is defined as:

$$v^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \,\middle|\, S_0 = s\right] \tag{1}$$

- The optimal value function is:

$$v^*(s) := \max_{\pi \in \Pi} v^{\pi}(s) \tag{2}$$

  which satisfies the Bellman Optimality Equation:

$$v^*(s) = \max_{a \in \mathbf{A}} \left(\mathbb{E}_{\nu_{s,a}}[R] + \gamma \mathbb{E}_{p_{s,a}}[v^*(S)]\right) \tag{3}$$

  where $R \sim \nu_{s,a}, S \sim p_{s,a}$.
- Optimal q-function:

$$q^*(s,a) := \mathbb{E}_{\nu_{s,a}}[R] + \gamma \mathbb{E}_{p_{s,a}}[v^*(S)] \tag{4}$$

  which satisfies:

$$q^*(s,a) = \mathbb{E}_{\nu_{s,a}}[R] + \gamma \mathbb{E}_{p_{s,a}}\left[\max_{b \in \mathbf{A}} q^*(S,b)\right] \tag{5}$$

- Optimal policy: $\pi^*(s) = \arg\max_{a \in \mathbf{A}} q^*(s,a) \Rightarrow$ Need to estimate q^*.

KL Divergence Constrained DR-RL

- For each $(s, a) \in \mathbf{S} \times \mathbf{A}$ and $\delta > 0$, define KL ambiguity set centered at $p_{s,a}$ and $\nu_{s,a}$ of radius δ by

$$\mathcal{P}_{s,a}(\delta) := \{p : D_{\mathrm{KL}}(p \,\|\, p_{s,a}) \leq \delta\}, \tag{6}$$

$$\mathcal{N}_{s,a}(\delta) := \{\nu : D_{\mathrm{KL}}(\nu \,\|\, \nu_{s,a}) \leq \delta\}. \tag{7}$$

- The DR Bellman operator $\mathcal{B}_\delta$ for value function is defined as

$$\mathcal{B}_\delta(v)(s) := \max_{a \in \mathbf{A}} \inf_{\substack{p \in \mathcal{P}_{s,a}(\delta) \\ \nu \in \mathcal{N}_{s,a}(\delta)}} (\mathbb{E}_\nu[R] + \gamma \mathbb{E}_p[v(S)]) \tag{8}$$

This is where robustness appears!

- The DR optimal value function v_δ^* is defined as

$$v_\delta^* = \mathcal{B}_\delta(v_\delta^*) \tag{9}$$

Dual Formulation

Lemma 1

Let X be a random variable and μ_0 be a probability measure on $(\Omega, \mathcal{F})$ s.t. X has a finite MGF in a neighborhood of zero. Then for any $\delta > 0$,

$$\inf_{\mu: D_{\mathrm{KL}}(\mu \| \mu_0) \leq \delta} \mathbb{E}_\mu X = \sup_{\alpha \geq 0} \left\{ -\alpha \log \mathbb{E}_{\mu_0} \left[e^{-X/\alpha} \right] - \alpha\delta \right\}. \tag{10}$$

Apply lemma to (9), we obtain:

$$v^*(s) = \max_{a \in \mathbf{A}} \left\{ \sup_{\alpha \geq 0} \left\{ -\alpha \log \mathbb{E}_{\nu_{s,a}} \left[e^{-R/\alpha} \right] - \alpha\delta \right\} + \gamma \sup_{\beta \geq 0} \left\{ -\beta \log \mathbb{E}_{p_{s,a}} \left[e^{-v^*(S)/\beta} \right] - \beta\delta \right\} \right\} \tag{11}$$

q-Function

- The optimal DR q-function is defined as

$$q^*(s,a) := \inf_{\substack{p\in\mathcal{P}_{s,a}(\delta)\\ \nu\in\mathcal{N}_{s,a}(\delta)}} \left(\mathbb{E}_\nu[R] + \gamma\mathbb{E}_p\left[v^*(S)\right]\right) \tag{12}$$

  where v^* is the DR optimal value function.
- The DR Bellman operator is defined as

$$\mathcal{T}(q)(s,a) := \inf_{\substack{p\in\mathcal{P}_{s,a}(\delta)\\ \nu\in\mathcal{N}_{s,a}(\delta)}} \left(\mathbb{E}_\nu[R] + \gamma\mathbb{E}_p\left[v(q)(S)\right]\right) \tag{13}$$

  where $v(q)(s) := \max_{b\in\mathbf{A}} q(s,b)$. The dual form is given by:

$$\begin{aligned}\mathcal{T}(q)(s,a) = &\sup_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}_{\nu_{s,a}}\left[e^{-R/\alpha}\right] - \alpha\delta\right\}\\ &+\gamma\sup_{\beta\geq 0}\left\{-\beta\log\mathbb{E}_{p_{s,a}}\left[e^{-v(q)(S)/\beta}\right] - \beta\delta\right\}\end{aligned} \tag{14}$$

- The following Bellman equation holds:

$$q^* = \mathcal{T}(q^*) \tag{15}$$

  and the optimal policy: $\pi^*(s) = \arg\max_{a\in\mathbf{A}} q^*(s,a) \Rightarrow$ Need to estimate q^*.

Table of Contents

Setting

- Denote $\nu_{s,a,n}$ and $p_{s,a,n}$ the estimate of $\mu_{s,a}$ and $p_{s,a}$ formed by n i.i.d. samples, i.e. for $f : \mathbf{U} \to \mathbb{R}$, where $\mathbf{U}$ could be $\mathbf{S}$ or $\mathbf{R}$,

$$\mathbb{E}_{\mu_{s,a,n}} f(U) := \frac{1}{n} \sum_{i=1}^{n} f(U_i) \tag{16}$$

  for $\mu = \nu, p$ and $U_i = R_i, S_i$ are i.i.d. across i.
- Define the empirical DR Bellman operator on n i.i.d. samples by

$$\begin{aligned} \mathbf{T}(q)(s,a) := & \sup_{\alpha \geq 0} \left\{ -\alpha \log \mathbb{E}_{\nu_{s,a,n}} \left[e^{-R/\alpha} \right] - \alpha\delta \right\} \\ & + \gamma \sup_{\beta \geq 0} \left\{ -\beta \log \mathbb{E}_{p_{s,a,n}} \left[e^{-v(q)(S)/\beta} \right] - \beta\delta \right\} \end{aligned} \tag{17}$$

  which is equivalent to (14) where the sets $\mathcal{P}_{s,a}(\delta), \mathcal{N}_{s,a}(\delta)$ are replaced with $\{p : D_{\mathrm{KL}}(p \,\|\, p_{s,a,n}) \leq \delta\}, \{\nu : D_{\mathrm{KL}}(\nu \,\|\, \nu_{s,a,n}) \leq \delta\}$.
- Biased estimator: $\mathbb{E}[\mathbf{T}(q)] \neq \mathcal{T}(q)$ for generic q.

Distributionally Robust Q-learning Algorithm

Algorithm 1 Distributionally Robust Q-Learning

Input: the total times of iteration k_0 and a batch size n_0.
Initialization: $q_1 \equiv 0$; $k = 1$.
for $1 \leq k \leq k_0$ **do**
Sample $\mathbf{T}_{k+1}$ the n_0-sample empirical DR Bellman operator as in Definition 5.
Compute the Q-learning update

$$q_{k+1} = (1 - \lambda_k) q_k + \lambda_k \mathbf{T}_{k+1}(q_k) \tag{3.4}$$

with stepsize $\lambda_k = 1/(1 + (1 - \gamma)k)$.
end for
return q_{k_0+1}.

Figure 1: Distributionally Robust Q-learning Algorithm. Source: [Wang et al., 2024]

Compare with classical Q-learning update:

$$q_{k+1}(s, a) = (1 - \lambda_k) q_k(s, a) + \lambda_k \left(R_{k+1} + \gamma \, v(q_k)(S_{k+1}) \right) \tag{18}$$

Key Assumptions

- Define the minimum support probability as

$$\mathfrak{p}_\wedge := \min_{s,a \in \mathbf{S}\times\mathbf{A}} \min \left\{ \min_{r \in \mathbf{R}: \nu_{s,a}(r)>0} \nu_{s,a}(r), \min_{s' \in \mathbf{S}: p_{s,a}(s')>0} p_{s,a}(s') \right\} \tag{19}$$

- **Assumption 1 (Limited Adversarial Power):** Suppose the adversary's power δ satisfies $\delta < \frac{1}{24}\mathfrak{p}_\wedge$.
- **Assumption 2 (Reward Bound):** Reward $\mathbf{R} \subset [0, 1]$.

Sample Complexity Upper Bound

Theorem 1

Denote $d := |\mathbf{S}||\mathbf{A}|\,(|\mathbf{S}| \vee |\mathbf{R}|)$. Assume Assumptions 1 and 2. Then the DR Q-learning Algorithm with parameters $k_0 = c_0 \frac{1}{(1-\gamma)^3\varepsilon} \log\left(\frac{3d}{(1-\gamma)\eta\varepsilon}\right)^2$ and $n_0 = c_0 \frac{1}{\mathfrak{p}_\wedge^3 (1-\gamma)^2 \varepsilon} \log\left(\frac{3dk_0}{\eta}\right)^2$ computes a solution q_{k_0+1} s.t. $\|q_{k_0+1} - q^*\|_\infty \leq \varepsilon$ w.p. at least $1-\eta$ using

$$\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}||\mathbf{A}|}{\mathfrak{p}_\wedge^3 (1-\gamma)^5 \varepsilon^2}\right) \tag{20}$$

number of samples.

This bound completely removes the dependence on δ!

Variance-Reduced Distributionally Robust Q-learning Algorithm

Intuition: Rather than spending a lot at each step for rough estimates, it's better to pay once for a solid reference point and then fine-tune cheaply around it.

Algorithm 2 Variance-Reduced Distributionally Robust Q-Learning

Input: the number of epochs l_{vr}, a sequence of recentering sample size $\{m_l\}_{l=1}^{l_{\text{vr}}}$, an epoch length k_{vr} and a batch size n_{vr}.
Initialization: $\hat{q}_0 \equiv 0$; $l = 1$; $k = 1$.
for $1 \le l \le l_{\text{vr}}$ **do**
  Compute $\widetilde{\mathbf{T}}_l$, m_l-sample empirical DR Bellman operator as in Definition 5.
  Set $q_{l,1} = \hat{q}_{l-1}$.
  for $1 \le k \le k_{\text{vr}}$ **do**
    Sample $\mathbf{T}_{l,k+1}$ an n_{vr}-sample empirical Bellman operator.
    Compute the recentered Q-learning update

$$q_{l,k+1} = (1-\lambda_k) q_{l,k} + \lambda_k \left(\mathbf{T}_{l,k+1}(q_{l,k}) - \mathbf{T}_{l,k+1}(\hat{q}_{l-1}) + \widetilde{\mathbf{T}}_l(\hat{q}_{l-1}) \right) \tag{3.5}$$

    with stepsize $\lambda_k = 1/(1+(1-\gamma)k)$.
  end for
  Set $\hat{q}_l = q_{l,k_{\text{vr}}+1}$.
end for
return $\hat{q}_{l_{\text{vr}}}$

Figure 2: Variance-Reduced DR Q-learning Algorithm. Source: [Wang et al., 2024]

Sample Complexity Upper Bound

Theorem 2

Assume Assumptions 1 and 2. For $\varepsilon < (1-\gamma)^{-1}$, the variance-reduced DR Q-learning Algorithm with specific parameters computes a solution $\hat{q}_{l_{\mathrm{vr}}}$ such that $\|\hat{q}_{l_{\mathrm{vr}}} - q^*\|_\infty \leq \varepsilon$ w.p. at least $1-\eta$ using

$$\widetilde{\mathcal{O}}\left(\frac{|\mathbf{S}|\,|\mathbf{A}|}{\mathfrak{p}_\wedge^3 (1-\gamma)^4 \min(1,\varepsilon^2)}\right) \tag{21}$$

number of samples.

This is superior to the bound (20) in terms of $1-\gamma$.

Table of Contents

Analyzing the Title

- **Prescriptive Analytics:** Not to determine "what happened" (descriptive) or "what will happen" (predictive), but rather "what we should do."
- **Adherence-aware:** Core problem of the paper. When making decisions, model must consider in advance: "Will this person adhere to my advice?" **Example:**
  - Algorithm: Path A is shortest in distance, recommend it.
  - Mailman: Path A is bumpy, choose path B instead.
  - Consequence: Mailman deviated from the recommended route, causing the platform's ETA to be inaccurate, resulting in negative customer experience.

  Model's objective is no longer finding the "objectively optimal solution," but rather the one with the "highest compliance", while ensuring that its "objective quality" remains reasonably good.

Analyzing the Title (cont'd)

- **Inverse Optimization:**
  - Forward optimization: Find the optimal path given costs to all paths.
  - Inverse optimization: Given the path mailman has taken, infer his perceived cost for each path.
  - Example:
    - Observed 1000 times, 90% mailman choose path A.
    - Standard IO yields a **point estimate:** "In mailman's perception, overall cost of path A (e.g., $5 \times$ time $+ 10 \times$ fatigue) must be high."
    - **Drawback:** Point estimate does not represent the population.
- **Conformal Prediction:** A statistical method that outputs an uncertainty set, instead of a single prediction value.

Table of Contents

Setting (Shop Storage Management Example)

- **Decision Maker:** Shop's manager.
- **Background Parameters $\mathbf{u}_k$:** Headquarters' algorithm-suggested list of products to be removed from shelves (e.g.: A, B).
- **Decision $\hat{\mathbf{x}}_k$:** Manager's actual decision: Removed A but kept B.
- **Ground-truth Parameter $\boldsymbol{\theta}^*$:** The actual sales data and profitability of the products across all shops.
- **Perceived Parameter $\hat{\boldsymbol{\theta}}_k$:** Manager's perceived value of the products, as he may possess local information unknown to the algorithm.
- **Goal of IO:** By observing which products the manager has kept, infer the "local value" he perceives for those products.
- **Final Goal:** Improve algorithm to consider both $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}_k$.

Given dataset $\mathcal{D} := \{(\hat{\mathbf{x}}_k, \mathbf{u}_k) \,|\, k \in [N]\}$, seek policy $\bar{\mathbf{x}} : \mathcal{U} \to \mathbb{R}^n$ s.t. $\bar{\mathbf{x}}(\mathbf{u})$ is of high quality w.r.t. $\boldsymbol{\theta}^*$ & future $\hat{\boldsymbol{\theta}}$.

Forward Optimization Problem

- Consider the forward optimization problem:

$$\mathbf{FOP}(\boldsymbol{\theta}, \mathbf{u}) : \min_{\mathbf{x} \in \mathcal{X}(\mathbf{u})} f(\boldsymbol{\theta}, \mathbf{x}) \tag{22}$$

- Decision maker obtains decision $\hat{\mathbf{x}}$ by solving $\mathbf{FOP}(\hat{\boldsymbol{\theta}}, \mathbf{u})$ where $\hat{\boldsymbol{\theta}}$ is his perception of the ground-truth parameter $\boldsymbol{\theta}^*$, which is unknown to him.
- Denote $\widetilde{\mathbf{x}} : \Theta \times \mathcal{U} \to \mathbb{R}^n$ the optimal solution to **FOP:** $\widetilde{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{u}) \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}) := \arg\min_{\mathbf{x} \in \mathcal{X}(\mathbf{u})} \{f(\boldsymbol{\theta}, \mathbf{x})\}$
- Assume f is linear in θ:

$$f(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i \in [d]} \theta_i f_i(\mathbf{x}) \tag{23}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ for all $i \in [d]$ are some known continuous basis functions. This results in scale-invariance.

Evaluation Metrics

- **Actual Optimality Gap (AOG)** of a decision policy $\bar{\mathbf{x}}$ is defined as

$$\mathrm{AOG}(\bar{\mathbf{x}}) := \mathbb{E}_{\mathbf{u}}\left[f(\boldsymbol{\theta}^*, \bar{\mathbf{x}}(\mathbf{u})) - \min_{\mathbf{x}\in\mathcal{X}(\mathbf{u})} f(\boldsymbol{\theta}^*, \mathbf{x})\right] \tag{24}$$

- **Perceived Optimality Gap (POG)** of a decision policy $\bar{\mathbf{x}}$ is defined as

$$\mathrm{POG}(\bar{\mathbf{x}}) := \mathbb{E}_{\hat{\boldsymbol{\theta}},\mathbf{u}}\left[f(\hat{\boldsymbol{\theta}}, \bar{\mathbf{x}}(\mathbf{u})) - \min_{\mathbf{x}\in\mathcal{X}(\mathbf{u})} f(\hat{\boldsymbol{\theta}}, \mathbf{x})\right] \tag{25}$$

- **Goal:** Design $\bar{\mathbf{x}}$ that achieves low AOG and POG simultaneously, so the recommendations generated are of high quality and likely to be implemented.

Necessary Assumptions

- **Assumption 3:** Dataset $\mathcal{D}$ is generated using $\hat{\mathbf{x}}_k := \widetilde{\mathbf{x}}(\hat{\boldsymbol{\theta}}_k, \mathbf{u}_k)$, where $(\hat{\boldsymbol{\theta}}_k, \mathbf{u}_k)$ are i.i.d. samples from $\mathbb{P}_{(\hat{\boldsymbol{\theta}}, \mathbf{u})}$ for all $k \in [N]$.
- **Assumption 4:** There exists a constant $\sigma \in [0, 2]$ such that $\|\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^*\|_2 \leq \sigma$.
- Define $\eta \in \mathbb{R}_+$ to be a constant such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \leq \eta$ for any $\mathbf{u} \in \mathcal{U}$, $\hat{\mathbf{x}} \in \mathcal{X}(\mathbf{u})$, and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta^{\mathrm{OPT}}(\hat{\mathbf{x}}, \mathbf{u})$, where $\Theta^{\mathrm{OPT}}(\mathbf{x}, \mathbf{u}) := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; \mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}),\ \|\boldsymbol{\theta}\|_2 = 1 \right\}$.

Standard Inverse Optimization Pipeline

First use IO to obtain a point estimate $\bar{\boldsymbol{\theta}}$ of the unknown parameters and then employ a policy $\bar{\mathbf{x}}_{\mathrm{IO}}(\mathbf{u}) := \widetilde{\mathbf{x}}(\bar{\boldsymbol{\theta}}, \mathbf{u})$, i.e., solving **FOP** with the estimated $\bar{\boldsymbol{\theta}}$ and $\mathbf{u}$.

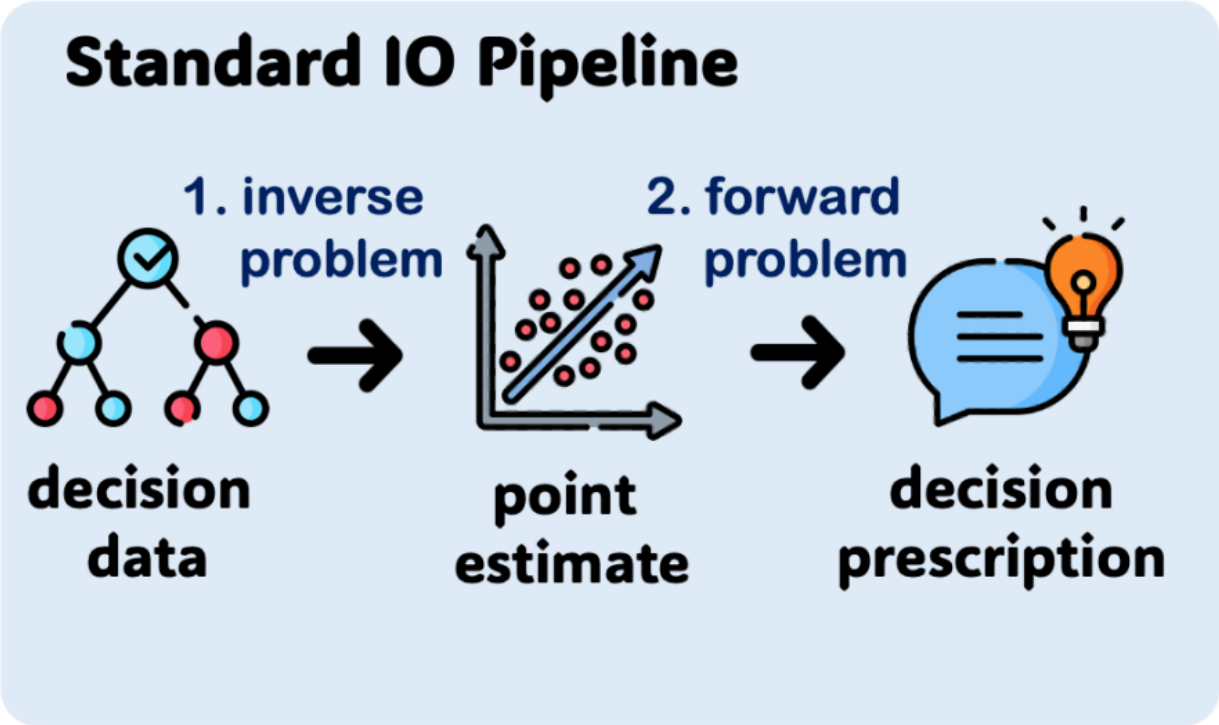

Figure 3: Standard IO Pipeline. Source: [Chan et al., 2024]

Standard Inverse Optimization Pipeline (cont'd)

- Given $\mathcal{D} := \{(\hat{\mathbf{x}}_k, \mathbf{u}_k) \mid k \in [N]\}$, estimate the parameters by solving the following **inverse optimization problem**:

$$\mathbf{IOP}(\mathcal{D}) : \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{k \in [N]} \ell\big(\hat{\mathbf{x}}_k, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}_k)\big)\,. \tag{26}$$

  where ℓ is a non-negative loss function that returns 0 only when $\hat{\mathbf{x}}_k \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u}_k)$

- E.g., the **sub-optimality loss** of $\boldsymbol{\theta}$ is given by

$$\ell_{\mathrm{S}}\big(\hat{\mathbf{x}}, \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})\big) = f(\boldsymbol{\theta}, \hat{\mathbf{x}}) - \min_{\mathbf{x} \in \mathcal{X}^{\mathrm{OPT}}(\boldsymbol{\theta}, \mathbf{u})} f(\boldsymbol{\theta}, \mathbf{x}). \tag{27}$$

  However, the paper demonstrated that this loss function can lead to decision policies with arbitrarily large AOG and POG.

Incorporating Robustness

- **Robust** forward optimization problem:

$$\mathbf{RFOP}\big(\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha), \mathbf{u}\big) : \min_{\mathbf{x} \in \mathcal{X}(\mathbf{u})} \max_{\boldsymbol{\theta} \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)} f(\boldsymbol{\theta}, \mathbf{x}). \tag{28}$$

  where the uncertainty set $\mathcal{C}$ is defined as:

$$\mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha) := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \|\boldsymbol{\theta}\|_2 = 1,\, \boldsymbol{\theta}^\top \bar{\boldsymbol{\theta}} \geq \cos \alpha \right\}. \tag{29}$$

  which contains all the unit vectors in $\mathbb{R}^d$ that are within angle α of the point estimate $\bar{\boldsymbol{\theta}}$.
- Only care about direction of $\boldsymbol{\theta}$, not magnitude.

Value of Robustness

Theorem 3

Given a coverage level $\gamma \in [0,1]$ and a point estimate $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$, choose the uncertain angle α_γ such that

$$\mathbb{P}_{(\hat{\boldsymbol{\theta}},\mathbf{u})}\big(\mathcal{C}(\bar{\boldsymbol{\theta}},\alpha_\gamma) \cap \Theta^{\mathrm{OPT}}(\hat{\mathbf{x}},\mathbf{u}) \neq \emptyset\big) \geq \gamma \tag{30}$$

for a random sample $(\hat{\boldsymbol{\theta}},\mathbf{u})$ from $\mathbb{P}_{(\hat{\boldsymbol{\theta}},\mathbf{u})}$ and $\hat{\mathbf{x}} = \tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}},\mathbf{u})$. For any $\mathbf{u}' \in \mathcal{U}$, let $\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u}';\alpha_\gamma)$ be an optimal solution to $\mathrm{RFOP}(\mathcal{C}(\bar{\boldsymbol{\theta}},\alpha_\gamma),\mathbf{u}')$. If Assumption 4 holds, then

$$\mathrm{POG}(\bar{\mathbf{x}}_{\mathrm{CIO}}) \leq \varepsilon_{\mathrm{POG}}(\gamma) := (2+\eta-2\gamma\cos 2\alpha_\gamma)\hat{\mu} + (2+\eta\gamma-2\gamma)\mu_{\mathrm{CIO}}(\alpha_\gamma), \tag{31}$$

$$\mathrm{AOG}(\bar{\mathbf{x}}_{\mathrm{CIO}}) \leq \varepsilon_{\mathrm{AOG}}(\gamma) := (2+\eta-2\gamma\cos 2\alpha_\gamma+\sigma)\mu^* + (2+\eta\gamma-2\gamma+\sigma)\mu_{\mathrm{CIO}}(\alpha_\gamma), \tag{32}$$

where $\hat{\mu} := \mathbb{E}_{(\hat{\boldsymbol{\theta}},\mathbf{u})}\big(\nu[\tilde{\mathbf{x}}(\hat{\boldsymbol{\theta}},\mathbf{u})]\big)$, $\quad \mu_{\mathrm{CIO}}(\alpha_\gamma) := \mathbb{E}_{\mathbf{u}}\big(\nu[\bar{\mathbf{x}}_{\mathrm{CIO}}(\mathbf{u};\alpha_\gamma)]\big)$, and $\mu^* := \mathbb{E}_{\mathbf{u}}\big(\nu[\tilde{\mathbf{x}}(\boldsymbol{\theta}^*,\mathbf{u})]\big)$.

This theorem haracterizes the performance of our policy in terms of AOG and POG when the uncertainty set used for decision prescription can "rationalize" an unseen, future decision with some probability, but **how to achieve this?**.

Table of Contents

Complete Procedure

1. **Data Split:** Split data into $\mathcal{D}_{\text{train}}$ & $\mathcal{D}_{\text{val}}$.
2. **Point Estimation:** Solve **IOP**$(\mathcal{D}_{\text{train}})$ with sub-optimality loss to obtain point estimate $\bar{\boldsymbol{\theta}}$.
3. **Uncertainty Set Construction:** Given $\bar{\boldsymbol{\theta}}$, construct uncertainty set using $\mathcal{D}_{\text{val}}$ s.t. with a specified probability, contains parameters that rationalize the next unseen decision.
4. **Final Decision:** Solve **RFOP** to prescribe new decisions.

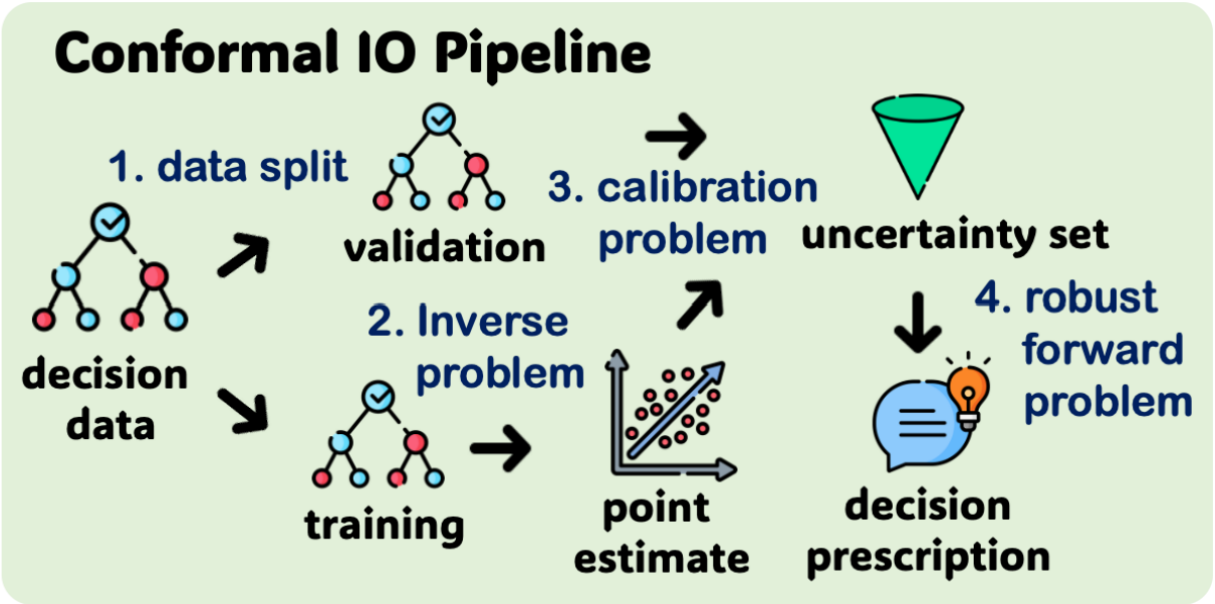

Figure 4: Conformal IO Pipeline. Source: [Chan et al., 2024]

Uncertainty Set Construction/Calibration

Learn the smallest uncertainty set that achieves the desired probability. The uncertainty set calibration problem is

$$
\begin{aligned}
\mathbf{CP}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{val}}, \gamma): \quad & \min_{\alpha, \{\boldsymbol{\theta_k}\}_{k \in \mathcal{K}_{\text{val}}}} \quad \alpha \\
& \text{subject to} \quad \hat{\mathbf{x}}_k \in \mathcal{X}^{\text{OPT}}(\boldsymbol{\theta_k}, \mathbf{u}_k), \quad \forall\, k \in \mathcal{K}_{\text{val}} \\
& \qquad \sum_{k \in \mathcal{K}_{\text{val}}} \mathbb{1}[\boldsymbol{\theta}_k \in \mathcal{C}(\bar{\boldsymbol{\theta}}, \alpha)] \geq \gamma(N_{\text{val}} + 1) \\
& \qquad \|\boldsymbol{\theta}_k\|_2 = 1, \quad \forall\, k \in \mathcal{K}_{\text{val}} \\
& \qquad 0 < \alpha \leq \pi
\end{aligned}
$$

References I


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


Panaganti, K. and Kalathil, D. (2022).
Sample complexity of robust reinforcement learning with a generative model.
In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR.


Shi, L. and Chi, Y. (2024).
Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity.
Journal of Machine Learning Research, 25(200):1–91.

References II


Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2024).
Sample complexity of variance-reduced distributionally robust q-learning.
Journal of Machine Learning Research, 25(341):1–77.


Yang, W., Zhang, L., and Zhang, Z. (2022).
Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics.
The Annals of Statistics, 50(6):3223–3248.


Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021).
Finite-sample regret bound for distributionally robust offline tabular reinforcement learning.
In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR.

Thank you!
Any questions?