SPO Extensions

Hong-Seok Choe

1 Introduction

- Downstream optimization task depends on uncertain parameter $\xi \in \mathbb{R}^m$
- Decision maker observes contextual feature vector $x \in \mathbb{R}^p$
- Goal:
$$\min_{w \in S(\xi)} \mathbb{E}_{\xi \sim \mathcal{D}_x} [c(w, \xi) \mid x]$$
- $\mathcal{D}_x$: conditional distribution $\xi \mid x$
- $c(\cdot, \xi) : \mathbb{R}^d \to \mathbb{R}$: objective function
- $S(\xi) \subseteq \mathbb{R}^d$: constraint set
- Access to historical data $\{(x_i, \xi_i)\}_{i=1}^n$

2 Related Work

Decision-Driven Regularization: A Blended Model for Predict-then-Optimize (Loki et al.)

- Setting

  Predict z from context feature x and optimize $\min_{y \in \mathcal{Y}} c(y, \hat{z})$. Real relationship $z = g(x)$ estimated by class of function $f(x, w)$ parameterized by weights w
- Key Idea: Balances two possible cost formulations
  - $c(y; \tilde{z})$ (empirical cost)
    Uses observed outcomes $\tilde{z}$
  - $c(y; f(\tilde{x}, w))$ (estimated cost)
    Uses estimated outcomes given weight w $\hat{z} := f(\tilde{x}, w)$
- Objective Function:
$$\arg\min_w L(w) - \lambda \frac{1}{n} \sum_{n=1}^{n} \min_{y_n \in \mathcal{Y}} \{\mu c(y_n; \tilde{z_n}) + (1 - \mu) c(y_n; f(\tilde{x_n}, w))\}$$
  Captures multiple frameworks (SPO, SP, JERO)
- Other Keywords: contextual optimization, empirical optimization, regret minimization
- Key Papers listed (to-see):
  - Joint Estimation and Robustness Optimization (Zhu et al.)
  - From data to decisions: Distributionally robust optimization is optimal (Van Parys et al.)
  - A general framework for optimal data-driven optimization (Sutter et al.)

- Small-data, large-scale linear optimization with uncertain objectives (Gupta & Rusmevichientong)
- Small-data, large-scale linear optimization with uncertain objectives (Gupta & Rusmevichientong)
- Machine Learning with Operational Costs (Tulabandhula and Rudin)

* RO Papers
  - D. Bertsimas, D. B. Brown, and C. Caramanis, Theory and applications of robust optimization
  - B. L. Gorissen et al., A practical guide to robust optimization
  - V. Gabrel, C. Murat, and A. Thiele, Recent advances in robust optimization: An Overview

Ideas in DRO: construct an empirical distribution of data points and consider an ambiguity set under divergence measure or Wasserstein metric

3 Predict-then-optimize with Robustness

In predict-then-optimize, cost vector c and feasible region S may both depend on unknown parameter ξ where feasibility can be violated with constraint predictions. Is there a way to address feasibility along with uncertainty using a different loss function?

SPO streamlined (Predict) – (Optimize). Here, when the constraint set $S(\xi)$ is uncertain, we want to streamline

$$\text{(Predict)} - \text{(Check Feasiblilty)} - \text{(Optimize)} \tag{1}$$
$$\text{(Predict)} - \text{(Robust Optimize)} \tag{2}$$

* If (1) is concerned, once $\hat{S}(x)$ is predicted, we can somehow obtain feasible set $\tilde{S}(x)$ based on the prediction and do the downstream optimization.
* If (2) is concerned, we have predicted constraints $\hat{S}(x)$, and can construct an uncertainty set using the predicted nominal constraints. Here, by constructing a proper uncertainty set, feasibility issue can be resolved?

Motivating example:

* (Maximum flow with capacity prediction) Consider a graph with edge set E and let the uncertain capacity for any arc $e \in E$ be, say, $c_e = \bar{c}_e(1 + \rho\xi_e)$, where $\bar{c}_e$ denotes the predicted nominal value and $\rho > 0$ is a robustness parameter. The uncertain vector ξ lies in the uncertainty set $\mathcal{U}(x)$ that depends on the given feature vector x (and/or realized observation of ξ's)
* We may construct the uncertainty set in the way it achieves feasibility by adjusting ρ. Note that this is more like predict-then-RO (i.e., (2)).

Possible Formulations

* Non-random ξ (Robust optimization formulation):

$$\begin{aligned} \min_{w} \quad & \sup_{\xi \in \mathcal{U}(x)} c(w, \xi) \\ \text{s.t.} \quad & w \in S(\xi) \qquad \forall \xi \in \mathcal{U}(x) \end{aligned}$$

  Here, the uncertainty set depends on the feature vector x (e.g., $\xi \in [u_1(x), u_2(x)]$ interval endpoints may depend on x)

* Random ξ (Contextual stochastic optimization formulation)

$$\min_{w \in S(\xi)} \mathbb{E}_{\xi \sim \mathcal{D}_x} \left[c(w, \xi) \mid x \right]$$

  Here, $\xi \approx f(x)$ and we want to learn f (similar assumptions as in supervised learning problems (x_i, ξ_i)'s are i.i.d. sampled from some underlying distribution $P(X, \Xi)$)

- Random ξ (DRO Formulation?)

$$\begin{aligned} \min_{w} \quad & \sup_{P \in \mathcal{P}} \mathbb{E}_{\xi \sim P} \left[c(w, \xi) \right] \\ \text{s.t.} \quad & w \in S(\xi) \end{aligned} \tag{*}$$

where $\mathcal{P}$ is the family of probability measures supported on the domain of ξ. In empirical setting, $\mathcal{P}$ is typically chosen as a ball of radius ϵ around the empirical sample $\hat{\mathbb{P}}_n$: $\mathcal{P} = \{\mathbb{Q} : d(\mathbb{Q}, \mathbb{P}_n) \leq \epsilon\}$ (common choices of d include ϕ-divergence and Wasserstein distance)

Q. Equivalent (or "safe") way of dealing with (*)

$$\mathbb{E}_P[g(w, \xi) \leq 0] \iff w \in S(\xi)$$

With $S(\xi)$ being polytope (i.e., with linear constraints), there seems to be an equivalent representation

Problems:

1. Predicting constraints $\hat{S}(x)$: How to estimate the true feasible region $S(\xi)$ given feature vector x

   For now, suppose that there exists an oracle that returns $\hat{S}(x)$ given a feature vector x. Indeed, $\hat{S}$ need not be feasible; should come up with another proxy $\tilde{S}(x)$ that best preserves the predicted values.

   - Brute-force remove potentially irrelavant constraints w.r.t loss function
   - Enlarge/scale the feasible region (e.g., $b + \Delta$ with $\Delta > 0$ for an RHS vector in linear constraints) w.r.t loss function
   - Probabilistic constraints and chance-constrained set approximation methods (See Nemirovski & Shapiro)

2. Designing uncertainty set $\mathcal{U}(x)$: should depend on the data $\{(x_i, \xi_i)\}_{i=1}^n$

   Zhu et al. (2020) proposes joint estimation and robustness optimization (JERO) where they use the gap between the prediction loss $L(\cdot)$ and the minimum value of loss function:

$$\mathcal{U}(\rho) := \{L(\xi) - \min_{\xi} L(\xi) \leq \rho\}$$

   Also see paper "Optimization under Decision-Dependent Uncertainty"

3. Designing corresponding loss function: What is the corresponding "regret"

   Should capture the disappointment between estimated cost from the predicted feasible set and the optimal cost from the true feasible set

   Minimax regret formulation: $\sup_{\xi}\{f(x, \xi) - OPT(\xi)\}$

$$\ell(\tilde{S}, S) := \ell(\tilde{S}(x), S(\xi)) = \sup_{\xi \in \mathcal{U}(x)} \left\{ \min_{w \in \tilde{S}(x)} c(w, \xi) - \min_{w \in S(\xi)} c(w, \xi) \right\}$$

   ERM principle w.r.t the loss above loss yields the following optimization problem for selecting $\tilde{S}$:

$$\min_{\tilde{S} \in \mathcal{S}} \frac{1}{N} \sum_{i=1}^{N} \ell(\tilde{S}(x_i), S(\xi_i))$$

   where $\mathcal{S}$ being a set of functions that gives feasible set. Obtaining $\mathcal{S}$ is non-trivial and highly depends on the structure of the problem. Indeed, a tractable $\mathcal{U}$ should also be given to solve the ERM problem.

Hereafter consider linear objective with linear constraints

$$\min_{w \in S(\xi)} \mathbb{E}_{\xi \sim \mathcal{D}_x} \left[c(\xi)^{\top} w \mid x \right]$$

Discrete non-random case with with uncertainty set $\mathcal{U} = \{[A_1; b_1], \ldots, [A_k; b_k]\}$

Things to be addressed

- Solid, sensible mathematical formulation (how should we even write down the problem with linear objective and linear constraints)
- Sanity check with toy example (especially with the proposed loss function)
- Tradeoff between feasibility and quality of solution

Let (P) be our downstream optimization problem:

$$\begin{aligned} \min \quad & -x \\ \text{s.t.} \quad & -w \leq 0 \\ & w \leq b \end{aligned} \tag{P}$$

Note that we have a nonnegativity constraint in our toy example which is intrinsic to the problem (denote it S_0, the "ground" constraint given) We can think of it as a flow preservation constraint at each node in the max flow problem when we are trying to predict the flow constraint values.

Setting: Observe feature vector $x \in \mathcal{X}$ and the real b is defined by the true mapping $f^*(x) : \mathcal{X} \to \mathbb{R}$.

Suppose further that $f^*(x) = 2^{-x}$ with $x \geq 0$ and we have data $(x_1, b_1) = (0, 1)$ and $(x_2, b_2) = (1, 0.5)$. In predict-then-optimize, for instance, we first predict b using linear regression and obtain $\hat{f}(x) = -0.5x + 1$. If the new observation is $x = 3$, we get $\hat{b} = -0.5$ which makes (P) infeasible.

Obs. Hypothesis class of linear functions does not work if apply ERM (even with Tikhonov-type regularizations)

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), b_i) + \lambda \|f\|_{\mathcal{H}}^2$$

Should come up with $\mathcal{H}$ in the way that the ERM problem well-posed AND ensures feasibility

- Idea 1. Using point estimate of upstream prediction

  Given new observation $\breve{x}$, we predict $\breve{b}$. In the previouse example, if the new observation is $\breve{x} = 3$, we get the predicted value $\breve{b} = -0.5$, which is not feasible at the moment. Consider $\breve{b} \pm \epsilon$ in the way that makes (P) feasible. (i.e., take $\epsilon = 0.5$)

  In general, we can consider a ball centered at $\breve{b}$. Let

  $$\begin{aligned} B_\epsilon(\breve{b}) &:= \{b : \|b - \breve{b}\| \leq \epsilon\} \\ S_\epsilon(\breve{b}) &:= \{w : a^\top w \leq b, b \in B_\epsilon(\breve{b})\} \end{aligned}$$

  We can take ϵ s.t. $S_0 \cup S_\epsilon(\breve{b}) \neq \emptyset$. Note that if S_0 is compact, there exists ϵ s.t. the intersection is a singleton (i.e., lies on the boundary of S_0)

  JERO type argument to set value of ϵ??

  This is more like predict then robust optimize (heavily depends on the prediction methodologies)

- Idea 2. One-shot way (regret-based loss function approach)

  Want to capture the discrepancy of estimated cost (or estimated feasible set) and the optimal cost (or optimal feasible set)

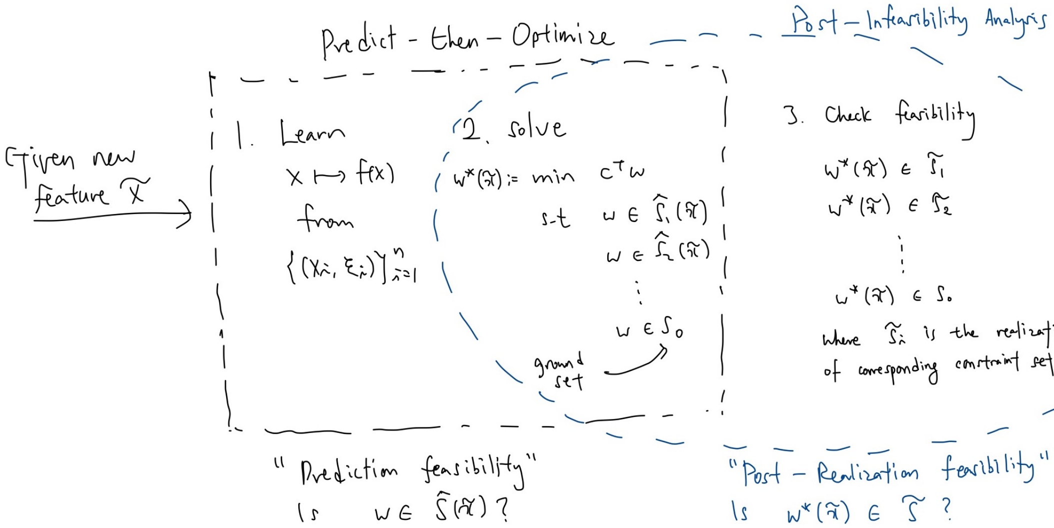
  What should be the appropriate regret?

  Types of costs:

  (i) Naïve cost: $\min_{w \in S_0} c^\top w$
  Note: This may be unbounded.

  (ii) Good feasible cost (from the feasible set formulation above): $\min_{w \in S_0 \cap S_\epsilon(\breve{b})} c^\top w$

  (iii) Closest feasible cost: $\min_{w \in S_0 \cap S_{i^*}} c^\top w$
  Going back to the previous example. We can just choose the closest feasible set from the data. That is, choose b_i s.t. $\|\breve{x} - x_i\|$ is minimized (w.r.t certain metric) where $\breve{x}$ is the new observation.



If $S(\hat{b})$ is well-defined using (ii),

$$
\begin{aligned}
\ell(\hat{b}, b) :&= \left| \min_{w \in S(\hat{b})} c^\top w - \min_{w \in S(b)} c^\top w \right| \\
&= \left| \min_{w \in S_0 \cap S_\epsilon(\hat{b})} c^\top w - \min_{w \in S(b)} c^\top w \right|
\end{aligned}
$$

ERM:

$$
\frac{1}{n} \sum_{i=1}^n \sum_{i=1}^n \ell(f(x_i), b_i) = \frac{1}{n} \sum_{i=1}^n \left| \min_{w \in S_0 \cap S_\epsilon(f(x_i))} c^\top w - \min_{w \in S(b_i)} c^\top w \right|
$$

Choosing the right cost with plausible feasible set (first term in RHS) and right hypothesis class $\mathcal{H}$ remains funky...

- Idea 3. Contextual Chanced Constrained Optimization

  Assume $\{(x_i, b_i)\}_{i=1}^n$ is i.i.d. from a joint distribution $p(x, b)$.

$$
\begin{aligned}
\min_w \quad & c^\top w \\
\text{s.t.} \quad & p_x(w) := P_{B|X}(h(w, b) \le 0 \mid X = x) \ge 1 - \alpha \\
& w \in S_0
\end{aligned}
$$

  where $P_{B|X}$ is the conditional probability of B given X. For each $X = x$, we have a solution $w^*(x)$ to CCCO problem.

  Possible DRO extension? Define a empirical distribution $\hat{\mathbb{P}}_n$ and consider an ambiguity set $\mathcal{P}_r := \{\mathbb{P} : d(\mathbb{P}, \hat{\mathbb{P}}_n) \le r\}$ along with contextual information $X = x$

- Idea 4. Algorithmic approach?

Two sources of infeasibility:

1. Infeasibility from Prediction:

   Learned mapping should return feasible constraint set (This may be compromised in some sense at the cost of suboptimality)

2. Post-Realization Infeasibility:

   Given feature $\tilde{x}$, our optimal solution $w^*(\tilde{x})$ solved over the predicted constraints may not be feasible for the true constraints associated with x.

3.1 Prediction Infeasibility

- Limiting output space of $f : \mathcal{X} \to S$
  - Coarse hypothesis space $\mathcal{H}$ in the output domain
- Design loss function that captures feasibility
  - If we want to use regret-based (i.e., $|predicted - optimal|$) formulation, this may heavily depend on the feasible set construction in the prediction step.
- Minimal Adjustment Ideas (possibly from point estimates of upstream prediction)
  - (Naïvely) Enlarge the infeasible constraint set: Similar to Idea 1. above
  - Distance to ill-posedness (Vera 1998, Ordonez & Freund 2003, Renegar 1994)
    $d = (A, b, c^*)$ tuple of LP instance:

$$\begin{aligned} \sup_{x \in X} \quad & c^* x \\ \text{s.t.} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

The norm of d is defined as:

$$\|d\| := \max\{\|A\|, \|b\|, \|c^*\|\}$$

We further define distances to the primal and dual infeasible sets as follows:

$$\rho_P(d) := \mathrm{dist}(d, \mathrm{Pri}\emptyset) := \inf\{\|d - \tilde{d}\| : \tilde{d} \in \mathrm{Pri}\emptyset)\}$$
$$\rho_D(d) := \mathrm{dist}(d, \mathrm{Dual}\emptyset) := \inf\{\|d - \tilde{d}\| : \tilde{d} \in \mathrm{Duali}\emptyset)\}$$

Theorem 1 *Assume X is reflexive, C_X and C_Y are closed. Assume $d = (A, b, c^*) \in D$. If d satisfies $dist(d, Pri\emptyset) > 0$ then statements (1) and (2) are true*

1. *There exists $x \in Feas(d)$ s.t.*

$$\|x\| \leq \frac{\|b\|}{dist(d, Pri\emptyset)}$$

2. *If $x' \in Feas(d + \Delta d)$ where $\Delta d := (0, \Delta b, 0)$ (i.e., perturbation of b alone), then there exists $x \in Feas(d)$ satisfying*

$$\|x - x'\| \leq \|\Delta b\| \frac{\max\{1, \|x'\|\}}{dist(d, Pri\emptyset)}$$

Note: With linear equations (when cones C_X and C_Y are subspaces) the term $\max\{1, \|x'\|\}$ *can be replaced by 1.*

The quantity $C(d)$ is called the "condition measure" or the "condition number" of the problem instance d and is a (positively) scale-invariance reciprocal of the smallest data perturbation Δd that will render the perturbed data instance either primal or dual infeasible

$$C(d) := \frac{\|d\|}{\min\{\rho_P(d), \rho_D(d)\}}$$

a problem is "ill-posed" if $\min\{\rho_P(d), \rho_D(d)\} = 0$ i.e., $C(d) = \infty$
When a feasible instance of (FPd) : $Ax = b, x \in C_X$ is well-posed ($C(d) < \infty$), there exists a point x feasible for (FPd) which satisfies $\|x\| \leq C(d)$.

- Infeasibility Repair
  - Isolate an Irreducible Infeaasible System (IIS)
    An IIS is a minimal set of constraints and variable bounds which is infeasible, but becomes feasible if any constraint or bound in it is removed.

- Find a Maximum Feasible Subset (Max FS)
  Maximum cardinality subset of constraints that is feasible (NP-hard)
- Find “Best fix” for infeasible constraints
  Different matrix norms for measuring “best fix”
  * Shifting constraints: smallest number (same as Max FS, weighted/unweighted)
  * Shifting constraints: smallet total penalty (minimizing ℓ_1 norm; same as elastic program, weighted/unweighted)
  * Altering constraint body: minimize a matrix norm

$$\min \phi(H, p) \text{ s.t. } (A + H)x \leq b + p$$

Further notes on shifting constraints (elasticizing Constraints in LP + related optimization methods):

- Make all constraints elastic by adding elastic variables e_i
- Elastic objective: $\min \sum_i e_i$
  Original constraint $g(x) \leq b_i$ becomes $g(x) \leq b_i + e_i$
- For a matrix M, ℓ_1 matrix norm is defined $\|M\|_{\ell_1} = \sum_{ij} |d_{ij}|$. For a system of $Ax \geq b$, a full elastic program is identical to minimizing an ℓ_1 matrix norm subject to elastic constraints i.e., $\min \|(b - Ax)^+\|_{\ell_1}$ ($(\cdot)^+$ denote the component-wise application of $\max\{0, \cdot\}$
- Least-squares method: find the point that has the smallest sum of squared constraint viloations

$$\min \sum_{i=1}^{m} [(b_i - a_i x)^+]^2$$

Automatic repair of convex optimization problems (Barratt et al., 2020)

Parametric convex optimizattion

$$\begin{aligned} (1)\ \min_x \quad & f_0(x; \theta) \\ \text{s.t.} \quad & f_i(x; \theta) \leq 0, i = 1, \ldots, m \\ & g_j(x; \theta) = 0, j = 1, \ldots, p \end{aligned}$$

Let $p^* = \inf\{f_0(x; \theta) : x \in S\}$

- Solvable if p^* is finite and attainable
- Infeasible if $p^* = \infty$
- Unbounded if $p^* = -\infty$
- Pathologcal if p^* is finite but unattainable by any x

Performance metric $r : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ (e.g., $r(\theta) := \|\theta - \theta_0\|_2$ where θ_0 is the initial parameter vector): The goal of the paper is to repair an unsolvable problem by adjusting the parameter θ so that it becomes solvable.

$$\begin{aligned} \min_\theta \quad & r(\theta) \\ \text{s.t.} \quad & \text{problem (1) is solvable} \end{aligned}$$

Canonical form of (1):

$$\begin{aligned} (\text{P})\ \min_x \quad & c(\theta)^\top x \\ \text{s.t.} \quad & A(\theta)x + s = b(\theta) \\ & s \in \mathcal{K} \end{aligned}$$

$$
\begin{aligned}
\text{(D)} \min_{x} \quad & -b(\theta)^\top y \\
\text{s.t.} \quad & A(\theta)^\top y + c(\theta) = 0 \\
& y \in \mathcal{K}^*
\end{aligned}
$$

Solvability here implies (P) feasible, (D) feasible, and strong duality holds.

(x^*, y^*, s^*) is a solution if

$$
\begin{bmatrix} A(\theta)x^* + s^* \\ A(\theta)^\top y^* + c(\theta) \\ c(\theta)^\top x^* + b(\theta)^\top y^* \end{bmatrix} = \begin{bmatrix} b(\theta) \\ 0 \\ 0 \end{bmatrix}, (s^*, y^*) \in \mathcal{K} \times \mathcal{K}^*
$$

Primal-dual embedding:

$$
\begin{aligned}
t^*(\theta) := \min \quad & t \\
\text{s.t.} \quad & \begin{bmatrix} A(\theta)x^* + s^* - b(theta) \\ A(\theta)^\top y^* + c(\theta) \\ c(\theta)^\top x^* + b(\theta)^\top y^* \end{bmatrix} \leq t \\
& s \in \mathcal{K}, y^* \in \mathcal{K}^*
\end{aligned}
$$

Note: $t^*(\theta) = 0$ iff (1) is solvable.

Reformulation:

$$
\begin{aligned}
\min \quad & r(\theta) \\
\text{s.t.} \quad & t^*(\theta) = 0
\end{aligned}
$$

This is NP-hard, so rely on heuristic to find an approximate solution (also assumes differentiability of t^* s.t. $\nabla t^*(\theta)$ exists.

3.2 Post-Realization Infeasibility

3.2.1 Contextual Chance Constrained Programming

Sample Average Approximation

$$
\begin{aligned}
\inf_{u \in \mathcal{U}} \quad & f(u) \\
\text{s.t.} \quad & P^N\{G(u, \xi) \leq 0\} = \frac{1}{N} \sum_{i=1}^{N} 1_{(-\infty, 0]}\{G_i(i, \xi^i)\} \geq 1 - \alpha
\end{aligned}
$$

CCP in data-rich world: "contextual information" associated with the random parameters ξ_i

$$
\begin{aligned}
\inf_{u \in \mathcal{U}} \quad & f(u) \\
\text{s.t.} \quad & P_{\xi|X}(G(u, \xi) \leq 0 \mid X = x) \geq 1 - \alpha
\end{aligned}
$$

x observed covariate realization of random X, and a minimizer will be $u^*(x)$ Cons:

- Multi-dimensional integrals
- Don't know the conditional distribution $\xi \mid X = x$

Pros:

- Historical data $D^n\{(x^i, \xi^i)\}_{i=1}^N$: covariates x^i along with the realizations of random parameter of ξ^i
- Observe a new covariate x and make a decision u given x but uncertain ξ

Numerical example shows that SAA can go wrong (returns infeasible solution i.e., the calculated conditional probability is less than $1-\alpha$)

Given historical data $D^n\{(x^i,\xi^i)\}_{i=1}^N$ and $X=x$, we estimate the probability distribution of $\xi \mid X=x$ as

$$P_{\xi|X}^N = \sum_{i=1}^N w_i(x)\delta_{\xi^i}$$

where δ_{ξ^i} is the Dirac point mass on ξ^i $(i=1,\ldots,N)$

Classical SAA: uniform weight

$$\frac{1}{N}\sum_{i=1}^N 1_{(-\infty,0]}\{G_i(i,\xi^i)\} \geq 1-\alpha$$

Data-driven Contextual CC: weights depending on feature vector

$$\sum_{i=1}^N w_i(x)1_{(-\infty,0]}\{G(u,\xi^i)\} \geq 1-\alpha$$

ML-Based Weighting Functions (follows Bertsimas & Kallus)

- Kernel function: Nadaraya-Watson estimation
- k-NN
- CART (or RF)

Note: True chance constraint can be written as

$$\mathbb{E}[1_{(-\infty,0]}\{G(u,\xi)\}] \geq 1-\alpha$$

Note: Feasibility depends on level α

- True feasible region
$$\mathcal{U}_\alpha := \{u \in U : \mathbb{E}[1_{(-\infty,0]}\{G(u,\xi)\}] \geq 1-\alpha\}$$
- Approximated feasible region:
$$\mathcal{U}_\alpha^n := \{u \in U : \sum_{i=1}^N w_i(x)1_{(-\infty,0]}\{G(u,\xi^i)\} \geq 1-\alpha\}$$

Things to consider (results from large deviation theory for weighted sum of random variables):

- Consistency:
$$\text{Prob}(\mathcal{U}_\gamma^N \subseteq \mathcal{U}_\alpha)$$
i.e., $(1-\gamma)$-feasible solutions to the approximated problem are $(1-\alpha)$-feasible to the true problem
- Finite sample guarantees: Estimate the sample size N s.t.
$$\text{Prob}(\mathcal{U}_\gamma^N \subseteq \mathcal{U}_\alpha) \geq 1-\delta$$
i.e., $(1-\gamma)$-feasible solutions to the approximated problem are $(1-\alpha)$-feasible to the true problem with a probability at least $(1-\delta)$

Proves results for finite $\mathcal{U}$, infinite but bounded $\mathcal{U}$

Future directions: weighting functions are ambiguous (distributionally robust framework: consider ambiguity set for the weighting functions)

3.2.2 Distributionally Robust Chance-Constrained Programs with Right-Hand Side Uncertainty under Wasserstein Ambiguity

Setting:

- $\mathcal{X} \subset \mathbb{R}^L$: a compact domain for the decision variables x
- $\mathcal{S}(x) \subseteq \mathbb{R}^K$: a decision-dependent safety-set
- $\xi \in \mathbb{R}^K$: a random vector with distribution $\mathbb{P}^*$
- $\epsilon \in (0,1)$: the risk tolerance for the random variable ξ falling outside the safety set $\mathcal{S}(x)$

$$
\begin{aligned}
\text{(CCP)} \quad \min_{x} \quad & c^\top x \\
\text{s.t.} \quad & \mathbb{P}^*[\xi \notin \mathcal{S}(x)]) \leq \epsilon \\
& x \in \mathcal{X}
\end{aligned}
$$

- In practice, the distribution $\mathbb{P}^*$ in the chance constraint in (CCP) is often unavailable to the optimizer
- Evaluating $\mathbb{P}^*[\xi \in \mathcal{S}(x)]$ exactly is often difficult even when $\mathbb{P}^*$ is available
- Instead, i.i.d. samples $\{\xi_i\}_{i\in[n]}$ are drawn from $\mathbb{P}^*$ and $\mathbb{P}^*$ is approximated using the empirical distribution $\mathbb{P}_N$ on these samples

$$
\begin{aligned}
\text{(SAA)} \quad \min_{x} \quad & c^\top x \\
\text{s.t.} \quad & \frac{1}{N}\sum_{i=1}^{N} 1(\xi_i \notin \mathcal{S}(x)) \leq \epsilon \\
& x \in \mathcal{X}
\end{aligned}
$$

Sample Average Approximation (SAA) of CCP sometimes can be refomulated as an MIP.

Still, the out-of-sample performace of the solution from (SAA) is sensitive to the specific sample $\{\xi_i\}_{i\in[N]}$ and can result in high variance when N is small.

Remedy for the out-of-sample performance? Use distributionally robust CCP!

$$
\begin{aligned}
\text{(DR-CCP)} \quad \min_{x} \quad & c^\top x \\
\text{s.t.} \quad & \sup_{\mathbb{P}\in\mathcal{F}_N(\theta)} \mathbb{P}[\xi \notin \mathcal{S}(x)]) \leq \epsilon \\
& x \in \mathcal{X}
\end{aligned}
$$

where

- $\mathcal{F}_N(\theta)$: an ambiguity set of distribution of $\mathbb{R}^K$ that contains the empirical distribution $\mathbb{P}_N$
- θ: a parameter that governs the size of the ambiguity set (i.e., the conservatism of (DR-CCP))

Ambiguity set:

Wasserstein ambiguity set with radius θ:

$$\mathcal{F}_N(\theta) := \{\mathbb{P} : d_W(\mathbb{P}_N, \mathbb{P}) \leq \theta\}$$

where

$$d_W(\mathbb{P}, \mathbb{P}') := \inf_{\Pi} \left\{ \mathbb{E}_{(\xi,\xi')\sim\Pi}[\|\xi - \xi'\|] : \Pi \text{ has a marginal distributions } \mathbb{P}, \mathbb{P}' \right\}$$

Distance to the unsafe set: Given a decision $x \in \mathcal{X}$ and a realization $\xi \in \mathbb{R}^K$, the distance from ξ to the unsafe set is

$$\operatorname{dist}(\xi, \mathcal{S}(x)) := \inf_{\xi' \in \mathbb{R}^K} \{\|\xi - \xi'\| : \xi' \notin \mathcal{S}(x)\}$$

Consider the following safety set:

- Safety set defined by linear inequalities with RHS uncertainty:

  $$\mathcal{S}(x) := \left\{\xi : b_p^\top \xi + d_p - a_p^\top x > 0, p \in [P]\right\}$$

  for given b_p, d_p, a_p

- Distance to safety set:

  $$\operatorname{dist}(\xi, \mathcal{S}(x)) = \max\left\{0, \min_{p \in [P]} \frac{b_p^\top \xi + d_p - a_p^\top x}{\|b_p\|_*}\right\}$$

Reformulation of (DR-CCP):

The feasible set

$$\mathcal{X}_{DR}(\mathcal{S}) := \left\{x \in \mathcal{X} : \sup_{\mathbb{P} \in \mathcal{F}_N(\theta)} \mathbb{P}[\xi \notin \mathcal{S}(x)] \leq \epsilon\right\}$$

is equivalent to

$$\mathcal{X}_{DR}(\mathcal{S}) = \left\{ x \in \mathcal{X} : \begin{array}{l} \exists t \geq 0, r \geq 0, \\ \operatorname{dist}(\xi_i, \mathcal{S}(x)) \geq t - r_i, i \in [N], \\ \epsilon t \geq \theta + \frac{1}{N} \sum_{i \in [N]} r_i \end{array} \right\}$$

can linearize the dist constraint using MIP

Example: Computational study (Chen et al., 2018)

Given set of factories $[F]$ with capacities m_f, $f \in [F]$, a set of distribution centers $[D]$ must meet the random demands ξ_d, $d \in [D]$ w.h.p at minimum cost

$$\begin{aligned} \min_x \quad & c^\top x \\ \text{s.t.} \quad & \mathbb{P}\left[\xi \notin \mathcal{S}(x)]\right) \geq 1 - \epsilon, && \mathbb{P} \in \mathcal{F}(\theta) \\ & \sum_{d \in [D]} x_{fd} \leq m_f, && f \in [F] \\ & x_{fd} \geq 0, \quad , && f \in [F], d \in [D] x \in \mathcal{X} \end{aligned}$$

3.3 Further Thoughts

Ideas from recourse (stochastic) programs and fuzzy programming?
Ideas using duality?
DRO using maximum mean discrepancy? (seems like finding the mean difference align with finding the difference of predicted and optimal values in the loss function)
Two-stage stochastic optimization to deal with "post-realization" problems