Sparse Optimization with Distributionally Robust Chance Constraint

D2 Zhicheng Zhang

@IPS, Fujisaki Lab

July 19, 2022

Sparse Decision-Making Under Uncertainty

Sparse chance constraint optimization (SCCO)

Consider the sparse chance constraint optimization as

$$
\begin{aligned}
\min_{x} \quad & \|x\|_0 && \text{(sparse cost)} \\
\text{s.t.} \quad & x \in \mathcal{X} && \text{(deterministic constraint)} \\
& \mathbb{P}\{h(x,\delta) \leq 0\} \geq 1-\epsilon && \text{(chance constraint)} \\
\Leftrightarrow \quad & \mathbb{P}\{h(x,\delta) > 0\} < \epsilon && \text{(violation)}
\end{aligned}
$$

- Non-convex cost function
- Uncertainty δ is assumed to be a random variable governed by probability $\mathbb{P}$ supported on $\Delta \subseteq \mathbb{R}^{n_\delta}$
- Violation event does not exceed a risk level $\epsilon \in (0,1)$
- Multivariate integral computation and non-convex feasible set

Sparsity

Definition (Sparsity)

A vector $x \in \mathbb{R}^n$ is sparse if it contains many 0's, or has ℓ_0 "norm"

$$\|x\|_0 = |\mathrm{supp}(x)| \qquad \Rightarrow \qquad \|x\|_0 \leq s \quad \text{(s-sparse)},$$

where $\mathrm{supp}(x) \doteq \{l \in \{1, \cdots, n\} : x_l \neq 0\}$ denotes the number of the nonzero elements in x.

Convex Relaxation (ℓ_1 norm)

The ℓ_1 norm convex set is defined as

$$\Sigma_s := \{x \in \mathbb{R}^n : \|x\|_1 \leq s\}, \qquad \text{(Lasso)}$$

where the ℓ_1 norm constraints $\|x\|_1 = \sum_{l=1}^n |x_l|$.

- When the assumption $\|x\|_\infty \leq 1$ holds, then the biconjugate function of ℓ_0 norm gives the result $\|x\|_0^{\star\star}(r) = \|r\|_1 \leq \|x\|_0$.

Conjugate Function

- The conjugate function of $\|x\|_0$ is defined by

$$\|x\|_0^{\star}(y) := \sup_x \{\langle x, y\rangle - \|x\|_0\} = \max\left(\sum_i |y_i|, 0\right),$$

it is always a convex form even the $\|x\|_0$ is non-convex.

- The biconjugate function of $\|x\|_0$ is as follows

$$\|x\|_0^{\star\star}(r) := \sup_y \{\langle r, y\rangle - \|x\|_0^{\star}(y)\} = \|r\|_1$$

- Hence, ℓ_1 norm is a *convex relaxation* of ℓ_0 norm in some sense due to the fact that $\|x\|_0^{\star\star}(r) \leq \|x\|_0$.

Scenario Approximation for Uncertainty

Sparse Scenario Approximation (SA) [Calafiore & Campi, 05, 06]

The SA entails randomized algorithms to randomly generate N samples i.i.d. from the probability $\mathbb{P}$, and the uncertainty set Δ is replaced by a finite samples $\{\delta^{(i)}\}_{i=1}^{N}$, then (SCCO) becomes sparse SA

$$J_N^* = \min_x \left\{ \|x\|_1 : x \in \mathcal{X},\ h(x, \delta^i) \leq 0,\ \forall i = 1, \cdots, N \right\},$$

which may be recast as the following *epigraphic form*

$$J_N^{s^*} = \min_{x,s} \left\{ s : x \in \mathcal{X}_s,\ \max_{i=1,\cdots,N} h(x, \delta^i) \leq 0 \right\},\ \text{where } \mathcal{X}_s = \{\|x\|_1 \leq s \cap \mathcal{X}\}$$

Sample complexity [Campi & Garatti, 13]

Given $\epsilon, \beta \in (0,1)$ and $q = \dim(\mathcal{X}_s) < n$. The sample complexity is

$$N \geq \frac{2}{\epsilon}\left(\ln \frac{1}{\beta} + q \ln \frac{n \cdot e}{q} \right), \quad e \approx 2.718\cdots$$

then it holds that $\mathbb{P}^N\{V(x_N^*(s^*)) > \epsilon\} \leq \beta$.

Sample Average Approximation for Uncertainty

Sample Average Approximation (SAA) [Luedke & Ahmed, 08]

The SAA is a probabilistic constraint relaxation for *out-of-sample* in SA under a significance level $\alpha \in (0,1)$, that is, $\widehat{p}_N(x) \leq \alpha$, where $0 \leq \alpha < \epsilon$. Therefore, the sparse SAA program is as follows

$$J^*_{N,\alpha} = \min_x \{\|x\|_1 : x \in \mathcal{X},\ \widehat{p}_N(x) \leq \alpha,\ i = 1, \cdots, N\}$$

- Chance constraint $p(x) := \mathbb{P}\big(\delta : h(x,\delta) > 0\big) = \mathbb{E}_{\mathbb{P}}\big[\mathbb{I}(h(x,\delta))\big]$
- Estimate the "true" probability via the *discrete empirical distribution*

$$\widehat{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(h(x,\delta^i)\right), \quad i = 1, \cdots, N.$$

- If $\alpha = 0$, then sparse SAA program reduces to sparse SA program
- Assess worst case probability of two-sided failure for given $\hat{\epsilon}, \hat{\beta} \in (0,1)$

$$\mathbb{P}^N\Big\{\sup_{x\in\mathcal{X}} |p(x) - \widehat{p}_N(x)| > \hat{\epsilon}\Big\} < \hat{\beta} \qquad \text{(VC theory)}$$

Exact Sparsity via SAA

Proposition (Exact Sparsity via SAA)

The exact ℓ_0 norm constraint is equivalent to a SAA of chance constraint

$$\|x\|_0 \leq s \;\Leftrightarrow\; \frac{1}{n}\sum_{l=1}^{n} \mathbb{I}\big(|x_l| \leq 0\big) \geq 1 - \frac{s}{n} \;\Leftrightarrow\; \frac{1}{n}\sum_{l=1}^{n} \mathbb{I}\big(|x_l| > 0\big) \leq \frac{s}{n}$$

- n scenarios are with *equal* probability $\frac{1}{n}$
- The l-th scenario index set is $\mathcal{S}^l := \{x : x_l = 0,\ \forall l = 1, \cdots, n\}$
- $\|u\|_0 \leq s$ means that at most s out of the n scenarios are violated.

SCCO via SAA Reformulation

The problem (SCCO) can be recast as a SAA formulation in sparse cost and chance constraint, that is,

$$\min_{s} \left\{ s : \; \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(h(u, \delta^i)\right) \leq \alpha,\; i \in \mathcal{N},\; \frac{1}{n_u}\sum_{l=1}^{n_u} \mathbb{I}\big(|u_l| \leq 0\big) \geq 1 - \frac{s}{n_u} \right\}$$

Mixed Integer Programming

Proposition (Exact Sparsity via MIP)

The exact ℓ_0 norm constraint is equivalent to a MIP formulation

$$\|x\|_0 \leq s \Leftrightarrow \ |x_l| \leq M_l z_l, \ \sum_{l=1}^{n} z_l \leq s, \ z_l \in \{0,1\}, \ l = 1, \cdots, n$$

$$\Leftrightarrow \ |x| \leq Mz, \ e^\top z \leq s, \ z \in \{0,1\}^n, \ e = [1 \cdots 1]$$

- Binary variables z_l is an auxiliary variable to evaluate the sparsity
- $z_l = 1$ counts the nonzero elements
- Solve big-M coefficients
- Boolean convex relaxation $z \in [0,1]^n$ (big-M free)

Chance Constraint via MIP

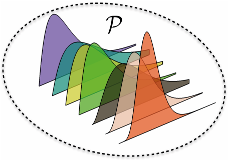
MIP for Chance constraint

We introduce an auxiliary binary variables $v_i \in \{0,1\}$ for each $i \in \mathcal{N}$, where $v_i = 1$ assures that the safety or reliability event $h(x, \delta^i) \leq 0$ holds; and otherwise $v_i = 0$ indicates the violation event. Thus, we express a MIP form

$$
\begin{aligned}
& h(x,\delta^i) \leq \eta_i(1 - v_i), && \big(\Leftrightarrow h(x,\delta^i) + \eta_i v_i \geq 0\big) \\
& \sum_{i=1}^{N} p_i v_i \geq 1 - \alpha, && \Big(\Leftrightarrow \sum_{i=1}^{N} p_i(1 - v_i) \leq \alpha\Big) \\
& v_i \in \{0,1\}, \quad \forall i = 1, \cdots, N,
\end{aligned}
$$

where $\eta_i \in \mathbb{R}$ and $0 < p_i = \mathbb{P}\{\xi = \xi^i\}$ are the *non-equal* probability of possible outcomes as scenarios and satisfy $\sum_{i=1}^{N} p_i = 1$. We refer these more generic constraints as *knapsack constraints*.

Distributionally Robust Chance Constraint [Calafiore & El Ghaoui, 06]



Sparse and distributionally robust optimization (SDRO)

Consider a sparse distributionally robust (chance constrained) optimization

$$\begin{aligned} \min_{x} \quad & \|x\|_0 \\ \text{s.t.} \quad & \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{\delta\in\Delta : f(x,\delta)\leq 0\} \geq 1-\epsilon \qquad \text{(safety)} \\ & \Leftrightarrow \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{\delta\in\Delta : h(x,\delta) > 0\} \leq \epsilon \end{aligned}$$

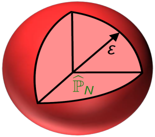
- "Ambiguity set" $\mathcal{P}$ = a family of probability distributions.
- Moment ambiguity set $\mathcal{P}(\mu,\Sigma)$, and metric ambiguity set $\mathcal{P}(\varepsilon)$

Data-Driven Wasserstein Ambiguity Set

Definition (Wasserstein ambiguity set) [Gao & Kleywegt, 17; Esfahani & Kuhn, 18]

The Wasserstein ambiguity set $\mathcal{P}^W$ can be defined as

$$\mathcal{P}^W := \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N) = \left\{ \mathbb{Q} : W_p(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \varepsilon \right\}, \quad \widehat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\hat{\delta}(i)} \text{ (empirical)}$$



Contains every $\mathbb{Q}$ obtainable by re-shaping $\widehat{\mathbb{P}}_N$ at a cost of at most ε

- $W_p(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)} \left(\int_{\Delta \times \Delta} \|\delta_1 - \delta_2\|^p \; \pi(\mathrm{d}\delta_1, \mathrm{d}\delta_2), \right)^{\frac{1}{p}}$, and Π is the set of couplings with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$. [Villani, 08]
- Be capable of comparing a continuous and a discrete distribution (weak convergence and convergence in the p-th moment)

Feasibility Analysis

$$\inf_{\mathbb{P}\in\mathcal{P}^W} \mathbb{P}\{h(x,\delta)\le 0\} \ge 1-\epsilon \quad \Leftrightarrow \quad \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{P}\{h(x,\delta)>0\} \le \epsilon$$

Primal Problem (Wassesterin metric)

The worst case ambiguous (violation) uncertainty quantification

$$(P) \qquad J^P = \sup_{\mathbb{P}\in\mathcal{P}^W} \mathbb{P}\big(h(x,\delta)>0\big) = \sup_{\mathbb{P}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{P}}\big[\mathbb{I}(h(x,\delta))\big]$$

$$= \begin{cases} \sup\limits_{\Pi,\mathbb{P}} \int_\Delta \mathbb{I}(h(x,\delta))\mathbb{P}(\mathrm{d}\delta) \\ \text{s.t. } W_p(\widehat{\mathbb{P}}_N,\mathbb{P}) \le \varepsilon \end{cases}$$

$$= \begin{cases} \sup\limits_{\mathbb{P}_i} \int_\Delta \mathbb{I}(h(x,\delta))\mathbb{P}_i(\mathrm{d}\delta) \\ \text{s.t. } \frac{1}{N}\sum_{i=1}^N \int_\Delta \|\delta-\hat{\delta}^i\|_p \mathbb{P}_i(\mathrm{d}\delta) \le \varepsilon \end{cases}$$

where the optimal value of primal problem (P) denotes J^P.

Tractable Performance Reformulation

Dual Problem

Using a standard Lagrangian dual variable $\lambda \geq 0$, the dual of primal problem (P) is as follows

$$
(D) \qquad J^D = \inf_{\lambda \geq 0} \left\{ \lambda \varepsilon - \int_{\Delta} \inf_{\delta \in \Delta} \left[\lambda \| \delta - \hat{\delta}^i \|_p - \mathbb{I}(h(x,\delta)) \right] \widehat{\mathbb{P}}_N(\mathrm{d}\hat{\delta}^i) \right\}
$$

$$
= \inf_{\lambda \geq 0} \left\{ \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} \sup_{\delta \in \Delta} \left[\mathbb{I}(h(x,\delta)) - \lambda \| \delta - \hat{\delta}^i \|_p \right] \right\}
$$

where the optimal value of dual problem (D) denotes J^D.

- Strong duality: $J^P = J^D$
- Introduce epigraphical auxiliary variables ζ_i, $\forall i \in \mathcal{N}$, then (D) is as

$$
\inf_{\lambda, \zeta_i} \left\{ \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} \zeta_i \quad \text{s.t.} \ \sup_{\delta \in \Delta} \left[\mathbb{I}(h(x,\delta)) - \lambda \| \delta - \hat{\delta}^i \|_p \right] \leq \zeta_i, \ \lambda \geq 0 \right\}
$$

Details

Reformulation [Gao & Kleywegt, 17; Esfahani & Kuhn, 18]

For $i \in \mathcal{N}$, the optimal value of (P) is equal to the optimal value of (D) under the Wasserstein ambiguity set $\mathcal{P}^W = \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, that is

$$\sup_{\mathbb{P}\in\mathcal{P}^W} \mathbb{P}\big(h(u,\delta) > 0\big) = \inf_{\lambda\geq 0}\left\{\lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\sup_{\delta\in\Delta}\left[\mathbb{I}(h(x,\delta)) - \lambda\|\delta - \hat{\delta}^i\|_p\right]\right\}$$

- Break down indicator function in infimum/supremum by discussing the condition of taking "zero" and "one"
- In general, the unsafety event can be redefined as

$$\sup_{\mathbb{P}\in\mathcal{P}^W} \mathbb{P}\big(x(\delta) \notin \mathcal{S}(\delta)\big) \leq \epsilon \quad \Rightarrow \quad x(\hat{\delta}^i) \text{ via } \widehat{\mathbb{P}}_N$$

- $\mathbb{I}(h(x,\delta))$ is the violation (or unsafety) event that governed by a metric/distance between a random point and the unsafety set.

Take Home Message

- SCCO can be approximated by scenario approximation,sample average approximation and Data-driven Wasserstein ball setting.
- SCCO can be recast as a SAA form
- SCCO can be recast as a MIP form

Future work

- The selection of performance function $h(x, \delta)$.
- Wasserstein ambiguity feasible set analysis.
- Using MIP to solve sparse distributionally robust chance constrained program.