Distributionally Robust Optimization (DRO) with Moment-based Ambiguity Set

CHEN Li

IORA, NUS

Seminar

August 8th, 2020

Outline

Decision making under uncertainty

- Stochastic programming

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x},\tilde{\boldsymbol{z}})\right]$$

- Robust Optimization

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{z}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{z})$$

- Risk-averse optimization

$$\min_{\boldsymbol{x}\in\mathcal{X}} \rho[f(\boldsymbol{x},\tilde{\boldsymbol{z}})]$$

- Distributionally robust optimization

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x},\tilde{\boldsymbol{z}})\right]$$

- Stochastic dynamic programming
- Reinforcement learning
- Online optimization
- ...

Stochastic programming

Given $\tilde{\boldsymbol{z}} \sim \mathbb{P}$, find decision $\boldsymbol{x}$ to minimize the objective function.

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{\mathbb{P}} \left[f(\boldsymbol{x}, \tilde{\boldsymbol{z}}) \right] \tag{SP}$$

Example: Newsvendor problem

- A newsvendor buy x unit of newspaper with unit cost $c > 0$ and sell newspapers at price $p > c$. The demand $\tilde{d} \sim \mathbb{P}$ is stochastic. The goal is to maximize the expected profit

$$\max_{x \geq 0} \mathbb{E}_{\mathbb{P}} \left[p \min\{x, \tilde{d}\} - cx \right]$$

Optimal solution: $x^* = F^{-1}(\frac{p-c}{p})$ where $F(\cdot)$ is the cdf of $\tilde{d}$.

The most popular SP model

- Two-stage linear SP

$$\min_{\boldsymbol{x} \geq \boldsymbol{0}: \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}} \boldsymbol{c}'\boldsymbol{x} + \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right] \tag{1}$$

where

$$\begin{aligned} Q(\boldsymbol{x}, \tilde{\boldsymbol{z}}) = \quad & \inf \quad \boldsymbol{d}(\tilde{\boldsymbol{z}})'\boldsymbol{y} \\ & \text{s.t.} \quad \boldsymbol{T}(\tilde{\boldsymbol{z}})\boldsymbol{x} + \boldsymbol{W}(\tilde{\boldsymbol{z}})\boldsymbol{y} = \boldsymbol{h}(\tilde{\boldsymbol{z}}) \\ & \qquad\ \ \boldsymbol{y} \geq \boldsymbol{0} \\ = \quad & \sup \quad \boldsymbol{\xi}'\left(\boldsymbol{h}(\tilde{\boldsymbol{z}}) - \boldsymbol{T}(\tilde{\boldsymbol{z}})\boldsymbol{x}\right) \\ & \text{s.t.} \quad \boldsymbol{\xi}'\boldsymbol{W}(\tilde{\boldsymbol{z}}) \leq \boldsymbol{d}(\tilde{\boldsymbol{z}})' \end{aligned}$$

- Relative complete recourse: Second-stage optimization is feasible for any feasible first-stage decision $\boldsymbol{x}$
- Fixed recourse: coefficient of $\boldsymbol{y}$ does not depend on $\tilde{\boldsymbol{z}}$
- Special case: $Q(\boldsymbol{x}, \boldsymbol{z}) = \max_{k \in [K]}\{\boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{z} + b_k(\boldsymbol{x})\}$

Solving stochastic programming

- Sample average approximation (SAA)[1]

$$\min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}, \boldsymbol{z}_i) \qquad \text{(SAA-}N\text{)}$$

  If $\mathcal{X}$ is nonempty and compact; $f(\cdot, \tilde{\boldsymbol{z}})$ is continuous w.p.1, and there exists $g(\cdot)$ such that $\sup_{x\in\mathcal{X}} f(\boldsymbol{x}, \tilde{\boldsymbol{z}}) \leq g(\tilde{\boldsymbol{z}})$ w.p.1, and $\mathbb{E}_{\mathbb{P}}\left[g(\tilde{\boldsymbol{z}})\right] < +\infty$; and $\boldsymbol{z}_i$ are i.i.d., then
  - Bias: $\mathbb{E}_{\mathbb{P}}\left[p_N^*\right] \leq \mathbb{E}_{\mathbb{P}}\left[p_{N+1}^*\right] \leq p^*$ where p_N^*, p^* are the optimal values of SAA-N and SP.
  - Consistency: $p_N^* \to p^*$ w.p.1 and the limiting point of x_N^* solves SP w.p.1 where x_N^* is the optimal solution of SAA-N.
  - Finite-sample complexity, convergence rate, ...
- Stochastic approximation[2]

[1] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[2] Arkadi Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on optimization* 19.4 (2009), pp. 1574–1609.

Algorithm for two-stage linear SP[3],[4]

- Extensive form

$$\min_{\boldsymbol{x}\geq \boldsymbol{0}:\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}} \boldsymbol{c}'\boldsymbol{x} + \frac{1}{N}\sum_{i=1}^{N} Q(\boldsymbol{x}, \boldsymbol{z}_i) \qquad (2)$$

where

$$\begin{aligned} Q(\boldsymbol{x}, \boldsymbol{z}_i) = \quad & \inf \quad \boldsymbol{d}_i'\boldsymbol{y} \\ & \text{s.t.} \quad \boldsymbol{T}_i\boldsymbol{x} + \boldsymbol{W}_i\boldsymbol{y} = \boldsymbol{h}_i \\ & \qquad\; \boldsymbol{y} \geq \boldsymbol{0} \end{aligned}$$

- L-shaped method (cutting plane/ Benders decomposition)
- Progressive hedging (Douglas-Rachford splitting)
- Interior point method

[3] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[4] Andrzej Ruszczyński. "Decomposition Methods". In: *Stochastic Programming*. Vol. 10. Handbooks in Operations Research and Management Science. Elsevier, 2003, pp. 141–211.

Early history: minmax SP

- Minmax SP[5]: $\mathbb{P}$ is not known exactly but

$$\mathbb{P} \in \mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}_+) \;\middle|\; \begin{array}{l} \tilde{d} \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{d}\right] = \mu \\ \mathbb{E}_{\mathbb{P}}\left[(\tilde{d} - \mu)^2\right] = \sigma^2 \\ \mathbb{P}\left[\tilde{d} \geq 0\right] = 1 \end{array} \right\}$$

  - The inner miminzation problem $\inf_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[p\min\{x, \tilde{d}\} - cx\right]$

$$= \begin{cases} \frac{\mu^2 px}{\sigma^2+\mu^2} - cx & \text{if } x \leq \frac{\sigma^2+\mu^2}{2\mu} \\ \frac{p}{2}(x + \mu - \sqrt{(x-\mu)^2 + \sigma^2}) - cx & \text{if } x \geq \frac{\sigma^2+\mu^2}{2\mu} \end{cases}$$

  - Solving $\max_{x\geq 0} \inf_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[p\min\{x, \tilde{d}\} - cx\right]$ for optimal policy:

$$x^* = \begin{cases} \mu + \sigma g(c/p) & \text{if } \frac{\mu^2}{\sigma^2+\mu^2} \geq \frac{c}{p} \\ 0 & \text{if } \frac{\mu^2}{\sigma^2+\mu^2} < \frac{c}{p} \end{cases} \quad \text{where } g(a) := \frac{1-2a}{2\sqrt{a(1-a)}}$$

[5] Herbert E Scarf. *A min-max solution of an inventory problem*. Tech. rep. RAND CORP SANTA MONICA CALIF, 1957.

From SP to DRO

- Generalized moment bounds for SP[6]
- Unknown distribution but some descriptive statistics or reference distribution[7]
  - Moment-based ambiguity set: Mean, absolute deviation, covariance, semi-variance, moment generating function, etc.
  - Statistical distance based ambiguity set: ϕ-divegence, Wasserstein metric, etc.
  - Other structures: symmetry, unimodality, independence, etc.

In the following, we assume

$$f(\boldsymbol{x}, \boldsymbol{z}) = \max_{k \in [K]} \{\boldsymbol{a}_k(\boldsymbol{x})' \boldsymbol{z} + b_k(\boldsymbol{x})\}$$

where $\boldsymbol{a}_k(\boldsymbol{x})$ and $b_k(\boldsymbol{x})$ are affine in $\boldsymbol{x}$ for simplicity unless it is specified.

[6] John R Birge and Roger J-B Wets. "Computing bounds for stochastic programming problems by means of a generalized moment problem". In: *Mathematics of Operations Research* 12.1 (1987), pp. 149–162.

[7] Hamed Rahimian and Sanjay Mehrotra. "Distributionally robust optimization: A review". In: *arXiv preprint arXiv:1908.05659* (2019).

Outline

How about multidimentisonal case?

Given ambiguity set

$$\mathbb{P} \in \mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^{I_z}) \left| \begin{array}{l} \tilde{\boldsymbol{z}} \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{z}}\right] = \boldsymbol{\mu} \\ \mathbb{E}_{\mathbb{P}}\left[(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma} \\ \mathbb{P}\left[\tilde{\boldsymbol{z}} \geq \mathbf{0}\right] = 1 \end{array} \right. \right\}$$

the DRO problem $\min_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right]$ can be reformulated as a copositive cone program:

$$\begin{array}{ll} \inf & \alpha + \boldsymbol{\beta}'\boldsymbol{\mu} + \boldsymbol{\Gamma} \bullet \boldsymbol{\Sigma} \\ \text{s.t.} & \alpha + \boldsymbol{\beta}'\boldsymbol{z} + \boldsymbol{\Gamma} \bullet (\boldsymbol{z}\boldsymbol{z}') \geq \boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{z} + b_k(x), \forall \boldsymbol{z} \geq \mathbf{0}, \forall k \in [K] \\ & \left(\Longleftrightarrow \begin{bmatrix} \alpha - b_k(x) & (\boldsymbol{\beta} - \boldsymbol{a}_k(\boldsymbol{x}))'/2 \\ (\boldsymbol{\beta} - \boldsymbol{a}_k(\boldsymbol{x}))'/2 & \boldsymbol{\Gamma} \end{bmatrix} \succeq_{co} \mathbf{0}, \forall k \in [K] \right) \end{array}$$

which is intractable.

Proof

$$\begin{aligned} & \alpha + \boldsymbol{\beta}'\boldsymbol{z} + \boldsymbol{\Gamma} \bullet (\boldsymbol{z}\boldsymbol{z}') \geq \boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{z} + b_k(x), \forall \boldsymbol{z} \geq \boldsymbol{0} \\ \iff & \begin{bmatrix} \alpha - b_k(x) & (\boldsymbol{\beta} - \boldsymbol{a}_k(\boldsymbol{x}))'/2 \\ (\boldsymbol{\beta} - \boldsymbol{a}_k(\boldsymbol{x}))'/2 & \boldsymbol{\Gamma} \end{bmatrix} \succeq_{co} \boldsymbol{0} \end{aligned}$$

- $\Longleftarrow$: take $\begin{bmatrix} 1 \\ \boldsymbol{z} \end{bmatrix} \geq \boldsymbol{0}$
- $\Longrightarrow$: For any $\begin{bmatrix} z_0 \\ \boldsymbol{z} \end{bmatrix} \geq \boldsymbol{0}$, if $z_0 > 0$, we consider $\begin{bmatrix} 1 \\ \boldsymbol{z}/z_0 \end{bmatrix}$; if $z_0 = 0$, it suffices to prove $\boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z} \geq 0$ for any $\boldsymbol{z} \geq \boldsymbol{0}$. Suppose not, then $\exists \bar{\boldsymbol{z}} \geq 0$ s.t. $\bar{\boldsymbol{z}}'\boldsymbol{\Gamma}\bar{\boldsymbol{z}} < 0$. Consider $\lambda\bar{\boldsymbol{z}}$ where $\lambda \geq 0$ and let $\lambda \uparrow +\infty$, we find

$$\alpha - b_k(x) + \lambda(\boldsymbol{\beta} - \boldsymbol{a}_k(\boldsymbol{x}))'\bar{\boldsymbol{z}} + \lambda^2\boldsymbol{\Gamma} \bullet (\bar{\boldsymbol{z}}\bar{\boldsymbol{z}}') < 0,$$

a contradiction.

A projection theorem

Consider ambiguity set without support constraints

$$\mathbb{P} \in \mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^{I_z}) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{z}} \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{z}}\right] = \boldsymbol{\mu} \\ \mathbb{E}_{\mathbb{P}}\left[(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma} \end{array} \right\}$$

Theorem 1 (A projection theorem[8])

For any $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^{I_z}$ and $\boldsymbol{\Sigma} \succeq \boldsymbol{0}$, the projection from space of random vector $\tilde{\boldsymbol{z}}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, defined by

$$\tilde{\boldsymbol{z}} \mapsto \tilde{z}_x = \boldsymbol{x}'\tilde{\boldsymbol{z}}$$

to univariate random variables with mean $\mu_x = \boldsymbol{x}'\boldsymbol{\mu}$ and variance $\sigma_x^2 = \boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}$ is onto.

Let $\mathcal{F}_x = \left\{\mathbb{P} \in \mathcal{P}(\mathbb{R}) : \mathbb{E}_{\mathbb{P}}\left[\tilde{z}_x\right] = \mu_x, \mathbb{E}_{\mathbb{P}}\left[(\tilde{z}_x - \mu_x)^2\right] = \sigma_x^2\right\}$

$$\sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x}'\tilde{\boldsymbol{z}})\right] = \sup_{\mathbb{P} \in \mathcal{F}_x} \mathbb{E}_{\mathbb{P}}\left[f(\tilde{z}_x)\right]$$

[8] Ioana Popescu. "Robust mean-covariance solutions for stochastic optimization". In: *Operations*

Moment based ambiguity set

- Mean, covariance, support[9]

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^{I_z}) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{z}} \sim \mathbb{P} \\ (\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}}] - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}}] - \boldsymbol{\mu}) \leq \lambda_1 \\ \mathbb{E}_{\mathbb{P}}[(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})'] \preceq \lambda_2 \boldsymbol{\Sigma} \\ \mathbb{P}[\tilde{\boldsymbol{z}} \in \mathcal{W}] = 1 \end{array} \right\} \tag{3}$$

where $\mathcal{W}$ is closed convex, $\lambda_1 \geq 0$, $\lambda_2 \geq 1$, $\boldsymbol{\Sigma} \succ \mathbf{0}$. Assume $\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \boldsymbol{z})] < +\infty$ for all $\mathbb{P} \in \mathcal{F}$.

  - Theoretical complexity
    - complexity of inner maximization: $\Psi(\boldsymbol{x}, \lambda_1, \lambda_2) := \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{x}, \tilde{\boldsymbol{z}})]$
    - complexity of whole DRO
  - Justification: high confidence of $\mathbb{P} \in \mathcal{F}$ in data-driven setting
    - Concentration inequalities

[9] Erick Delage and Yinyu Ye. "Distributionally robust optimization under moment uncertainty with application to data-driven problems". In: *Operations research* 58.3 (2010), pp. 595–612.

Equivalent formulation of inner maximization

$$
\begin{array}{lll}
\inf\limits_{r,t,\boldsymbol{q},\boldsymbol{\Gamma}} & r+t & \\
\text{s.t.} & r \geq f(\boldsymbol{x},\tilde{\boldsymbol{z}}) - \boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z} - \boldsymbol{q}'\boldsymbol{z}, \quad \forall \boldsymbol{z} \in \mathcal{W} \\
& t \geq \sqrt{\lambda_1}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{q}+\mathbf{2}\boldsymbol{\Gamma}\boldsymbol{\mu})\|_2 + \boldsymbol{q}'\boldsymbol{\mu} + (\lambda_2\boldsymbol{\Sigma}+\boldsymbol{\mu}'\boldsymbol{\mu})\bullet\boldsymbol{\Gamma} \\
& \boldsymbol{\Gamma} \succeq \mathbf{0}
\end{array}
\tag{Inner-dual}
$$

Proof: By conic duality,

$$
\begin{array}{ll}
\inf & \alpha + \boldsymbol{\Gamma}\bullet\lambda_2\boldsymbol{\Sigma} \\
\text{s.t.} & f(\boldsymbol{x},\tilde{\boldsymbol{z}}) \leq \alpha - \beta_0\sqrt{\lambda_1} + \boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z}-\boldsymbol{\mu}) + \boldsymbol{\Gamma}\bullet(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})' \\
& \quad \forall \boldsymbol{z} \in \mathcal{W} \\
& \boldsymbol{\Gamma} \succeq \mathbf{0}, \beta_0 \geq \|\boldsymbol{\beta}\|_2
\end{array}
$$

Note $\beta_0 = \|\boldsymbol{\beta}\|_2$, let $\boldsymbol{q} = -(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta} + 2\boldsymbol{\Gamma}\boldsymbol{\mu})$ and rearrange terms.
Strong duality holds since Dirac measure $\delta_{\boldsymbol{\mu}} \in \mathrm{ri}(\mathcal{F})$[10] .

[10] Alexander Shapiro. "On duality theory of conic linear problems". In: *Semi-infinite programming*. Springer, 2001, pp. 135–165.

A theoretical complexity result

Lemma 1 (Complexity via ellipsoid method[11])

Consider a convex optimization problem of the form $\min_{\boldsymbol{z}\in\mathcal{Z}} \boldsymbol{c}'\boldsymbol{z}$ *with linear objective and convex feasible set* $\mathcal{Z}$*. Given that the set of optimal solutions is non-empty, the problem can be solved to any precision* ϵ *in time polynomial in* $\log\left(\frac{1}{\epsilon}\right)$ *and in the size of the problem by using the ellipsoid method if and only if* $\mathcal{Z}$ *satisfies the following two conditions :*

1. *for any* $\boldsymbol{z}$*, it can be verified whether* $\boldsymbol{z}$ *or not in time polynomial in the dimension of* $\boldsymbol{z}$*;*
2. *for any infeasible* $\boldsymbol{z}$*, a hyperplane that separates* $\boldsymbol{z}$ *from the feasible set* $\mathcal{Z}$ *can be generated in time polynomial in the dimension of* $\boldsymbol{z}$*.*

[11] Martin Grötschel, László Lovász, and Alexander Schrijver. "The ellipsoid method and its consequences in combinatorial optimization". In: *Combinatorica* 1.2 (1981), pp. 169–197.

Assumptions

1. The support set $\mathcal{W} \subseteq \mathcal{R}^{I_z}$ is convex and compact, and it is equipped with an oracle that can for any $\boldsymbol{z} \in \mathbb{R}^{I_z}$ either confirm that $\boldsymbol{z} \in \mathcal{W}$ or provide a hyperplane that separates $\boldsymbol{z}$ from $\mathcal{W}$ in time polynomial in I_z.
2. The function $f(\boldsymbol{x}, \boldsymbol{z})$ has the form $f(\boldsymbol{x}, \boldsymbol{z}) = \max_{k \in [K]} f_k(\boldsymbol{x}, \boldsymbol{z})$ such that for each k, $f_k(\boldsymbol{x}, \boldsymbol{z})$ is concave in $\boldsymbol{z}$. In addition, given a pair $(\boldsymbol{x}, \boldsymbol{z})$, it is assumed that one can in polynomial time:
   (1) evaluate the value of $f_k(\boldsymbol{x}, \boldsymbol{z})$ in $\boldsymbol{z}$;
   (2) find a super-gradient of $f_k(\boldsymbol{x}, \boldsymbol{z})$ in $\boldsymbol{z}$.

   Furthermore, for any $\boldsymbol{x}, \boldsymbol{q}$, and any $\boldsymbol{\Gamma} \succeq \boldsymbol{0}$, the set $\{y \in \mathbb{R} : y \leq f(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{q}'\boldsymbol{z} - \boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z}\}$ is closed.

Complexity cont.

Lemma 2

Let function $f(\boldsymbol{x}, \boldsymbol{z})$ be concave in $\tilde{\boldsymbol{z}}$ and be such that one can provide a super-gradient of $\boldsymbol{z}$ in time polynomial in I_z. Then, under Assumption 1, for any fixed assignment $\boldsymbol{x}, \boldsymbol{q}, \boldsymbol{\Gamma} \succeq \boldsymbol{0}$ one can find an assignment $\boldsymbol{z}$ that is ϵ-optimal with respect to the problem

$$\begin{array}{ll} \max\limits_{t,\boldsymbol{z}} & t \\ \text{s.t.} & t \leq f(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z} - \boldsymbol{q}'\boldsymbol{z} \\ & z \in \mathcal{W} \end{array}$$

in time polynomial in $\log\left(\frac{1}{\epsilon}\right)$ and the size of the problem.

Proof sketch:

- Assumption 1: $\mathcal{W}$ satisfies conditions of Lemma 1
- Assumption 2: For $(t, \boldsymbol{z})$, the condition of Lemma 1 satisfies for constraint $t \leq f(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z} - \boldsymbol{q}'\boldsymbol{z}$.

Complexity of inner maximization

Proposition 1

Suppose $\mathcal{W}$ satisfies assumption 1 and $f(\boldsymbol{x}, \boldsymbol{z})$ satisfies assumption 2, then the problem Inner-dual is a convex optimization problem and its optimal value $\Psi(\boldsymbol{x}; \lambda_1, \lambda_2)$ can be solved to within any accuracy ϵ in time polynomial in $\log\left(\frac{1}{\epsilon}\right)$ and problem size.

Proof sketch:

- The convexity is easy from the formulation of Inner-dual.
- The problem Inner-dual is feasible.
- The problem Inner-dual is bounded below. Hence $\Psi(\boldsymbol{x}; \lambda_1, \lambda_2)$ is finite and the set of optimal solutions to problem Inner-dual must be non-empty.
- Structure $f(\boldsymbol{x}, \boldsymbol{z}) = \max_{k \in [K]} f_k(\boldsymbol{x}, \boldsymbol{z})$ allows decoupling the first constraint in Inner-dual into K constraints satisfying conditions of Lemma 1 (Implication of Lemma 2). Apply Lemma 1 to Inner-dual to conclude.

Complexity of DRO

Assumptions:

3. The set $\mathcal{X} \subseteq \mathbb{R}^{I_x}$ is convex and compact, and it is equipped with an oracle that can for any $\boldsymbol{x} \in \mathbb{R}^{I_x}$ either confirm that $\boldsymbol{x} \in \mathcal{X}$ or provide a hyperplane that separates $\boldsymbol{x}$ from $\mathcal{X}$ in time polynomial in I_x.
4. The function $f(\boldsymbol{x}, \boldsymbol{z})$ is convex in $\boldsymbol{x}$. In addition, it is assumed that one can find in polynomial time a sub-gradient of $f(\boldsymbol{x}, \boldsymbol{z})$ in $\boldsymbol{x}$.

Proposition 2

Given that assumptions 1, 2, 3, and 4 hold, then the DRO

$$\min_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}} \left[f(\boldsymbol{x}, \tilde{\boldsymbol{z}}) \right]$$

can be solved to any accuracy ϵ in time polynomial in $\log\left(\frac{1}{\epsilon}\right)$ and the sizes of $\boldsymbol{x}$ and $\boldsymbol{z}$.

Proof sketch

Note the DRO is equivalent to problem

$$
\begin{array}{ll}
\min\limits_{\boldsymbol{x},r,t,\boldsymbol{q},\boldsymbol{\Gamma}} & r+t \\
\text{s.t.} & r \geq f_k(\boldsymbol{x},\boldsymbol{z}) - \boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z} - \boldsymbol{q}'\boldsymbol{z}, \quad \forall \boldsymbol{z} \in \mathcal{W}, \forall k \in [K] \\
 & t \geq \sqrt{\lambda_1}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{q}+\boldsymbol{2}\boldsymbol{\Gamma}\boldsymbol{\mu})\|_2 + \boldsymbol{q}'\boldsymbol{\mu} + (\lambda_2\boldsymbol{\Sigma}+\boldsymbol{\mu}'\boldsymbol{\mu}) \bullet \boldsymbol{\Gamma} \\
 & \boldsymbol{\Gamma} \succeq \boldsymbol{0} \\
 & \boldsymbol{x} \in \mathcal{X}
\end{array}
\tag{DRO-dual}
$$

- Assumption 3: $\mathcal{W}$ satisfies conditions of Lemma 1
- Assumption 4: For $(\boldsymbol{x}, r, \boldsymbol{\Gamma}, \boldsymbol{q})$, the condition of Lemma 1 satisfies for constraint $r \geq f_k(\boldsymbol{x},\boldsymbol{z}) - \boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z} - \boldsymbol{q}'\boldsymbol{z}$.

Implications

- Theoretical side
- Practical side: relation between robust optimization and DRO. Deal with

$$\sup_{\boldsymbol{z}\in\mathcal{W}} f_k(\boldsymbol{x},\boldsymbol{z}) - \boldsymbol{z}'\boldsymbol{\Gamma}\boldsymbol{z} - \boldsymbol{q}'\boldsymbol{z} \le r$$

using robust counterpart[12] or cutting plane.
If $f_k(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{z} + b_k(\boldsymbol{x})$, then it is equivalent to

$$\sup_{\boldsymbol{z}\in\mathcal{W}} \begin{bmatrix} 1 & \boldsymbol{z}' \end{bmatrix} \begin{bmatrix} b_k(\boldsymbol{x}) - r & (\boldsymbol{a}_k(\boldsymbol{x}) - \boldsymbol{q})'/2 \\ (\boldsymbol{a}_k(\boldsymbol{x}) - \boldsymbol{q})/2 & -\boldsymbol{\Gamma} \end{bmatrix} \begin{bmatrix} 1 \\ \boldsymbol{z} \end{bmatrix} \le 0$$

If in addition, $\mathcal{W} = \mathbb{R}^{I_z}$, it is equivalent to

$$\begin{bmatrix} r - b_k(\boldsymbol{x}) & (\boldsymbol{q} - \boldsymbol{a}_k(\boldsymbol{x}))'/2 \\ (\boldsymbol{q} - \boldsymbol{a}_k(\boldsymbol{x}))/2 & \boldsymbol{\Gamma} \end{bmatrix} \succeq \boldsymbol{0}$$

[12] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Vol. 28. Princeton University Press, 2009.

What about non-negativity constraints?

If $\mathcal{W} = \mathbb{R}_+^{I_z}$, we have equivalence

$$\inf_{\boldsymbol{z} \geq \boldsymbol{0}} \boldsymbol{\Gamma} \bullet \boldsymbol{z}\boldsymbol{z}' + \boldsymbol{q}'\boldsymbol{z} + r - \boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{z} - b_k(\boldsymbol{x}) \geq 0$$

Recall $\boldsymbol{\Gamma} \succeq \boldsymbol{0}$, we introduce one more matrix variable $\boldsymbol{U}$ and reformulate the problem above:

$$\begin{aligned}
& \inf_{\boldsymbol{z} \geq \boldsymbol{0}, \boldsymbol{U} \succeq \boldsymbol{z}\boldsymbol{z}'} \boldsymbol{\Gamma} \bullet \boldsymbol{U} + \boldsymbol{q}'\boldsymbol{z} + r - \boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{z} - b_k(\boldsymbol{x}) \geq 0 \\
\iff \quad & \inf_{\boldsymbol{z} \geq \boldsymbol{0}, \begin{bmatrix} 1 & \boldsymbol{z}' \\ \boldsymbol{z} & \boldsymbol{U} \end{bmatrix} \succeq \boldsymbol{0}} \boldsymbol{\Gamma} \bullet \boldsymbol{U} + \boldsymbol{q}'\boldsymbol{z} + r - \boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{z} - b_k(\boldsymbol{x}) \geq 0 \\
\iff \quad & \boldsymbol{\xi} \geq \boldsymbol{0}, \begin{bmatrix} r - b_k(\boldsymbol{x}) & (\boldsymbol{q} - \boldsymbol{a}_k(\boldsymbol{x}) - \boldsymbol{\xi})'/2 \\ (\boldsymbol{q} - \boldsymbol{a}_k(\boldsymbol{x}) - \boldsymbol{\xi})/2 & \boldsymbol{\Gamma} \end{bmatrix} \succeq \boldsymbol{0}
\end{aligned}$$

Lifted ambiguity set[13]

In the previous example, lifted ambiguity set is more convenient

Theorem 2 (Lifting Theorem)

Let $\boldsymbol{h} \in \mathbb{R}^m$ and $\boldsymbol{g} : \mathbb{R}^{I_z} \to \mathbb{R}^m$ be a function with a conic representable $\mathcal{K}$-epigraph, and consider the ambiguity set

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^{I_z}) \mid \mathbb{E}_{\mathbb{P}}\left[g(\tilde{\boldsymbol{z}})\right] \preceq_{\mathcal{K}} \boldsymbol{h}, \mathbb{P}\left[\tilde{\boldsymbol{z}} \in \mathcal{W}\right] = 1 \right\}$$

and lifted ambiguity set

$$\mathcal{G} = \left\{ \mathbb{Q} \in \mathcal{P}(\mathbb{R}^{I_z} \times \mathbb{R}^m) \mid \mathbb{E}_{\mathbb{Q}}\left[\tilde{\boldsymbol{u}}\right] = \boldsymbol{h}, \mathbb{Q}\left[\tilde{\boldsymbol{z}} \in \mathcal{W}, g(\tilde{\boldsymbol{z}}) \preceq_{\mathcal{K}} \tilde{\boldsymbol{u}}\right] = 1 \right\},$$

then $\mathcal{F} = \prod_{\tilde{\boldsymbol{z}}} (\mathcal{G})$.

- For any $\mathbb{P} \in \mathcal{F}$, let $\tilde{\boldsymbol{z}} \sim \mathbb{P}$, consider $\tilde{\boldsymbol{u}} = \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{h} - \boldsymbol{g}(\tilde{\boldsymbol{z}})\right] + \boldsymbol{g}(\tilde{\boldsymbol{z}})$ and let $(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{u}}) \sim \mathbb{Q}$, then $\mathbb{Q} \in \mathcal{G}$.
- For any $\mathbb{Q} \in \mathcal{G}$, let $(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{u}}) \sim \mathbb{Q}$, then $\mathbb{E}_{\mathbb{P}}\left[g(\tilde{\boldsymbol{z}})\right] \preceq_{\mathcal{K}} \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{u}}\right] = \boldsymbol{h}$

[13] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. "Distributionally robust convex optimization". In: *Operations Research* 62.6 (2014), pp. 1358–1376.

Lifted linear decision rule

Suppose

$$f(\boldsymbol{x}, \tilde{\boldsymbol{z}}) = \begin{array}{ll} \min & \boldsymbol{d}'\boldsymbol{y} \\ \text{s.t.} & \boldsymbol{A}(\tilde{\boldsymbol{z}})\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} \geq \boldsymbol{b}(\tilde{\boldsymbol{z}}) \end{array}$$

Then

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right] = \begin{array}{l} \min_{\boldsymbol{y}(\tilde{\boldsymbol{z}})} \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{d}'\boldsymbol{y}(\tilde{\boldsymbol{z}})\right] \\ \text{s.t.} \boldsymbol{A}(\tilde{\boldsymbol{z}})\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y}(\tilde{\boldsymbol{z}}) \geq \boldsymbol{b}(\tilde{\boldsymbol{z}}), \forall \tilde{\boldsymbol{z}} \in \mathcal{W} \end{array}$$

where $\boldsymbol{y}(\cdot)$ is a decision rule dependent on $\tilde{\boldsymbol{z}}$.

- Lifted linear decision rule[14]: Restrict

$$\boldsymbol{y}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{u}}) = \boldsymbol{y}^0 + \sum_{i\in I_z} \tilde{z}_i \boldsymbol{y}^i + \sum_{j=1}^{m} \tilde{u}_j \boldsymbol{y}_{j+I_z}$$

- Under restrictive assumpiton like complete recourse and one-dim recourse decision y, it is optimal[15].

[14] Angelos Georghiou, Wolfram Wiesemann, and Daniel Kuhn. "Generalized decision rule approximations for stochastic programming via liftings". In: *Mathematical Programming* 152.1-2 (2015), pp. 301–338.

[15] Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. "Adaptive distributionally robust optimization". In: *Management Science* 65.2 (2019), pp. 604–618.

Short comment

- Tractability:
  - size of SP grows with sample size while size of DRO preserves the problem size of its deterministic counterpart
  - But this is not true for DRO with statistical distance-based ambiguity set like Wasserstein ball
- Performance:
  - Stochastic setting: known distribution, available sampler
    - SAA wins as long as we can solve it with large enough sample size; also elegant sampling complexity and convergence.
  - Data-driven setting: Unknown distribution, limited data possibly corrupted by noise, etc.
    - DRO often wins but moment-based ambiguity set usually lacks of theoretical guarantee
  - Endogeneous uncertainty: decision affects uncertainty like pricing, promotion may affect distribution of demand
    - Sampling based method usually does not work since we must make decision before sampling
    - DRO can work by modeling the dependence relation, although it typically leads to non-convex formulation.

Outline

Moment

- The design of ambiguity set is problem-dependent.
- Typically we want it to be small by incorporating more probability information so that the DRO is less conservative.
- Q: How to incorporate stochastic independence?
- A: In general, requiring independence would destroy the convexity of ambiguity set and lead to intractability[16].

Consider the entropic dominance ambiguity set[17]:

$$\mathcal{F}_E = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^{I_z}) \,\middle|\, \begin{array}{l} \tilde{\boldsymbol{z}} \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}}] = \boldsymbol{\mu} \\ \ln \mathbb{E}_{\mathbb{P}}[\exp(\boldsymbol{q}'(\tilde{\boldsymbol{z}} - \boldsymbol{\mu}))] \leq \phi(\boldsymbol{q}), \forall \boldsymbol{q} \in \mathbb{R}^{I_z} \\ \mathbb{P}[\tilde{\boldsymbol{z}} \in \mathcal{W}] = 1 \end{array} \right\} \tag{4}$$

where $\phi(\cdot)$ is some convex and twice continuously differentiable function that satisfies $\phi(\mathbf{0}) = 0$ and $\nabla \phi(\mathbf{0}) = \mathbf{0}$. Independence implies $\phi(\boldsymbol{q}) = \sum_{i \in I_z} \phi(q_i)$ is separable.

[16] Grani A Hanasusanto et al. "A distributionally robust perspective on uncertainty quantification and chance constrained programming". In: *Mathematical Programming* 151.1 (2015), pp. 35–62.

[17] Zhi Chen, Melvyn Sim, and Huan Xu. "Distributionally robust optimization with infinitely constrained ambiguity sets". In: *Operations Research* 67.5 (2019), pp. 1328–1344.

Related to covariance-based ambiguity set

Proposition 3

If $\ln \mathbb{E}_{\mathbb{P}}\left[\exp(\boldsymbol{q}'(\tilde{\boldsymbol{z}} - \boldsymbol{\mu}))\right] \leq \phi(\boldsymbol{q}), \forall \boldsymbol{q} \in \mathbb{R}^{I_z}$*, then* $\mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{z}}\right] = \boldsymbol{\mu}$ *and* $\mathbb{E}_{\mathbb{P}}\left[(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})'\right] \preceq \nabla^2 \phi(\boldsymbol{0})$*.*

Proof: Using Taylor's expansion and Tonelli's theorem.

- $\mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{z}}\right] = \boldsymbol{\mu}$
    - Note that $\mathbb{E}_{\mathbb{P}}\left[\tilde{z}_i - \mu_i\right] = \lim_{\lambda \to 0} \frac{\mathbb{E}_{\mathbb{P}}[\exp(\lambda(\tilde{z}_i - \mu_i))] - 1}{\lambda}$
    - $\lim_{\lambda \downarrow 0} \frac{\mathbb{E}_{\mathbb{P}}[\exp(\lambda(\tilde{z}_i - \mu_i))] - 1}{\lambda} \leq \lim_{\lambda \downarrow 0} \frac{\exp(\phi(\lambda \boldsymbol{e}_i)) - 1}{\lambda} = \frac{\partial \phi(\boldsymbol{0})}{\partial q_i} = 0$
    - $\lim_{\lambda \uparrow 0} \frac{\mathbb{E}_{\mathbb{P}}[\exp(\lambda(\tilde{z}_i - \mu_i))] - 1}{\lambda} \geq \lim_{\lambda \uparrow 0} \frac{\exp(\phi(\lambda \boldsymbol{e}_i)) - 1}{\lambda} = \frac{\partial \phi(\boldsymbol{0})}{\partial q_i} = 0$
- $\mathbb{E}_{\mathbb{P}}\left[((\tilde{\boldsymbol{z}} - \boldsymbol{\mu})'\boldsymbol{q})^2\right] \leq \boldsymbol{q}'\nabla^2\phi(\boldsymbol{0})\boldsymbol{q}, \forall \boldsymbol{q} \in \mathbb{R}^{I_z}$
    - $\lim_{\lambda \to 0} \frac{\mathbb{E}_{\mathbb{P}}\left[\exp\left(\lambda \boldsymbol{q}'(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})\right)\right] - \exp(\phi(\lambda \boldsymbol{q}))}{\lambda^2} \leq 0$
    - $\mathbb{E}_{\mathbb{P}}\left[\exp\left(\lambda \boldsymbol{q}'(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})\right)\right] = 1 + 0 + \frac{(\lambda \boldsymbol{q}'(\tilde{\boldsymbol{z}} - \boldsymbol{\mu}))^2}{2} + o(\lambda^2)$
    - $\exp\left(\phi(\lambda \boldsymbol{q})\right) = 1 + 0 + \frac{\lambda^2 \boldsymbol{q}'\nabla^2\phi(\boldsymbol{0})\boldsymbol{q}}{2} + o(\lambda^2)$

Solving DRO with $\mathcal{F}_E$

Consider $\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{F}_E} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x},\tilde{\boldsymbol{z}})\right]$ using constraint generation.

- Start from $\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{F}_E^J} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{x},\tilde{\boldsymbol{z}})\right]$ where $\mathcal{F}_E^J$ is the same as $\mathcal{F}_E$ except replacing $\mathbb{R}^{I_z}$ in $\ln \mathbb{E}_{\mathbb{P}}\left[\exp(\boldsymbol{q}'(\tilde{\boldsymbol{z}}-\boldsymbol{\mu}))\right] \le \phi(\boldsymbol{q}), \forall \boldsymbol{q}\in\mathbb{R}^{I_z}$ by $\{\boldsymbol{q}_1,\boldsymbol{q}_2,\dots,\boldsymbol{q}_J\}$. The finite approximation is equivalent to a problem involving exponential cone:

$$
\begin{array}{ll}
\inf & \alpha + \boldsymbol{\beta}'\boldsymbol{\mu} + \sum_{j\in[J]} \gamma_j \\
\text{s.t.} & \alpha - b_k(\boldsymbol{x}) + \sum_{j\in[J]} l_{kj}(\boldsymbol{q}_j'\boldsymbol{\mu} + \phi(\boldsymbol{q}_j)) - \sum_{j\in[J]} m_{k,j} - t_k \ge 0 \\
& \quad \forall k\in[K] \\
& \boldsymbol{\beta} - \boldsymbol{a}_k(\boldsymbol{x}) - \boldsymbol{r}_k - \sum_{j\in[J]} l_{kj}\boldsymbol{q}_j = \boldsymbol{0}, \quad \forall k\in[K] \\
& \boldsymbol{\gamma} - \boldsymbol{n}_k = \boldsymbol{0}, \quad \forall k\in[K] \\
& (n_{kj}, m_{kj}, l_{kj}) \in \mathcal{K}_{\exp}^*, \quad \forall k\in[K], j\in[J] \\
& (t_k, \boldsymbol{r}_k) \in \mathcal{K}(\mathcal{W})^* \\
& \boldsymbol{\gamma} \ge \boldsymbol{0}, \boldsymbol{x}\in\mathcal{X}
\end{array}
$$

Solving DRO with $\mathcal{F}_E$ cont.

- From the dual solution we can construct the worst-case distribution $\mathbb{P}^\dagger$.

$$
\begin{array}{ll}
\sup & \sum_{k\in[K]} (\boldsymbol{a}_k(\boldsymbol{x})'\boldsymbol{\xi}_k + b_k(\boldsymbol{x})\eta_k) \\
\text{s.t.} & \sum_{k\in[K]} \eta_k = 1 \\
& \sum_{k\in[K]} \boldsymbol{\xi}_k = \boldsymbol{\mu} \\
& \sum_{k\in[K]} \boldsymbol{\zeta}_k \le \boldsymbol{e} \\
& (\eta_k, \boldsymbol{\xi}_k) \in \mathcal{K}(\mathcal{W}) \quad \forall k \in [K] \\
& \Big(\zeta_{kj}, \eta_k, \boldsymbol{q}_j'(\boldsymbol{\xi}_k - \eta_k\boldsymbol{\mu}) - \eta_k\phi(\boldsymbol{q}_j)\Big) \in \mathcal{K}_{\exp} \quad \forall k \in [K], j \in [J]
\end{array}
$$

by $\mathbb{P}^\dagger\left[\tilde{\boldsymbol{z}} = \frac{\boldsymbol{\xi}_k^*}{\eta_k^*}\right] = \eta_k^*$ where $\eta_k^* > 0$.

- Then we solve the separation problem

$$
\max_{\boldsymbol{q}\in\mathbb{R}^{I_z}} \ln \mathbb{E}_{\mathbb{P}^\dagger}\left[\exp(\boldsymbol{q}'(\tilde{\boldsymbol{z}} - \boldsymbol{\mu}))\right] - \phi(\boldsymbol{q})
$$

If the optimal value is positive, we find new $\boldsymbol{q}$ to add, otherwise we restart with different initialization until finding violated constraints or reaching max number of restarting.

Sub-Gaussian estimation

Assume sub-Gaussian random vector $\tilde{\boldsymbol{z}}$, i.e.,

$$\ln \mathbb{E}_{\mathbb{P}} \left[\exp(\boldsymbol{y}'(\tilde{\boldsymbol{z}} - \boldsymbol{\mu})) \right] \leq \frac{1}{2} \boldsymbol{y}' \boldsymbol{\Sigma} \boldsymbol{y}, \quad \forall \boldsymbol{y} \in \mathbb{R}^{I_z}$$

- How to estimate $\boldsymbol{\mu}$ and variance proxy $\boldsymbol{\Sigma}$ given empirical data $\{\boldsymbol{z}_i\}_{i=1}^{N}$?
- Given empirical distribution $\mathbb{P}_N$, we can do

$$\begin{array}{ll} \min\limits_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} & \operatorname{tr}(\boldsymbol{\Sigma}) \\ \text{s.t.} & \ln \sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}' \boldsymbol{z}_i) \leq \boldsymbol{\mu}' \boldsymbol{y} + \frac{1}{2} \boldsymbol{y}' \boldsymbol{\Sigma} \boldsymbol{y} \quad \forall \boldsymbol{y} \in \mathbb{R}^{I_z} \\ & \boldsymbol{\Sigma} \succeq \mathbf{0} \end{array} \tag{5}$$

- Suppose the data points are i.i.d. sampled from a sub-Gaussian distribution $\mathbb{P}$, note the estimation is biased since $\mathbb{E}_{\mathbb{P}} \left[\ln \sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}' \boldsymbol{z}_i) \right] \leq \ln \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}' \boldsymbol{z}_i) \right]$

Sub-Gaussian estimation

- Suppose the data points are i.i.d. sampled from a sub-Gaussian distribution $\mathbb{P}$, note the estimation is biased since $\mathbb{E}_{\mathbb{P}}\left[\ln \sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}'\boldsymbol{z}_i)\right] \leq \ln \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}'\boldsymbol{z}_i)\right]$
- Given empirical distribution $\mathbb{P}_N$, we can do

$$
\begin{array}{ll}
\min\limits_{\boldsymbol{\mu},\boldsymbol{\Sigma}} & \operatorname{tr}(\boldsymbol{\Sigma}) \\
\text{s.t.} & \ln \sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}'\boldsymbol{z}_i) \leq \boldsymbol{\mu}'\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}'\boldsymbol{\Sigma}\boldsymbol{y} \quad \forall \boldsymbol{y} \in \mathbb{R}^{I_z} \\
& \boldsymbol{\Sigma} \succeq \hat{\boldsymbol{\Sigma}}
\end{array}
\tag{6}
$$

- If the underlying distribution $\mathbb{P}$ is NOT sub-Gaussian (note the empirical distribution is bounded, hence sub-Gaussian), we may restrict $\boldsymbol{y}$ in a ball, say $\|\boldsymbol{y}\|_\infty \leq r$. Intuitively it means the probability is concentrated around $\boldsymbol{\mu}$, but do not have very light tail.

Estimation by finite approximation

- Consider starting from finite approximation

$$
\begin{aligned}
\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \quad & \boldsymbol{\Sigma} \bullet \boldsymbol{I} \\
\text{s.t.} \quad & b_j \triangleq 2\ln \textstyle\sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}_j' \boldsymbol{z}_i) \leq \boldsymbol{\Sigma} \bullet \boldsymbol{y}_j \boldsymbol{y}_j' + 2\boldsymbol{\mu}' \boldsymbol{y}_j \quad \forall j \in [J] \\
& \boldsymbol{\Sigma} \succeq \hat{\boldsymbol{\Sigma}}
\end{aligned}
$$

  - A SDP can be solved via interior point with low-rank coefficient (SDPT3) or ADMM
- Then we need to add additional $\boldsymbol{y}_j$ by finding a violated constraint:

$$
\min_{\|\boldsymbol{y}\| \leq r} \quad \tfrac{1}{2}\boldsymbol{y}'\boldsymbol{\Sigma}\boldsymbol{y} + \boldsymbol{\mu}'\boldsymbol{y} - \left(\ln \textstyle\sum_{i=1}^{N} \frac{1}{N} \exp(\boldsymbol{y}'\boldsymbol{z}_i)\right)
$$

  - A non-convex smooth problem: projected gradient or trust-region method
- Repeat until the objective value converges or reaching `maxiter`.

Problem

- Estimation
  - Computation: slow improvement after a few iterations.
  - Biased estimation in contrast to covariance estimation
  - Choice of radius r
- Optimization
  - Extend to two-stage DRO
- Joint estimation and optimization

Thank You! Questions?