Conditional Robust Optimization

Saturday Seminar

Li Chen

May 2022

Outline

Conditional stochastic optimization

Robust optimization to conditional stochastic optimization

Motivation

Stochastic optimization: given uncertain parameters $\tilde{\boldsymbol{v}} \sim \mathbb{P}$, find a decision $\boldsymbol{x} \in \mathcal{X}$ to minimize the expected cost:

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

Motivation

Stochastic optimization: given uncertain parameters $\tilde{\boldsymbol{v}} \sim \mathbb{P}$, find a decision $\boldsymbol{x} \in \mathcal{X}$ to minimize the expected cost:

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

Example 1 (Newsvendor)

A newsvendor decides to order x units in face with stochastic demand $\tilde{v}$ to maximize the expected profit:

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

where c is the unit ordering cost, p is the unit revenue.

Motivation

Stochastic optimization: given uncertain parameters $\tilde{\boldsymbol{v}} \sim \mathbb{P}$, find a decision $\boldsymbol{x} \in \mathcal{X}$ to minimize the expected cost:

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

In practice, the distribution $\mathbb{P}$ is estimated from historical data $\{\hat{\boldsymbol{v}}_k\}_{k \in [N]}$, for example, the empirical distribution $\hat{\mathbb{P}} = \dfrac{1}{N} \displaystyle\sum_{k \in [N]} \delta_{\hat{\boldsymbol{v}}_k}$.

Motivation

Stochastic optimization: given uncertain parameters $\tilde{\boldsymbol{v}} \sim \mathbb{P}$, find a decision $\boldsymbol{x} \in \mathcal{X}$ to minimize the expected cost:

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

In practice, the distribution $\mathbb{P}$ is estimated from historical data $\{\hat{\boldsymbol{v}}_k\}_{k \in [N]}$, for example, the empirical distribution $\hat{\mathbb{P}} = \frac{1}{N} \sum_{k \in [N]} \delta_{\hat{\boldsymbol{v}}_k}$.

However, we might have auxiliary data $\{\hat{\boldsymbol{u}}_k\}_{k \in [N]}$ that can help predict $\tilde{\boldsymbol{v}}$. **Conditional stochastic optimization**: replacing distribution by conditional distribution and solve

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{\mathbb{P}} \left[g(\boldsymbol{x}, \tilde{\boldsymbol{v}}) \mid \tilde{\boldsymbol{u}} = \boldsymbol{u} \right]$$

is motivated to utilize these data for better decision-making.

Conditional stochastic optimization (CSO) framework

- $\tilde{\boldsymbol{u}}$ is observable before decision-making, it is called feature/ covariates/ side information/ explanatory variables/ contexts
- $\tilde{\boldsymbol{v}}$: the uncertain parameters of interest
- Assume $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \sim \mathbb{P}^\star$, then ideally, we would like to solve

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{v}}) \mid \tilde{\boldsymbol{u}} = \boldsymbol{u}\right] \tag{1}$$

  for each given $\boldsymbol{u} \in \mathcal{U}$.
- Equivalently, we write the CSO problem as

$$\begin{array}{lll} \min\limits_{\boldsymbol{x}(\cdot)} & \mathbb{E}_{\mathbb{P}^\star}\left[g(\boldsymbol{x}(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{v}})\right] & \\ \text{s.t.} & \boldsymbol{x}(\boldsymbol{u}) \in \mathcal{X} & \forall \boldsymbol{u} \in \mathcal{U} \end{array} \tag{2}$$

  where $\boldsymbol{x}(\cdot)$ is a mapping from feature to decision.

Examples

Example 1 (Regression)

$$g(\boldsymbol{x}, \boldsymbol{v}) = \|\boldsymbol{x} - \boldsymbol{v}\|_2^2$$

Example 2 (Joint production and procurement problem)

- Feature-demand historical data $\{(\hat{\boldsymbol{u}}_k, \hat{\boldsymbol{v}}_k)\}_{k \in [N]}$
- Production and procurement decision

$$\mathcal{X} = \{(\boldsymbol{x}, \boldsymbol{r}) \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{r}, \boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{r} \geq \boldsymbol{0}\},$$

- Cost function:

$$g(\boldsymbol{x}, \boldsymbol{r}, \boldsymbol{v}) = \boldsymbol{c}^\top \boldsymbol{r} - \sum_{i \in [n_x]} p_i \min\{x_i, v_i\}$$

Comparison to supervised learning

Recall the equivalent form of CSO:

$$\begin{array}{lll} \min\limits_{\boldsymbol{x}(\cdot)} & \mathbb{E}_{\mathbb{P}^\star}\left[g(\boldsymbol{x}(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{v}})\right] & \\ \text{s.t.} & \boldsymbol{x}(\boldsymbol{u}) \in \mathcal{X} & \forall \boldsymbol{u} \in \mathcal{U} \end{array}$$

- Mathematically, CSO can be regarded as a generalized supervised learning problem with more complex objective function and constraints.
- From modeling perspective, since supervised learning focus on **prediction** while CSO focus on **decision-making**, the latter requires more interpretability on the policy $\boldsymbol{x}(\cdot)$.

Comparison to supervised learning

Recall the equivalent form of CSO:

$$\begin{array}{lll} \min\limits_{\boldsymbol{x}(\cdot)} & \mathbb{E}_{\mathbb{P}^\star}\left[g(\boldsymbol{x}(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{v}})\right] & \\ \text{s.t.} & \boldsymbol{x}(\boldsymbol{u}) \in \mathcal{X} & \forall \boldsymbol{u} \in \mathcal{U} \end{array}$$

- Mathematically, CSO can be regarded as a generalized supervised learning problem with more complex objective function and constraints.
- From modeling perspective, since supervised learning focus on **prediction** while CSO focus on **decision-making**, the latter requires more interpretability on the policy $x(\cdot)$.

In business analytics, "prescriptive analytics" refers to CSO while "predictive analytics" refers to supervised learning (prediction).

Solution frameworks [BK20]

In practice, $\mathbb{P}^\star$ is unknown: we only have data $\{(\hat{\boldsymbol{u}}_k, \hat{\boldsymbol{v}}_k)\}_{k\in[N]}$ of feature-outcome pairs.

- Use density estimation to approximate conditional distribution

$$
\begin{aligned}
\mathbb{P}^\star\left[\tilde{\boldsymbol{v}} \mid \tilde{\boldsymbol{u}} = \boldsymbol{u}\right] \quad & \longrightarrow \sum_{k\in[N]} w_k(\boldsymbol{u})\delta_{\hat{\boldsymbol{v}}_k} \\
\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{v}}) \mid \tilde{\boldsymbol{u}} = \boldsymbol{u}\right] \quad & \longrightarrow \min_{\boldsymbol{x}\in\mathcal{X}} \sum_{k\in[N]} w_k(\boldsymbol{u}) g(\boldsymbol{x}, \hat{\boldsymbol{v}}_k)
\end{aligned}
$$

  - Pros: fix $\boldsymbol{u}$, CSO is the same as traditional stochastic optimization
  - Cons: density estimation is difficult for high-dim or small N cases
  - usually called "predict-then-optimize" approach

Solution frameworks [BK20]

In practice, $\mathbb{P}^\star$ is unknown: we only have data $\{(\hat{\boldsymbol{u}}_k, \hat{\boldsymbol{v}}_k)\}_{k\in[N]}$ of feature-outcome pairs.

- Use density estimation to approximate conditional distribution

$$\begin{aligned} \mathbb{P}^\star\left[\tilde{\boldsymbol{v}} \mid \tilde{\boldsymbol{u}} = \boldsymbol{u}\right] \quad &\longrightarrow \sum_{k\in[N]} w_k(\boldsymbol{u})\delta_{\hat{\boldsymbol{v}}_k} \\ \min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{v}}) \mid \tilde{\boldsymbol{u}} = \boldsymbol{u}\right] \quad &\longrightarrow \min_{\boldsymbol{x}\in\mathcal{X}} \sum_{k\in[N]} w_k(\boldsymbol{u}) g(\boldsymbol{x}, \hat{\boldsymbol{v}}_k) \end{aligned}$$

  - Pros: fix $\boldsymbol{u}$, CSO is the same as traditional stochastic optimization
  - Cons: density estimation is difficult for high-dim or small N cases
  - usually called "predict-then-optimize" approach
- Approximate joint distribution $\mathbb{P}^\star$ by $\hat{\mathbb{P}}$ and optimize over a chosen policy set $\mathcal{A}$:

$$\begin{array}{cll} \displaystyle\min_{\boldsymbol{x}(\cdot)\in\mathcal{A}} & \mathbb{E}_{\hat{\mathbb{P}}}\left[g(\boldsymbol{x}(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{v}})\right] & \\ \text{s.t.} & \boldsymbol{x}(\boldsymbol{u}) \in \mathcal{X} & \forall \boldsymbol{u} \in \mathcal{U} \end{array}$$

  - Pros: directly output a policy; fully utilize available data
  - Cons: choice of $\mathcal{A}$ is difficult; more complicated optimization
  - Rmk: same as Empirical Risk Minimization (ERM)

CSO via density estimation

Given $\boldsymbol{u} \in \mathcal{U}$, solve

$$\min_{\boldsymbol{x} \in \mathcal{X}} \sum_{k \in [N]} w_k(\boldsymbol{u}) g(\boldsymbol{x}, \hat{\boldsymbol{v}}_k)$$

with conditional distribution estimation via

- Nadaraya-Watson Kernel density estimation (NW-KDE) [HPB10]

  $$w_k(\boldsymbol{u}) = \frac{K_h(\boldsymbol{u} - \hat{\boldsymbol{u}}_k)}{\sum_{j \in [N]} K_h(\boldsymbol{u} - \hat{\boldsymbol{u}}_j)}$$

  where $K(\cdot)$ is a Kernel function and h is the width hyperparameter.
- Local learning methods such as k-NN
- CART
- Random Forest
- Parametric statistical models + residual distribution

CSO via ERM

The choice of the policy set needs to balance the trade-off of tractability, interpretability, and optimality.

- Affine policy set [BR19]

$$\boldsymbol{x}(\boldsymbol{u}) = \boldsymbol{x}^0 + \boldsymbol{X}\boldsymbol{u}$$

- Decision tree policy set [BDM19]

$$\boldsymbol{x}(\boldsymbol{u}) = \sum_{\ell \in [L]} \boldsymbol{x}^{\ell} \mathbb{I}[\boldsymbol{u} \in \mathcal{U}_{\ell}]$$

  requires solving mixed-integer optimization.

- Reproducing Kernel Hilbert Space (RKHS) [BK22, NP22]

$$\boldsymbol{x}(\boldsymbol{u}) = \sum_{k \in [N]} K(\boldsymbol{u}, \boldsymbol{u}_k) \boldsymbol{x}^k$$

  the constraint $\boldsymbol{x}(\boldsymbol{u}) \in \mathcal{X}$ is difficult to enforce and often ignored.

- more can be explored...

Outline

Conditional stochastic optimization

Robust optimization to conditional stochastic optimization

Motivation for robust optimization

For machine learning or data-driven stochastic optimization, the goal is to obtain a predictor/ decision with good **out-of-sample** performance.

- Although many generalization bounds exist for ERM, it requires large N

Motivation for robust optimization

For machine learning or data-driven stochastic optimization, the goal is to obtain a predictor/ decision with good **out-of-sample** performance.

- Although many generalization bounds exist for ERM, it requires large N
- Distributionally robust optimization (DRO) accounts distribution shift (discrepancy of training set and test set)

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{v}})\right] \Longrightarrow \min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}:d(\mathbb{P},\hat{\mathbb{P}})\leq\Gamma} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{v}})\right]$$

  - focus on Type-I Wasserstein distance [EK18]

    $$d(\mathbb{P}_1,\mathbb{P}_2) = \inf_{\mathbb{Q}\in\mathcal{P}_0(\mathcal{V}\times\mathcal{V})} \left\{\mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{v}}_1-\tilde{\boldsymbol{v}}_2\|\right] : (\tilde{\boldsymbol{v}}_1,\tilde{\boldsymbol{v}}_2)\sim\mathbb{Q} \text{ with marginal } \mathbb{P}_1,\mathbb{P}_2\right\}$$

Motivation for robust optimization

For machine learning or data-driven stochastic optimization, the goal is to obtain a predictor/ decision with good **out-of-sample** performance.

- Although many generalization bounds exist for ERM, it requires large N
- Distributionally robust optimization (DRO) accounts distribution shift (discrepancy of training set and test set)

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{v}})\right] \Longrightarrow \min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}:d(\mathbb{P},\hat{\mathbb{P}})\leq\Gamma} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{v}})\right]$$

  - focus on Type-I Wasserstein distance [EK18]

    $d(\mathbb{P}_1,\mathbb{P}_2) = \inf_{\mathbb{Q}\in\mathcal{P}_0(\mathcal{V}\times\mathcal{V})} \left\{\mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{v}}_1-\tilde{\boldsymbol{v}}_2\|\right] : (\tilde{\boldsymbol{v}}_1,\tilde{\boldsymbol{v}}_2)\sim\mathbb{Q} \text{ with marginal } \mathbb{P}_1,\mathbb{P}_2\right\}$

- Regularization via robustification [GCK17, SAKE19]

Example 3 (Linear regression)

If $\tilde{\boldsymbol{u}}$ and $\tilde{v}$ has no support constraints, then

$$\min_{\boldsymbol{x}\in\mathbb{R}^{n_u}} \sup_{\mathbb{P}:d(\mathbb{P},\hat{\mathbb{P}})\leq\Gamma} \mathbb{E}_{\mathbb{P}}\left[\|\boldsymbol{x}^\top\tilde{\boldsymbol{u}}-\tilde{v}\|\right] \Longleftrightarrow \min_{\boldsymbol{x}\in\mathbb{R}^{n_u}} \mathbb{E}_{\hat{\mathbb{P}}}\left[\|\boldsymbol{x}^\top\tilde{\boldsymbol{u}}-\tilde{v}\|\right]+\Gamma\|[\boldsymbol{x};-1]\|_*$$

Conditional robust optimization (CRO)

- Robust optimization for "predict-then-optimize" model [SWHH21, EPM21, KBL20, BVP21]

$$\min_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P}: d(\mathbb{P}, \hat{\mathbb{P}}(\boldsymbol{u})) \leq \Gamma} \mathbb{E}_{\mathbb{P}} \left[g(\boldsymbol{x}, \tilde{\boldsymbol{v}}) \right]$$

where $\hat{\mathbb{P}}(\boldsymbol{u}) \triangleq \sum_{k \in [N]} w_k(\boldsymbol{u}) \delta_{\hat{\boldsymbol{v}}_k}$

Conditional robust optimization (CRO)

- Robust optimization for "predict-then-optimize" model [SWHH21, EPM21, KBL20, BVP21]

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}:d(\mathbb{P},\hat{\mathbb{P}}(\boldsymbol{u}))\leq\Gamma} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{v}})\right]$$

  where $\hat{\mathbb{P}}(\boldsymbol{u}) \triangleq \sum_{k\in[N]} w_k(\boldsymbol{u})\delta_{\hat{\boldsymbol{v}}_k}$

  [SWHH21] adopts a variance regularization and establish its connection to DRO.

  [EPM21] add additional layer of robustness on the reference distribution $\hat{\mathbb{P}}(\boldsymbol{u})$

  [KBL20] use Wasserstein DRO for the residual distribution with a parametric model

  [BVP21] analyze statistical performance of DRO models on bootstrapping sample sets.

Conditional robust optimization (CRO)

- Robust optimization for "predict-then-optimize" model [SWHH21, EPM21, KBL20, BVP21]

$$\min_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P}: d(\mathbb{P}, \hat{\mathbb{P}}(\boldsymbol{u})) \le \Gamma} \mathbb{E}_{\mathbb{P}} \left[g(\boldsymbol{x}, \tilde{\boldsymbol{v}}) \right]$$

where $\hat{\mathbb{P}}(\boldsymbol{u}) \triangleq \sum_{k \in [N]} w_k(\boldsymbol{u}) \delta_{\hat{\boldsymbol{v}}_k}$

  - The computation is the same as Wasserstein DRO
  - Most focus on statistical consistency

Conditional robust optimization (CRO)

- Robust optimization for "predict-then-optimize" model [SWHH21, EPM21, KBL20, BVP21]

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}:d(\mathbb{P},\hat{\mathbb{P}}(\boldsymbol{u}))\leq\Gamma} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{v}})\right]$$

  where $\hat{\mathbb{P}}(\boldsymbol{u}) \triangleq \sum_{k\in[N]} w_k(\boldsymbol{u})\delta_{\hat{\boldsymbol{v}}_k}$

  - The computation is the same as Wasserstein DRO
  - Most focus on statistical consistency
- Robust optimization for "ERM" model [CSZZ22]
  - lack of study

The robust ERM model

Consider solving a DRO model over a chosen policy set $\mathcal{A}$:

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}(\cdot)\in\mathcal{A}} & \sup\limits_{\mathbb{P}\in\mathcal{F}(\Gamma)} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}(\tilde{\boldsymbol{u}}),\tilde{\boldsymbol{v}})\right] \\ \text{s.t.} & \boldsymbol{x}(\boldsymbol{u})\in\mathcal{X}, \forall \boldsymbol{u}\in\mathcal{U} \end{array} \tag{RERM}$$

where the type-I Wassterstein ambiguity set is

$$\mathcal{F}(\Gamma) \triangleq \left\{ \mathbb{P}\in\mathcal{P}_0(\mathcal{Z}) \;\middle|\; \begin{array}{l} (\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}})\sim\mathbb{P} \\ \Delta(\mathbb{P},\hat{\mathbb{P}})\leq\Gamma \end{array} \right\},$$

$\mathcal{Z}=\mathcal{U}\times\mathcal{V}$ and

$$\Delta(\mathbb{P},\hat{\mathbb{P}}) \triangleq \inf_{\mathbb{Q}\in\mathcal{P}_0(\mathcal{Z}^2)} \left\{ \mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{u}}-\tilde{\boldsymbol{u}}_1\|+\|\tilde{\boldsymbol{v}}-\tilde{\boldsymbol{v}}_1\|\right] \;\middle|\; \begin{array}{l} (\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}},\tilde{\boldsymbol{u}}_1,\tilde{\boldsymbol{v}}_1)\sim\mathbb{Q}, \\ (\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}})\sim\mathbb{P}, (\tilde{\boldsymbol{u}}_1,\tilde{\boldsymbol{v}}_1)\sim\hat{\mathbb{P}} \end{array} \right\}.$$

The robust ERM model

Consider solving a DRO model over a chosen policy set $\mathcal{A}$:

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}(\cdot)\in\mathcal{A}} & \sup\limits_{\mathbb{P}\in\mathcal{F}(\Gamma)} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}(\tilde{\boldsymbol{u}}),\tilde{\boldsymbol{v}})\right] \\ \text{s.t.} & \boldsymbol{x}(\boldsymbol{u})\in\mathcal{X}, \forall \boldsymbol{u}\in\mathcal{U} \end{array} \tag{RERM}$$

where the type-I Wassterstein ambiguity set is

$$\mathcal{F}(\Gamma) \triangleq \left\{ \mathbb{P}\in\mathcal{P}_0(\mathcal{Z}) \;\middle|\; \begin{array}{l} (\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}})\sim\mathbb{P} \\ \Delta(\mathbb{P},\hat{\mathbb{P}})\leq\Gamma \end{array} \right\},$$

$\mathcal{Z}=\mathcal{U}\times\mathcal{V}$ and

$$\Delta(\mathbb{P},\hat{\mathbb{P}}) \triangleq \inf_{\mathbb{Q}\in\mathcal{P}_0(\mathcal{Z}^2)} \left\{ \mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{u}}-\tilde{\boldsymbol{u}}_1\|+\|\tilde{\boldsymbol{v}}-\tilde{\boldsymbol{v}}_1\|\right] \;\middle|\; \begin{array}{l} (\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}},\tilde{\boldsymbol{u}}_1,\tilde{\boldsymbol{v}}_1)\sim\mathbb{Q}, \\ (\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}})\sim\mathbb{P}, (\tilde{\boldsymbol{u}}_1,\tilde{\boldsymbol{v}}_1)\sim\hat{\mathbb{P}} \end{array} \right\}.$$

- We focus on linear optimization with recourse:

$$\begin{array}{rll} g(\boldsymbol{x},\boldsymbol{v}) = & \min & \boldsymbol{d}^\top\boldsymbol{y} \\ & \text{s.t.} & \boldsymbol{F}(\boldsymbol{v})\boldsymbol{x}+\boldsymbol{B}\boldsymbol{y}\geq\boldsymbol{f}(\boldsymbol{v}) \\ & & \boldsymbol{y}\in\mathbb{R}^{n_y}, \end{array}$$

The robust ERM model

Consider solving a DRO model over a chosen policy set $\mathcal{A}$:

$$\begin{aligned} \min_{\boldsymbol{x}(\cdot)\in\mathcal{A}} \quad & \sup_{\mathbb{P}\in\mathcal{F}(\Gamma)} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}(\tilde{\boldsymbol{u}}), \tilde{\boldsymbol{v}})\right] \\ \text{s.t.} \quad & \boldsymbol{x}(\boldsymbol{u}) \in \mathcal{X}, \forall \boldsymbol{u} \in \mathcal{U} \end{aligned} \tag{RERM}$$

where the type-I Wassterstein ambiguity set is

$$\mathcal{F}(\Gamma) \triangleq \left\{ \mathbb{P} \in \mathcal{P}_0(\mathcal{Z}) \;\middle|\; \begin{array}{l} (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \sim \mathbb{P} \\ \Delta(\mathbb{P}, \hat{\mathbb{P}}) \leq \Gamma \end{array} \right\},$$

$\mathcal{Z} = \mathcal{U} \times \mathcal{V}$ and

$$\Delta(\mathbb{P}, \hat{\mathbb{P}}) \triangleq \inf_{\mathbb{Q}\in\mathcal{P}_0(\mathcal{Z}^2)} \left\{ \mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{u}} - \tilde{\boldsymbol{u}}_1\| + \|\tilde{\boldsymbol{v}} - \tilde{\boldsymbol{v}}_1\|\right] \;\middle|\; \begin{array}{l} (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{v}}_1) \sim \mathbb{Q}, \\ (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \sim \mathbb{P}, (\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{v}}_1) \sim \hat{\mathbb{P}} \end{array} \right\}.$$

- We focus on linear optimization with recourse:

$$\begin{aligned} g(\boldsymbol{x}, \boldsymbol{v}) = \quad \min \quad & \boldsymbol{d}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{F}(\boldsymbol{v})\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} \geq \boldsymbol{f}(\boldsymbol{v}) \\ & \boldsymbol{y} \in \mathbb{R}^{n_y}, \end{aligned}$$

- Polyhedral feasibility set

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^{n_x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$$

Adaptive DRO perspective

- Technically, we can regard Problem (RERM) as a three-stage robust optimization problem:

$$\boldsymbol{w} \to \tilde{\boldsymbol{u}} \to \boldsymbol{x}(\tilde{\boldsymbol{u}}) \to \tilde{\boldsymbol{v}} \to \boldsymbol{y}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$$

  where $\boldsymbol{w}$ are auxiliary decisions that are independent of uncertainty.
- Standard duality approach leads to the DRO reformulation [EK18]

$$\begin{array}{lll} \min & \kappa\Gamma + \dfrac{1}{N}\displaystyle\sum_{k\in[N]} t_k & \\ \text{s.t.} & \displaystyle\sup_{(\boldsymbol{u},\boldsymbol{v})\in\mathcal{Z}} \{g(\boldsymbol{x}(\boldsymbol{u}),\boldsymbol{v}) - \kappa(\|\boldsymbol{v}-\hat{\boldsymbol{v}}_k\| + \|\boldsymbol{u}-\hat{\boldsymbol{u}}_k\|)\} \le t_k & \forall k\in[N] \\ & \boldsymbol{x}(\boldsymbol{u}) \in \mathcal{X} & \forall \boldsymbol{u}\in\mathcal{U} \\ & \boldsymbol{x}\in\mathcal{A}, \kappa \ge 0. & \end{array} \tag{3}$$

Scenario-wise lifted equivalence

Recall that

$$g(\boldsymbol{x},\boldsymbol{v}) = \begin{array}{ll} \min & \boldsymbol{d}^\top \boldsymbol{y} \\ \text{s.t.} & \boldsymbol{F}(\boldsymbol{v})\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} \geq \boldsymbol{f}(\boldsymbol{v}) \\ & \boldsymbol{y} \in \mathbb{R}^{n_y}, \end{array}$$

then for each $k \in [N]$:

$$\begin{array}{lll} & g(\boldsymbol{x}(\boldsymbol{u}),\boldsymbol{v}) - \kappa(\|\boldsymbol{v} - \hat{\boldsymbol{v}}_k\| + \|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\|)\} \leq t_k & \\ \iff & \boldsymbol{d}^\top \boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v}) - \kappa(\|\boldsymbol{v} - \hat{\boldsymbol{v}}_k\| + \|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\|)\} \leq t_k & \\ & \boldsymbol{F}(\boldsymbol{v})\boldsymbol{x}(\boldsymbol{u}) + \boldsymbol{B}\boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v}) \geq \boldsymbol{f}(\boldsymbol{v}) & \forall (\boldsymbol{u},\boldsymbol{v}) \in \mathcal{Z} \\ \iff & \boldsymbol{d}^\top \boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) - \kappa(\sigma + \nu) \leq t_k & \forall (\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \in \bar{\mathcal{Z}}_k \\ & \boldsymbol{F}(\boldsymbol{v})\boldsymbol{x}(\boldsymbol{u}) + \boldsymbol{B}\boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \geq \boldsymbol{f}(\boldsymbol{v}) & \forall (\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \in \bar{\mathcal{Z}}_k \end{array}$$

where the lifted uncertainty set is

$$\bar{\mathcal{Z}}_k \triangleq \{(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \in \mathcal{U} \times \mathcal{V} \times \mathbb{R} \times \mathbb{R} \mid \sigma \geq \|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\|, \nu \geq \|\boldsymbol{v} - \hat{\boldsymbol{v}}_k\|\}$$

Last equivalence: given $\boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v})$, construct $\bar{\boldsymbol{y}}_k(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) = \boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v})$;
given $\boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v},\sigma,\nu)$, construct $\bar{\boldsymbol{y}}_k(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v},\|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\|,\|\boldsymbol{v} - \hat{\boldsymbol{v}}_k\|)$.

Tractable approximations

The difficult constraints are

$$\begin{array}{ll} \boldsymbol{d}^\top \boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) - \kappa(\sigma+\nu) \le t_k & \forall(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \in \bar{\mathcal{Z}}_k \\ \boldsymbol{F}(\boldsymbol{v})\boldsymbol{x}(\boldsymbol{u}) + \boldsymbol{B}\boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \ge \boldsymbol{f}(\boldsymbol{v}) & \forall(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \in \bar{\mathcal{Z}}_k \end{array}$$

Note that even $\boldsymbol{x}(\cdot)$ is affine, we have bilinear optimization to solve.

- First approximation: Bi-affine approx of $\boldsymbol{y}_k(\cdot)$: Let

$$\begin{aligned} \boldsymbol{y}_k(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) = \quad & \boldsymbol{y}_k^0 + \sum_{i\in[n_u]} \boldsymbol{y}_k^i u_i + \sum_{j\in[n_v]} \boldsymbol{y}_k^{n_u+j} v_j + \boldsymbol{y}_k^{n_u+n_v+1}\sigma \\ & + \boldsymbol{y}_k^{n_u+n_v+2}\nu + \sum_{i\in[n_u]}\sum_{j\in[n_v]} \boldsymbol{y}_k^{ij} u_i v_j \end{aligned}$$

- It remains to approximate bilinear constraints of the form:

$$\boldsymbol{q}_1^\top \boldsymbol{u} + \boldsymbol{q}_2^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{Q}_3 \boldsymbol{u} \le q_4\sigma + q_5\nu + q_6, \quad \forall(\boldsymbol{u},\boldsymbol{v},\sigma,\nu) \in \bar{\mathcal{Z}}_k \quad (4)$$

Note q_4 and q_5 are necessarily non-negative.

Bilinear optimization via adaptive robust optimization

Assume $\mathcal{U}$ and $\mathcal{V}$ are non-empty bounded polytope, for simplicity consider

$$\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{n_u} \mid \underline{\boldsymbol{u}} \leq \boldsymbol{u} \leq \bar{\boldsymbol{u}}\}, \quad \mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^{n_v} \mid \underline{\boldsymbol{v}} \leq \boldsymbol{v} \leq \bar{\boldsymbol{v}}\},$$

Bilinear optimization via adaptive robust optimization

Assume $\mathcal{U}$ and $\mathcal{V}$ are non-empty bounded polytope, for simplicity consider

$$\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{n_u} \mid \underline{\boldsymbol{u}} \leq \boldsymbol{u} \leq \bar{\boldsymbol{u}}\}, \quad \mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^{n_v} \mid \underline{\boldsymbol{v}} \leq \boldsymbol{v} \leq \bar{\boldsymbol{v}}\},$$

Then we dualize $\boldsymbol{v}$ to obtain the following equivalence

$$\begin{aligned}
& \sup_{(\boldsymbol{u},\boldsymbol{v},\sigma,\nu)\in\bar{\mathcal{Z}}_k} \left\{\boldsymbol{q}_1^\top \boldsymbol{u} + \boldsymbol{q}_2^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{Q}_3 \boldsymbol{u} - q_4\sigma - q_5\nu\right\} \\
= & \sup_{\boldsymbol{u}\in\mathcal{U},\boldsymbol{v}\in\mathcal{V}} \left\{\boldsymbol{q}_1^\top \boldsymbol{u} + \boldsymbol{q}_2^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{Q}_3 \boldsymbol{u} - q_4\|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\| - q_5\|\boldsymbol{v} - \hat{\boldsymbol{v}}_k\|\right\} \\
= & \sup_{\boldsymbol{u}\in\mathcal{U}} \inf_{\overline{\boldsymbol{\rho}}\geq\boldsymbol{0},\underline{\boldsymbol{\rho}}\geq\boldsymbol{0}} \left\{\boldsymbol{q}_1^\top \boldsymbol{u} + \overline{\boldsymbol{\rho}}^\top \overline{\boldsymbol{v}} - \underline{\boldsymbol{\rho}}^\top \underline{\boldsymbol{v}} + (\boldsymbol{q}_2 + \boldsymbol{Q}_3\boldsymbol{u} + \underline{\boldsymbol{\rho}} - \overline{\boldsymbol{\rho}})^\top \hat{\boldsymbol{v}}_k \right. \\
& \qquad \left. - q_4\|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\| : \|\boldsymbol{q}_2 + \boldsymbol{Q}_3\boldsymbol{u} - \overline{\boldsymbol{\rho}} + \underline{\boldsymbol{\rho}}\|_* \leq q_5\right\}.
\end{aligned}$$

Bilinear optimization via adaptive robust optimization

Assume $\mathcal{U}$ and $\mathcal{V}$ are non-empty bounded polytope, for simplicity consider

$$\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{n_u} \mid \underline{\boldsymbol{u}} \le \boldsymbol{u} \le \bar{\boldsymbol{u}}\}, \quad \mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^{n_v} \mid \underline{\boldsymbol{v}} \le \boldsymbol{v} \le \bar{\boldsymbol{v}}\},$$

Then we dualize $\boldsymbol{v}$ to obtain the following equivalence

$$\begin{aligned}
& \sup_{(\boldsymbol{u},\boldsymbol{v},\sigma,\nu)\in\bar{\mathcal{Z}}_k} \left\{\boldsymbol{q}_1^\top \boldsymbol{u} + \boldsymbol{q}_2^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{Q}_3 \boldsymbol{u} - q_4\sigma - q_5\nu\right\} \\
= \quad & \sup_{\boldsymbol{u}\in\mathcal{U},\boldsymbol{v}\in\mathcal{V}} \left\{\boldsymbol{q}_1^\top \boldsymbol{u} + \boldsymbol{q}_2^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{Q}_3 \boldsymbol{u} - q_4\|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\| - q_5\|\boldsymbol{v} - \hat{\boldsymbol{v}}_k\|\right\} \\
= \quad & \sup_{\boldsymbol{u}\in\mathcal{U}} \inf_{\overline{\boldsymbol{\rho}}\ge\boldsymbol{0},\underline{\boldsymbol{\rho}}\ge\boldsymbol{0}} \Big\{\boldsymbol{q}_1^\top \boldsymbol{u} + \overline{\boldsymbol{\rho}}^\top \overline{\boldsymbol{v}} - \underline{\boldsymbol{\rho}}^\top \underline{\boldsymbol{v}} + (\boldsymbol{q}_2 + \boldsymbol{Q}_3\boldsymbol{u} + \underline{\boldsymbol{\rho}} - \overline{\boldsymbol{\rho}})^\top \hat{\boldsymbol{v}}_k \\
& \qquad -q_4\|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\| : \|\boldsymbol{q}_2 + \boldsymbol{Q}_3\boldsymbol{u} - \overline{\boldsymbol{\rho}} + \underline{\boldsymbol{\rho}}\|_* \le q_5\Big\}\,.
\end{aligned}$$

We regard it as an adaptive robust optimization problem with uncertainty $\boldsymbol{u}$ and recourse decision $\overline{\boldsymbol{\rho}}, \underline{\boldsymbol{\rho}}$.

Bilinear optimization via adaptive robust optimization

Assume $\mathcal{U}$ and $\mathcal{V}$ are non-empty bounded polytope, for simplicity consider

$$\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{n_u} \mid \underline{\boldsymbol{u}} \leq \boldsymbol{u} \leq \bar{\boldsymbol{u}}\}, \quad \mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^{n_v} \mid \underline{\boldsymbol{v}} \leq \boldsymbol{v} \leq \bar{\boldsymbol{v}}\},$$

We regard it as an adaptive robust optimization problem with uncertainty $\boldsymbol{u}$ and recourse decision $\overline{\boldsymbol{\rho}}, \underline{\boldsymbol{\rho}}$.

$$\inf_{\overline{\boldsymbol{\rho}}(\boldsymbol{u}) \geq \mathbf{0}, \underline{\boldsymbol{\rho}}(\boldsymbol{u}) \geq \mathbf{0}} \sup_{\boldsymbol{u} \in \mathcal{U}} \{\boldsymbol{q}_1^\top \boldsymbol{u} + \overline{\boldsymbol{\rho}}(\boldsymbol{u})^\top \overline{\boldsymbol{v}} - \underline{\boldsymbol{\rho}}(\boldsymbol{u})^\top \underline{\boldsymbol{v}} - q_4 \|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\| \\ + (\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} + \underline{\boldsymbol{\rho}}(\boldsymbol{u}) - \overline{\boldsymbol{\rho}}(\boldsymbol{u}))^\top \hat{\boldsymbol{v}}_k : \|\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{\rho}}(\boldsymbol{u}) + \underline{\boldsymbol{\rho}}(\boldsymbol{u})\|_* \leq q_5$$

Bilinear optimization via adaptive robust optimization

Assume $\mathcal{U}$ and $\mathcal{V}$ are non-empty bounded polytope, for simplicity consider

$$\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{n_u} \mid \underline{\boldsymbol{u}} \leq \boldsymbol{u} \leq \bar{\boldsymbol{u}}\}, \quad \mathcal{V} = \{\boldsymbol{v} \in \mathbb{R}^{n_v} \mid \underline{\boldsymbol{v}} \leq \boldsymbol{v} \leq \bar{\boldsymbol{v}}\},$$

We regard it as an adaptive robust optimization problem with uncertainty $\boldsymbol{u}$ and recourse decision $\overline{\boldsymbol{\rho}}, \underline{\boldsymbol{\rho}}$.

$$\inf_{\overline{\boldsymbol{\rho}}(\boldsymbol{u}) \geq \boldsymbol{0}, \underline{\boldsymbol{\rho}}(\boldsymbol{u}) \geq \boldsymbol{0}} \sup_{\boldsymbol{u} \in \mathcal{U}} \{\boldsymbol{q}_1^\top \boldsymbol{u} + \overline{\boldsymbol{\rho}}(\boldsymbol{u})^\top \overline{\boldsymbol{v}} - \underline{\boldsymbol{\rho}}(\boldsymbol{u})^\top \underline{\boldsymbol{v}} - q_4 \|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\|$$
$$+ (\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} + \underline{\boldsymbol{\rho}}(\boldsymbol{u}) - \overline{\boldsymbol{\rho}}(\boldsymbol{u}))^\top \hat{\boldsymbol{v}}_k : \|\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{\rho}}(\boldsymbol{u}) + \underline{\boldsymbol{\rho}}(\boldsymbol{u})\|_* \leq q_5$$

We apply affine recourse approx: $\overline{\boldsymbol{\rho}}(\boldsymbol{u}) = \overline{\boldsymbol{p}} + \overline{\boldsymbol{P}}\boldsymbol{u}$, $\underline{\boldsymbol{\rho}}(\boldsymbol{u}) = \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}}\boldsymbol{u}$.

$$\begin{aligned}
&\sup_{\boldsymbol{u} \in \mathcal{U}} \{\boldsymbol{q}_1^\top \boldsymbol{u} + (\overline{\boldsymbol{p}} + \overline{\boldsymbol{P}}\boldsymbol{u})^\top \overline{\boldsymbol{v}} - (\underline{\boldsymbol{p}} + \underline{\boldsymbol{P}}\boldsymbol{u})^\top \underline{\boldsymbol{v}} \\
&\qquad + (\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}}\boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}}\boldsymbol{u})^\top \hat{\boldsymbol{v}}_k - q_4 \|\boldsymbol{u} - \hat{\boldsymbol{u}}_k\|\} \leq q_6 \\
&\sup_{\boldsymbol{u} \in \mathcal{U}} \{\|\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}}\boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}}\boldsymbol{u}\|_*\} \leq q_5 \\
&\inf_{\boldsymbol{u} \in \mathcal{U}} \{\boldsymbol{e}_k^\top (\overline{\boldsymbol{p}} + \overline{\boldsymbol{P}}\boldsymbol{u})\} \geq 0 \qquad \forall i \in [n_v] \\
&\inf_{\boldsymbol{u} \in \mathcal{U}} \{\boldsymbol{e}_k^\top (\underline{\boldsymbol{p}} + \underline{\boldsymbol{P}}\boldsymbol{u})\} \geq 0 \qquad \forall i \in [n_v]
\end{aligned} \tag{5}$$

Maximize dual norm over a polytope

Recall $\mathcal{U}$ is a bounded polytope. However, norm maximization over polytope is difficult in general. Indeed,

$$\max_{\boldsymbol{x}} \left\{ \boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x} : \|\boldsymbol{x}\|_\infty \leq 1 \right\}$$

is NP-hard for general $\boldsymbol{\Lambda} \succeq \mathbf{0}$ [Hås01].

Maximize dual norm over a polytope

Recall $\mathcal{U}$ is a bounded polytope. However, norm maximization over polytope is difficult in general. Indeed,

$$\max_{\boldsymbol{x}} \left\{ \boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x} : \|\boldsymbol{x}\|_\infty \leq 1 \right\}$$

is NP-hard for general $\boldsymbol{\Lambda} \succeq \boldsymbol{0}$ [Hås01].

$$\begin{aligned}
& \sup_{\boldsymbol{u} \in \mathcal{U}} \left\{ \|\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}} \boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}} \boldsymbol{u}\|_* \right\} \leq q_5 \\
\iff \quad & \sup_{\boldsymbol{u} \in \mathcal{U}, \|\boldsymbol{\gamma}\| \leq 1} \left\{ \boldsymbol{\gamma}^\top (\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}} \boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}} \boldsymbol{u}) \right\} \leq q_5 \\
\iff \quad & \sup_{\|\boldsymbol{\gamma}\| \leq 1} \inf_{\overline{\boldsymbol{\theta}} \geq \boldsymbol{0}, \underline{\boldsymbol{\theta}} \geq \boldsymbol{0}} \left\{ \boldsymbol{\gamma}^\top (\boldsymbol{q}_2 - \overline{\boldsymbol{p}} + \underline{\boldsymbol{p}}) + \overline{\boldsymbol{u}}^\top \overline{\boldsymbol{\theta}} - \underline{\boldsymbol{u}}^\top \underline{\boldsymbol{\theta}} : \right. \\
& \qquad \left. (\boldsymbol{Q}_3 - \overline{\boldsymbol{P}} + \underline{\boldsymbol{P}})^\top \boldsymbol{\gamma} = \overline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}} \right\} \leq q_5
\end{aligned}$$

Maximize dual norm over a polytope

Recall $\mathcal{U}$ is a bounded polytope. However, norm maximization over polytope is difficult in general. Indeed,

$$\max_{\boldsymbol{x}} \left\{ \boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x} : \|\boldsymbol{x}\|_\infty \leq 1 \right\}$$

is NP-hard for general $\boldsymbol{\Lambda} \succeq \boldsymbol{0}$ [Hås01].

$$\begin{aligned}
& \sup_{\boldsymbol{u} \in \mathcal{U}} \left\{ \|\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}} \boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}} \boldsymbol{u}\|_* \right\} \leq q_5 \\
\iff \quad & \sup_{\boldsymbol{u} \in \mathcal{U}, \|\boldsymbol{\gamma}\| \leq 1} \left\{ \boldsymbol{\gamma}^\top (\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}} \boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}} \boldsymbol{u}) \right\} \leq q_5 \\
\iff \quad & \sup_{\|\boldsymbol{\gamma}\| \leq 1} \inf_{\overline{\boldsymbol{\theta}} \geq \boldsymbol{0}, \underline{\boldsymbol{\theta}} \geq \boldsymbol{0}} \left\{ \boldsymbol{\gamma}^\top (\boldsymbol{q}_2 - \overline{\boldsymbol{p}} + \underline{\boldsymbol{p}}) + \overline{\boldsymbol{u}}^\top \overline{\boldsymbol{\theta}} - \underline{\boldsymbol{u}}^\top \underline{\boldsymbol{\theta}} : \right. \\
& \qquad \left. (\boldsymbol{Q}_3 - \overline{\boldsymbol{P}} + \underline{\boldsymbol{P}})^\top \boldsymbol{\gamma} = \overline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}} \right\} \leq q_5
\end{aligned}$$

Again, we can apply affine recourse approx of $\overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}$ with uncertainty $\boldsymbol{\gamma}$:

$$\overline{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \boldsymbol{\theta} + \boldsymbol{\Theta}^\top \boldsymbol{\gamma}, \quad \underline{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \boldsymbol{\theta} + \boldsymbol{\Theta}^\top \boldsymbol{\gamma} - (\boldsymbol{Q}_3 - \overline{\boldsymbol{P}} + \underline{\boldsymbol{P}})^\top \boldsymbol{\gamma}$$

Maximize dual norm over a polytope

Recall $\mathcal{U}$ is a bounded polytope. However, norm maximization over polytope is difficult in general. Indeed,

$$\max_{\boldsymbol{x}} \left\{ \boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x} : \|\boldsymbol{x}\|_\infty \leq 1 \right\}$$

is NP-hard for general $\boldsymbol{\Lambda} \succeq \mathbf{0}$ [Hås01].

$$\begin{aligned}
& \sup_{\boldsymbol{u} \in \mathcal{U}} \left\{ \|\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}} \boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}} \boldsymbol{u}\|_* \right\} \leq q_5 \\
\iff & \sup_{\boldsymbol{u} \in \mathcal{U}, \|\boldsymbol{\gamma}\| \leq 1} \left\{ \boldsymbol{\gamma}^\top (\boldsymbol{q}_2 + \boldsymbol{Q}_3 \boldsymbol{u} - \overline{\boldsymbol{p}} - \overline{\boldsymbol{P}} \boldsymbol{u} + \underline{\boldsymbol{p}} + \underline{\boldsymbol{P}} \boldsymbol{u}) \right\} \leq q_5 \\
\iff & \sup_{\|\boldsymbol{\gamma}\| \leq 1} \inf_{\overline{\boldsymbol{\theta}} \geq \mathbf{0}, \underline{\boldsymbol{\theta}} \geq \mathbf{0}} \left\{ \boldsymbol{\gamma}^\top (\boldsymbol{q}_2 - \overline{\boldsymbol{p}} + \underline{\boldsymbol{p}}) + \overline{\boldsymbol{u}}^\top \overline{\boldsymbol{\theta}} - \underline{\boldsymbol{u}}^\top \underline{\boldsymbol{\theta}} : \right. \\
& \qquad \left. (\boldsymbol{Q}_3 - \overline{\boldsymbol{P}} + \underline{\boldsymbol{P}})^\top \boldsymbol{\gamma} = \overline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}} \right\} \leq q_5
\end{aligned}$$

Again, we can apply affine recourse approx of $\overline{\boldsymbol{\theta}}, \underline{\boldsymbol{\theta}}$ with uncertainty $\boldsymbol{\gamma}$:

$$\overline{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \boldsymbol{\theta} + \boldsymbol{\Theta}^\top \boldsymbol{\gamma}, \quad \underline{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \boldsymbol{\theta} + \boldsymbol{\Theta}^\top \boldsymbol{\gamma} - (\boldsymbol{Q}_3 - \overline{\boldsymbol{P}} + \underline{\boldsymbol{P}})^\top \boldsymbol{\gamma}$$

Rmk: For 1-norm $\|\cdot\|_1$ with dual norm $\|\cdot\|_\infty$, one can enumerate $2n_\gamma$ vertices of 1-norm unit ball.

Summary of deriving safe approximations

1. Start from Problem RERM (DRO model)
2. Obtain equivalent robust optimization problem (3) for each $k \in [N]$ by standard duality approach
3. For each $k \in [N]$, we adopt a lifted bi-affine approx of $\boldsymbol{y}(\boldsymbol{u}, \boldsymbol{v}, \sigma, \nu)$
4. Approximate the robust bilinear constraints via adaptive robust optimization (repeatedly use LP duality + affine recourse approx)

Summary of deriving safe approximations

1. Start from Problem RERM (DRO model)
2. Obtain equivalent robust optimization problem (3) for each $k \in [N]$ by standard duality approach
3. For each $k \in [N]$, we adopt a lifted bi-affine approx of $\boldsymbol{y}(\boldsymbol{u}, \boldsymbol{v}, \sigma, \nu)$
4. Approximate the robust bilinear constraints via adaptive robust optimization (repeatedly use LP duality + affine recourse approx)

- Close relation of adaptive robust opt and bilinear opt, some connection of affine recourse approx and reformulation-linearization-technique (RLT) is shown [ZMdM+22]
- Generalize the approx from bilinear opt to multi-linear opt

Conclusion

- Conditional stochastic optimization (CSO) unifies supervised learning and data-driven stochastic optimization
- Most literature focus on statistical properties
- Robustifying CSO brings additional challenges for optimization
- We study a RERM model and derive its tractable convex approximations
  - we use a tree-based piecewise affine policy in [CSZZ22]
  - more numerical study is needed
  - more research on modeling and optimization: $\boldsymbol{u}$ and $\boldsymbol{v}$ are not disjoint, non-linear decision rule with robust opt, endogenous uncertainty...

References I

- Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru, *Optimal prescriptive trees*, INFORMS Journal on Optimization **1** (2019), no. 2, 164–183.
- D. Bertsimas and N. Kallus, *From predictive to prescriptive analytics*, Management Science **66** (2020), no. 3, 1025–1044.
- Dimitris Bertsimas and Nihal Koduri, *Data-driven optimization: A reproducing kernel Hilbert space approach*, Operations Research **70** (2022), no. 1, 454–471.
- Gah-Yi Ban and Cynthia Rudin, *The big data newsvendor: Practical insights from machine learning*, Operations Research **67** (2019), no. 1, 90–108.
- Dimitris Bertsimas and Bart Van Parys, *Bootstrap robust prescriptive analytics*, Mathematical Programming (2021), 1–40.
- Li Chen, Melvyn Sim, Xun Zhang, and Minglong Zhou, *Robust explainable prescriptive analytics*, Available at SSRN 4106222 (2022).

References II

Peyman Mohajerin Esfahani and Daniel Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Mathematical Programming **171** (2018), no. 1-2, 115–166.

Adrián Esteban-Pérez and Juan M Morales, *Distributionally robust stochastic programs with side information based on trimmings*, Mathematical Programming (2021), 1–37.

R. Gao, X. Chen, and A.J Kleywegt, *Wasserstein distributional robustness and regularization in statistical learning*, Available at arXiv:1712.06050, 2017.

Johan Håstad, *Some optimal inapproximability results*, Journal of the ACM (JACM) **48** (2001), no. 4, 798–859.

Lauren Hannah, Warren Powell, and David Blei, *Nonparametric density estimation for stochastic optimization with an observable state variable*, Advances in Neural Information Processing Systems **23** (2010).

References III

Rohit Kannan, Güzin Bayraksan, and James R Luedtke, *Residuals-based distributionally robust optimization with covariate information*, arXiv preprint arXiv:2012.01088 (2020).

Pascal M Notz and Richard Pibernik, *Prescriptive analytics for flexible capacity management*, Management Science **68** (2022), no. 3, 1756–1775.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani, *Regularization via mass transportation*, Journal of Machine Learning Research **20** (2019), no. 103, 1–68.

Prateek R Srivastava, Yijie Wang, Grani A Hanasusanto, and Chin Pang Ho, *On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach*, arXiv preprint arXiv:2110.04855 (2021).

References IV

Jianzhe Zhen, Ahmadreza Marandi, Danique de Moor, Dick den Hertog, and Lieven Vandenberghe, *Disjoint bilinear programming: a two-stage robust optimization perspective*, INFORMS Journal on Computing (2022).