Data-Driven Conditional Robust Optimization

Abhilash Chenreddy Nymisha Bandi Erick Delage

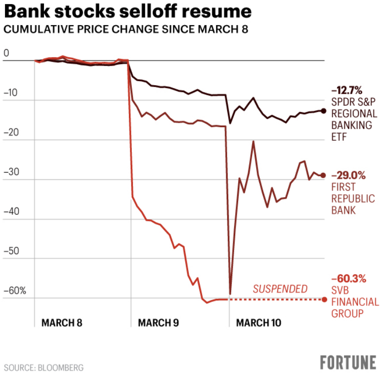
HEC Montréal, GERAD & McGill University
Montréal, Canada

ICSP Tutorial on End-to-end Learning
July 23rd, 2023

Presentation overview

Motivating example

- Returns of different assets are unknown but may depend on historical returns, economic factors, investor sentiments via social media.
- Portfolio manager can formulate an allocation problem to minimize the value-at-risk (VaR) of the portfolio while preserving an expected return above a given target.



What is contextual optimization?

- Optimization problems arising in practice almost always involve **unknown parameters** $\xi \in \mathbb{R}^{m_\xi}$
- Oftentimes, there is a relationship between unknown parameters and some **contextual data** $\psi \in \mathbb{R}^{m_\psi}$

What is contextual optimization?

- Optimization problems arising in practice almost always involve **unknown parameters** $\xi \in \mathbb{R}^{m_\xi}$
- Oftentimes, there is a relationship between unknown parameters and some **contextual data** $\psi \in \mathbb{R}^{m_\psi}$
- **Contextual Optimization**:
  - Optimizes a policy, $\boldsymbol{x} : \mathbb{R}^{m_\psi} \to \mathcal{X}$
    - I.e., action $x \in \mathcal{X}$ is adapted to the observed context ψ
  - Risk Neutral Contextual Optimization problem maximizes the expected cost of running the policy over the joint distribution of (ψ, ξ):

$$\text{(RN-CO)} \qquad \min_{\boldsymbol{x}(\cdot)} \mathbb{E}[c(\boldsymbol{x}(\psi), \xi)]$$

Relation to conditional stochastic optimization

- Interchangeability property (see Theorem 14.60 of Rockafellar and Wets [2009]) states that:

$$\boldsymbol{x}^*(\cdot) \in \arg\min_{\boldsymbol{x}(\cdot)} \mathbb{E}[c(\pi(\psi), \xi)] \quad \Leftrightarrow \quad \boldsymbol{x}^*(\psi) \in \arg\min_{x \in \mathcal{X}} \mathbb{E}[c(x, \xi) | \psi] \text{ a.s.}$$

Relation to conditional stochastic optimization

- Interchangeability property (see Theorem 14.60 of Rockafellar and Wets [2009]) states that:

$$\boldsymbol{x}^*(\cdot) \in \arg\min_{\boldsymbol{x}(\cdot)} \mathbb{E}[c(\pi(\psi), \xi)] \quad \Leftrightarrow \quad \boldsymbol{x}^*(\psi) \in \arg\min_{x \in \mathcal{X}} \mathbb{E}[c(x, \xi)|\psi] \text{ a.s.}$$

- An optimal policy for RNCO problem can therefore be obtained using the following **conditional stochastic optimization (CSO)** problem:

$$\boldsymbol{x}(\psi) := \operatorname*{argmin}_{x \in \mathcal{X}} \mathbb{E}[c(x, \xi)|\psi],$$

What is contextual/conditional robust optimization?

- We introduce a novel **Robust Contextual Optimization** paradigm for solving contextual optimization problems in a risk-averse setting:

$$\text{(Robust-CO)} \qquad \min_{\boldsymbol{x}(\cdot)} \max_{\psi \in \mathcal{V}, \xi \in \mathcal{U}(\psi)} c(\boldsymbol{x}(\psi), \xi)$$

where $\mathcal{U}(\psi)$ is an uncertainty set designed to contain with high probability the realization of ξ conditionally on observing ψ.

What is contextual/conditional robust optimization?

- We introduce a novel **Robust Contextual Optimization** paradigm for solving contextual optimization problems in a risk-averse setting:

$$\text{(Robust-CO)} \qquad \min_{\boldsymbol{x}(\cdot)} \max_{\psi\in\mathcal{V},\xi\in\mathcal{U}(\psi)} c(\boldsymbol{x}(\psi),\xi)$$

  where $\mathcal{U}(\psi)$ is an uncertainty set designed to contain with high probability the realization of ξ conditionally on observing ψ.
- A weaker interchangeability property states:

$$\boldsymbol{x}^*(\cdot) \in \arg\min_{\boldsymbol{x}(\cdot)} \max_{\psi\in\mathcal{V},\xi\in\mathcal{U}(\psi)} c(\pi(\psi),\xi)$$

$$\Leftarrow \quad \boldsymbol{x}^*(\psi) \in \underbrace{\arg\min_{x\in\mathcal{X}} \max_{\xi\in\mathcal{U}(\psi)} c(x,\xi)}_{\text{Conditional Robust Optimization (CRO)}} \quad , \forall\, \psi \in \mathcal{V}$$

Desirable coverage properties for $\mathcal{U}(\psi)$

The field of conformal prediction identifies two important properties for conditional uncertainty sets

- Marginal coverage property: $\mathbb{P}(\xi \in \mathcal{U}(\psi)) \geq 1 - \epsilon$
- Conditional coverage property: $\mathbb{P}(\xi \in \mathcal{U}(\psi)|\psi) \geq 1 - \epsilon$ a.s.
- Conditional coverage $\Rightarrow$ Marginal coverage

E.g., target coverage $1 - \epsilon = 90\%$:

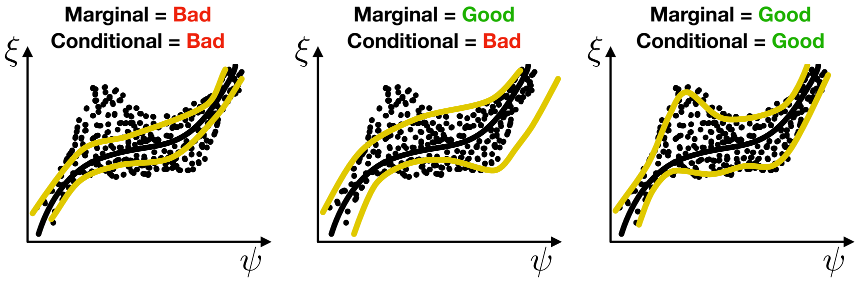


Image from Angelos and Bates, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, CoRR, 2021.

Connection to contextual optimization with VaR I

- Marginal coverage implies that CRO is a conservative approximation to:

$$\text{(Static VaR-CO)} \qquad \min_{\boldsymbol{x}(\cdot)} \text{VaR}_{1-\epsilon}(c(\boldsymbol{x}(\psi), \xi))$$

where $\text{VaR}_{1-\epsilon}(X)$ is the $1-\epsilon$ quantile of X when X is continuous

Connection to contextual optimization with VaR I

- Marginal coverage implies that CRO is a conservative approximation to:

$$\text{(Static VaR-CO)} \qquad \min_{\boldsymbol{x}(\cdot)} \text{VaR}_{1-\epsilon}(c(\boldsymbol{x}(\psi), \xi))$$

  where $\text{VaR}_{1-\epsilon}(X)$ is the $1-\epsilon$ quantile of X when X is continuous

- Proof: Let $\boldsymbol{x}^*_{\text{CRO}}(\cdot)$ be the CRO policy, and $\bar{v} := \text{esssup}\max_{\xi \in \mathcal{U}(\psi)} c(\boldsymbol{x}^*_{\text{CRO}}(\psi), \xi)$ then

$$\begin{aligned}
&\mathbb{P}(c(\boldsymbol{x}^*_{\text{CRO}}(\psi), \xi) \leq \bar{v}) \\
&\qquad \geq \mathbb{P}(\xi \in \mathcal{U}(\psi)) \cdot \mathbb{P}(c(\boldsymbol{x}^*_{\text{CRO}}(\psi), \xi) \leq \bar{v} | \xi \in \mathcal{U}(\psi)) \\
&\qquad \geq (1-\epsilon) \cdot \mathbb{P}\left(\max_{\xi' \in \mathcal{U}(\psi)} c(\boldsymbol{x}^*_{\text{CRO}}(\psi), \xi') \leq \bar{v} \middle| \xi \in \mathcal{U}(\psi) \right) \\
&\qquad \geq (1-\epsilon) \cdot 1 = 1 - \epsilon
\end{aligned}$$

  Hence $\text{VaR}_{1-\epsilon}(c(\boldsymbol{x}^*_{\text{CRO}}(\psi), \xi)) \leq \bar{v}$

Connection to contextual optimization with VaR II

- Conditional coverage implies that CRO is a conservative approximation to:

$$\text{(Nested VaR-CO)} \qquad \min_{\boldsymbol{x}(\cdot)} \mathbb{E}[\ \mathrm{VaR}_{1-\epsilon}(c(\boldsymbol{x}(\psi), \xi)|\psi)\ ]$$

Connection to contextual optimization with VaR II

- Conditional coverage implies that CRO is a conservative approximation to:

$$\text{(Nested VaR-CO)} \qquad \min_{\boldsymbol{x}(\cdot)} \mathbb{E}[\ \mathrm{VaR}_{1-\epsilon}(c(\boldsymbol{x}(\psi), \xi)|\psi)\ ]$$

- Proof: Let $\boldsymbol{x}^*_{\mathrm{CRO}}(\cdot)$ be the CRO policy, and $\bar{v}(\psi) := \max_{\xi\in\mathcal{U}(\psi)} c(\boldsymbol{x}^*_{\mathrm{CRO}}(\psi), \xi)$ then

$$\begin{aligned}
&\mathbb{P}(c(\boldsymbol{x}^*_{\mathrm{CRO}}(\psi), \xi) \leq \bar{v}(\psi)|\psi) \\
&\qquad \geq \mathbb{P}(\xi \in \mathcal{U}(\psi)|\psi) \cdot \mathbb{P}(c(\boldsymbol{x}^*_{\mathrm{CRO}}(\psi), \xi) \leq \bar{v}(\psi)|\psi, \xi \in \mathcal{U}(\psi)) \\
&\qquad \geq (1-\epsilon) \cdot \mathbb{P}(\max_{\xi\in\mathcal{U}(\psi)} c(\boldsymbol{x}^*_{\mathrm{CRO}}(\psi), \xi) \leq \bar{v}(\psi)|\psi, \xi \in \mathcal{U}(\psi)) \\
&\qquad = (1-\epsilon) \cdot 1 = 1-\epsilon
\end{aligned}$$

Hence $\mathrm{VaR}_{1-\epsilon}(c(\boldsymbol{x}^*_{\mathrm{CRO}}(\psi), \xi)|\psi) \leq \bar{v}(\psi)$ a.s. $\Rightarrow$

$$\mathbb{E}[\mathrm{VaR}_{1-\epsilon}(c(\boldsymbol{x}^*_{\mathrm{CRO}}(\psi), \xi|\psi))] \leq \mathbb{E}[\bar{v}(\psi)]$$

Related work in operations research literature

- Conditional Stochastic Optimization:
  - Hannah et al. [2010], Bertsimas and Kallus [2020], ...: Conditional distribution estimation used to formulate and solve the CSO problem.
  - Donti et al. [2017], Elmachtoub and Grigas [2022], ...: End to end paradigm applied to solve the data driven CSO problem.
- Distributionally robust CSO:
  - Bertsimas et al. [2022], McCord [2019], Wang and Jacquillat [2020], Kannan et al. [2020]: DRO approaches with ambiguity sets centered at the estimated conditional distribution
- Data-driven Robust Optimization:
  - Goerigk and Kurtz [2023], Johnstone and Cox [2021]: learns a traditional "non-contextual" uncertainty set using deep learning, and conformal prediction.
  - Ohmori [2021], Sun et al. [2023]: calibrates a box or ellipsoidal set to cover the realizations of a kNN-based or residual-based conditional distribution.
  - Chenreddy et al. [2022] learns a contextual uncertainty set using an integrated clustering then classification approach

Outline

Deep Data-Driven Robust Optimization (DDDRO)

- Classic **non-contextual** RO model is written as

$$\min_{x \in \mathcal{X}} \max_{\xi \in \mathcal{U}} c(x, \xi),$$

Deep Data-Driven Robust Optimization (DDDRO)

- Classic **non-contextual** RO model is written as

$$\min_{x \in \mathcal{X}} \max_{\xi \in \mathcal{U}} c(x, \xi),$$

- Goerigk and Kurtz [2023] describe the uncertainty set $\mathcal{U}$ in the form,

$$\mathcal{U}(W, R) = \{ \xi \in \mathbb{R}^{m_\xi} : \|f_W(\xi) - \bar{f}_0\| \leq R \},$$

where $f_W : \mathbb{R}^{m_\xi} \to \mathbb{R}^d$ is a DNN. The uncertainty set here is defined as a sphere of radius R centered at some $\bar{f}_0$ in the projected space.

Deep Data-Driven Robust Optimization (DDDRO)

- Classic **non-contextual** RO model is written as

$$\min_{x \in \mathcal{X}} \max_{\xi \in \mathcal{U}} c(x, \xi),$$

- Goerigk and Kurtz [2023] describe the uncertainty set $\mathcal{U}$ in the form,

$$\mathcal{U}(W, R) = \left\{ \xi \in \mathbb{R}^{m_\xi} : \|f_W(\xi) - \bar{f}_0\| \leq R \right\},$$

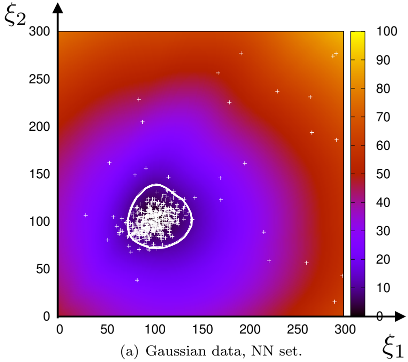
where $f_W : \mathbb{R}^{m_\xi} \to \mathbb{R}^d$ is a DNN. The uncertainty set here is defined as a sphere of radius R centered at some $\bar{f}_0$ in the projected space.

- Given a dataset $\mathcal{D}_\xi = \{\xi_1, \xi_2 \ldots \xi_N\}$, $\mathcal{U}$ is designed by training a NN to minimize the one-class classification loss

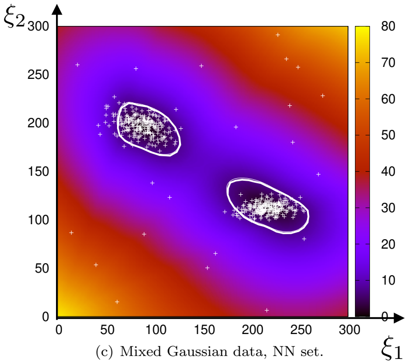
$$\min_W \frac{1}{N} \sum_{i=1}^{N} \|f_W(\xi_i) - \bar{f}_0\|^2,$$

where $\bar{f}_0 := (1/N) \sum_{i \in [N]} f_{W_0}(\xi_i)$ is the center of the projected points and the radius R of $\mathcal{U}$ is calibrated for $1 - \epsilon$ coverage on the data set.

Illustrative examples



(a) Gaussian data, NN set.



(c) Mixed Gaussian data, NN set.

Images from Goerigk and Kurtz. Data-driven robust optimization using deep neural networks. Computers and Operational Research, 151(C), 2023

Solving robust optimization with deep uncertainty sets

- When using piecewise affine activation functions, $\mathcal{U}(W, R)$ can be represented as:

$$\mathcal{U}(W,R) := \left\{ \xi \;\middle|\; \begin{array}{c} \exists u \in \{0,1\}^{d\times K\times L},\ \zeta \in \mathbb{R}^{d\times L},\ \phi \in \mathbb{R}^{d\times L} \\ \sum_{k=1}^{K} u_j^{k,\ell} = 1,\ \forall j,\ell \\ \phi^1 = W^1\xi \\ \zeta_j^\ell = \sum_{k=1}^{K} u_j^{k,\ell} a_k^\ell \phi_j^\ell + \sum_{k=1}^{K} u_j^{k,\ell} b_k^\ell,\ \forall j,\ell \\ \phi^\ell = W^\ell \zeta^{\ell-1},\ \forall \ell \geq 2 \\ \sum_{k=1}^{K} u_j^{k,\ell} \underline{\alpha}_k^\ell \leq \phi_j^\ell \leq \sum_{k=1}^{K} u_j^{k,\ell} \overline{\alpha}_k^\ell,\ \forall j,\ell \\ \|\zeta^L - \bar{f}_0\| \leq R \end{array} \right\}$$

- The problem $\max_{\xi\in\mathcal{U}(W,R)} c(x,\xi)$ can therefore be formulated as a mixed-integer linear program when $c(x,\xi)$ is linear.

Solving robust optimization with deep uncertainty sets

- When using piecewise affine activation functions, $\mathcal{U}(W, R)$ can be represented as:

$$\mathcal{U}(W,R) := \left\{ \xi \;\middle|\; \begin{array}{c} \exists u \in \{0,1\}^{d\times K\times L},\ \zeta \in \mathbb{R}^{d\times L},\ \phi \in \mathbb{R}^{d\times L} \\ \sum_{k=1}^{K} u_j^{k,\ell} = 1,\ \forall j,\ell \\ \phi^1 = W^1\xi \\ \zeta_j^\ell = \sum_{k=1}^{K} u_j^{k,\ell} a_k^\ell \phi_j^\ell + \sum_{k=1}^{K} u_j^{k,\ell} b_k^\ell,\ \forall j,\ell \\ \phi^\ell = W^\ell \zeta^{\ell-1},\ \forall \ell \geq 2 \\ \sum_{k=1}^{K} u_j^{k,\ell} \underline{\alpha}_k^\ell \leq \phi_j^\ell \leq \sum_{k=1}^{K} u_j^{k,\ell} \overline{\alpha}_k^\ell,\ \forall j,\ell \\ \|\zeta^L - \bar{f}_0\| \leq R \end{array} \right\}$$

- The problem $\max_{\xi\in\mathcal{U}(W,R)} c(x,\xi)$ can therefore be formulated as a mixed-integer linear program when $c(x,\xi)$ is linear.
- This can be integrated in a cutting plane method for solving the RO:

$$\begin{array}{rl} \min\limits_{x\in\mathcal{X},t} & t \\ \text{subject to} & c(x,\xi) \leq t,\ \forall \xi \in \mathcal{U}' \subset \mathcal{U}(W,R) \end{array}$$

Outline

Deep Cluster then Classify (DCC)

- We use $\mathcal{D} := \{(\psi_1, \xi_1), \ldots, (\psi_N, \xi_N)\}$ to design data-driven conditional uncertainty sets $\mathcal{U}(\psi)$.
- This approach reduces the side-information ψ to a set of K different clusters and designs customized sets, i.e., $\mathcal{U}(\psi) := \mathcal{U}_{a(\psi)}$
  - $a : \mathbb{R}^{m_\psi} \to [K]$ is a trained K-class cluster assignment function
  - Each $\mathcal{U}_k$, for $k = 1, \ldots, K$, is an uncertainty sets for ξ calibrated on the dataset $\mathcal{D}_\xi^k := \cup_{(\psi,\xi)\in\mathcal{D}:a(\psi)=k}\{\xi\}$ using one-class classification as in Goerigk and Kurtz [2023].

Deep clustering using auto-encoder/decoder networks

We use an auto-encoder and decoder network to identify $a(\cdot)$,

$$\mathcal{L}^1(V,\theta) := \frac{1-\alpha_K}{N}\sum_{i=1}^{N}\|g_{V_D}(g_{V_E}(\psi_i)) - \psi_i\|^2 + \frac{\alpha_K}{N}\sum_{i=1}^{N}\|g_{V_E}(\psi_i) - \theta^{a(\psi_i)}\|^2,$$

where

$$a(\psi) := \operatorname*{argmin}_{k\in[K]} \|g_{V_E}(\psi) - \theta^k\|$$

and V_E and V_D are the network parameters.

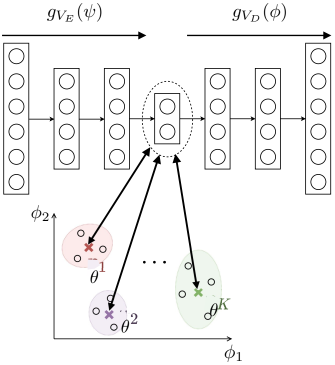


Image adapted from Fard et al. Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognition Letters, 138:185–192, 2020

Deep Cluster then Classify (DCC) shortcomings

1. Fails to tackle the conditional uncertainty set learning problem as a whole
   - I.e., low total variation in the projected ψ space for cluster does not imply low total variation is possible for projections of ξ.
2. Can struggle for cases where clear separation of clusters isn't possible.

Outline

Integrated Deep Cluster then Classify (IDCC)

IDCC addresses the shortcomings of DCC:

1. Optimize V_E, V_D, θ, and $\{W^k\}_{k=1}^K$ jointly using a loss function that trades-off between the objectives used for clustering and each of the K versions of one-class classifiers

Integrated Deep Cluster then Classify (IDCC)

IDCC addresses the shortcomings of DCC:

1. Optimize V_E, V_D, θ, and $\{W^k\}_{k=1}^K$ jointly using a loss function that trades-off between the objectives used for clustering and each of the K versions of one-class classifiers

2. The issue of hard assignment is handled by training a parameterized random assignment policy $\tilde{a}(\psi) \sim \pi(\psi)$:

$$\mathbb{P}(\tilde{a}(\psi) = k) = \pi_k(\psi) := \frac{\exp\{-\beta\|g_V(\psi) - \theta^k\|^2\}}{\sum_{k'=1}^K \exp\{-\beta\|g_V(\psi) - \theta^{k'}\|^2\}}$$

Integrated Deep Cluster then Classify (IDCC)

With these changes, the proposed loss function is of the form,

$$\mathcal{L}^2_\alpha(V,\theta,\{W^k\}_{k=1}^K) := \alpha_S\Big((1-\alpha_K)\mathbb{E}^\pi_{\mathcal{D}}[\|g_{V_D}(g_{V_E}(\psi_i))-\psi_i\|^2] + \alpha_K\mathbb{E}^\pi_{\mathcal{D}}[\text{TotalVar}^\pi_{\mathcal{D}}(g_{V_E}(\psi),\theta^{\tilde{a}(\psi)}\,|\tilde{a}(\psi))]\Big) + (1-\alpha_S)\frac{1}{K}\sum_{k=1}^K \min_{\vartheta^k}\text{TotalVar}^\pi_{\mathcal{D}}(f_{W^k}(\xi),\vartheta^k\,|\tilde{a}(\psi)=k)\,,$$

where $\text{TotalVar}^\pi_{\mathcal{D}}(\phi,\theta|\tilde{a}(\psi)) := \sum_{j=1}^d \mathbb{E}^\pi_{\mathcal{D}}[(\phi_j-\theta_j)^2|\tilde{a}(\psi)]$ is the conditional centered total variation of ϕ given $\tilde{a}(\psi)$.

The **random** uncertainty set is $\tilde{\mathcal{U}}(\psi) := \mathcal{U}(W^{\tilde{a}(\psi)}, R^{\tilde{a}(\psi)})$

IDCC conservatively approximates Value-at-Risk contextual optimization

Lemma

Under DCC and IDCC, if the uncertainty set is calibrated to satisfy:

$$\mathbb{P}^{\pi}_{\mathcal{D}}(\xi \in \tilde{\mathcal{U}}(\psi) | \tilde{a}(\psi) = k) \geq 1 - \epsilon, \ \forall k \,, \tag{1}$$

then the random policy $\tilde{\boldsymbol{x}}(\cdot)$ to the randomized CRO problem together with

$$v^* := \max_{k \in [K]} \min_{x \in \mathcal{X}} \max_{\xi \in \mathcal{U}(W^k, R^k)} c(x, \xi)$$

provide a conservative approximate solution to the VaR-CO problem:

$$\min_{\boldsymbol{x}(\cdot)} VaR^{\mathcal{D},\pi}_{1-\varepsilon}(c(\boldsymbol{x}(\psi), \xi)),$$

under the empirical measure $\mathbb{P}^{\pi}_{\mathcal{D}}$. Namely, $VaR^{\mathcal{D},\pi}_{1-\epsilon}(c(\tilde{\boldsymbol{x}}(\psi), \xi)) \leq v^*$.

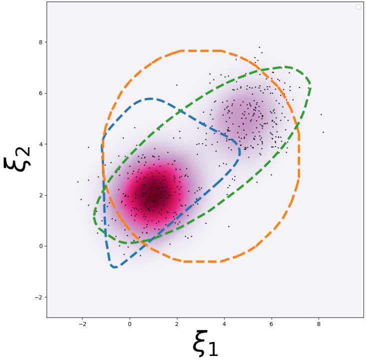
Outline

Experiments

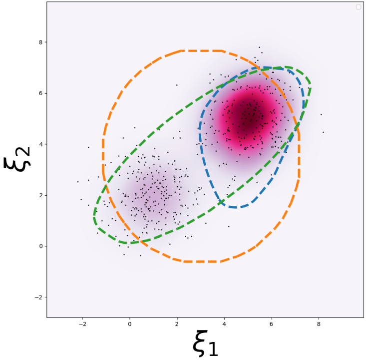
Two environments:

1. Simple Synthetic Environment:
   - $(\psi, \xi) \in \mathbb{R}^2 \times \mathbb{R}^2$ drawn from a mixture of two multivariate Gaussian distributions

Illustration of conditional uncertainty sets in synthetic environment



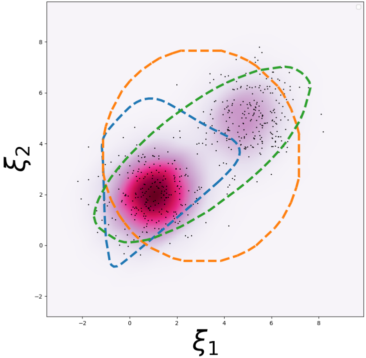
(a) $\tilde{a}(\psi) = 1$, 90% coverage



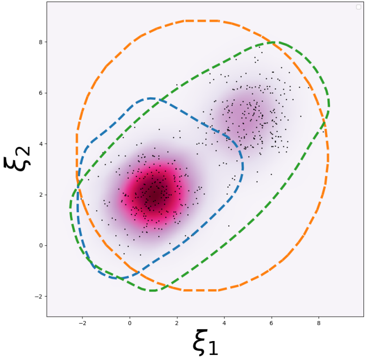
(b) $\tilde{a}(\psi) = 2$, 90% coverage

Ellipsoid DDDRO IDCC

Illustration of conditional uncertainty sets in synthetic environment



(c) $\tilde{a}(\psi) = 1$, 90% coverage



(d) $\tilde{a}(\psi) = 1$, 99% coverage

■ Ellipsoid ■ DDDRO ■ IDCC

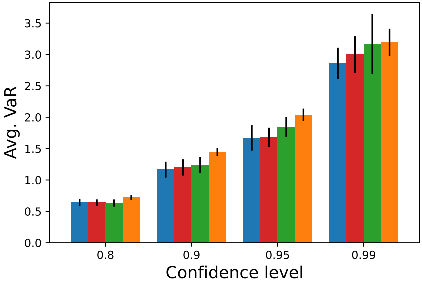
Experiments

Two environments:

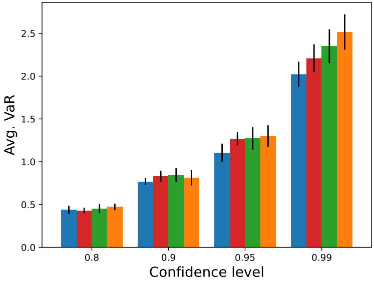
1. Simple Synthetic Environment:
   - $(\psi, \xi) \in \mathbb{R}^2 \times \mathbb{R}^2$ drawn from a mixture of two multivariate Gaussian distributions

2. Robust portfolio optimization with market data
   - Decision model: $\min_{x \in \mathcal{X}} \mathrm{VaR}_{1-\epsilon}(\xi^\intercal x)$ where $\mathcal{X} := \{x \in \mathbb{R}^n | \sum_{i=1}^n x_i = 1,\ x \geq 0\}$ captures the need to invest one unit of wealth among the available assets while minimizing risk exposure.
   - Contextual info: Trading volume of individual stocks, market indices such as volatility index (VIX), 10-year treasury yield index (TNX), oil index (CLF), S&P 500 (GSPC), gold price (GC=F), Dow Jones (DJI) as covariates.
   - Market data from Yahoo! Finance: 70 different stocks from 8 sectors during period from Jan. 1st 2012 to Dec. 31 2019.

Portfolio optimization: Comparison of Avg. VaR across portfolio simulations



(a) 2018



(b) 2019

Ellipsoid DDDRO DCC IDCC

Outline

End-to-End CRO

- The IDCC approach suffers from two issues:
  1. Training is done solely based on total variation measurements, disregarding entirely the out-of-sample performance of the solution obtained from robust optimization.
  2. While the calibration process encourages **marginal coverage** by making the coverage accurate for each cluster:

$$\mathbb{P}(\xi \in \mathcal{U}(\psi) | \tilde{a}(\psi) = k) \geq 1 - \epsilon \forall k \ ✔ \ \Rightarrow \ \mathbb{P}(\xi \in \mathcal{U}(\psi)) \geq 1 - \epsilon \ ✔$$

it does not promote **conditional coverage** over all ψ:

$$\mathbb{P}(\xi \in \mathcal{U}(\psi) | \psi) \geq 1 - \epsilon \text{ a.s. } ✖$$

- In this next part, we propose **End-to-end Conditional Robust Optimization** that promotes **conditional coverage**.

A sequential approach for continuous adaptation

- We consider a continuously adapted conditional ellipsoidal set:

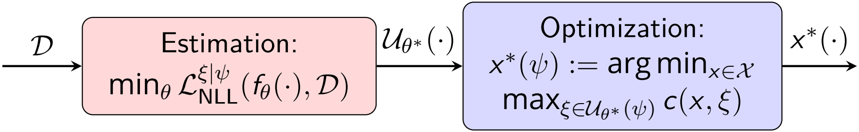
$$\mathcal{U}_\theta(\psi) := \{\ \xi \in \mathbb{R}^{m_\xi} : (\xi - \mu_\theta(\psi))^T \Sigma_\theta^{-1}(\psi)(\xi - \mu_\theta(\psi)) \leq R_\theta \},$$

A sequential approach for continuous adaptation

- We consider a continuously adapted conditional ellipsoidal set:

$$\mathcal{U}_\theta(\psi) := \{ \, \xi \in \mathbb{R}^{m_\xi} : (\xi - \mu_\theta(\psi))^T \Sigma_\theta^{-1}(\psi)(\xi - \mu_\theta(\psi)) \le R_\theta \} \, ,$$

- Given a data set $\mathcal{D} = \{(\psi_1, \xi_1), (\psi_2, \xi_2) \ldots (\psi_N, \xi_N)\}$, a sequential learning and optimization approach takes the form:



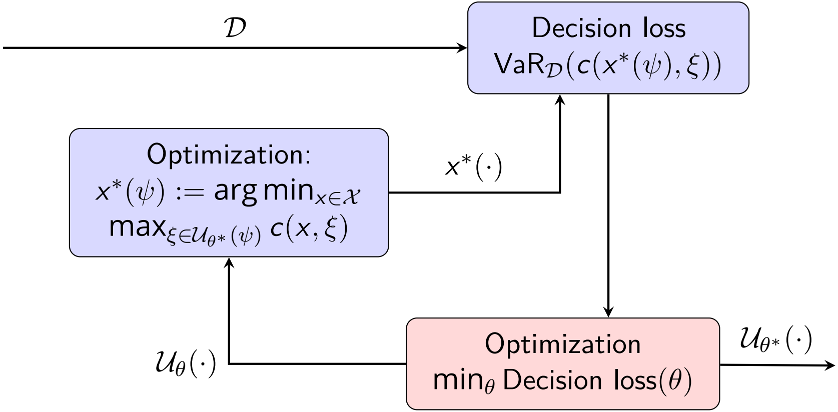
where $\mathcal{L}_{\text{NLL}}^{\xi|\psi}$ is the negative log likelihood for a conditional Gaussian density estimator (see Barratt and Boyd [2021]):

$$\xi \sim f_\theta(\psi) := \mathcal{N}(\mu_\theta(\psi), \Sigma_\theta(\psi))$$

and R_θ s.t. $\mathbb{P}_\mathcal{D}(\xi \in \mathcal{U}_\theta(\psi)) = 1 - \epsilon$

End-to-end objective

An end-to-end approach learns the estimator by trying to minimize the decision loss, e.g. the portfolio risk based on $\text{VaR}_{1-\epsilon}$



Conditional coverage objective I

Lemma

An uncertainty set $\mathcal{U}_\theta(\psi)$ *has an a.s. conditional coverage of* $1-\epsilon$ *if and only if*

$$\mathcal{L}_{CC}(\theta) := \mathbb{E}[\,(\mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) - (1-\epsilon))^2\,] = 0$$

Conditional coverage objective I

Lemma

An uncertainty set $\mathcal{U}_\theta(\psi)$ has an a.s. conditional coverage of $1-\epsilon$ if and only if

$$\mathcal{L}_{CC}(\theta) := \mathbb{E}[\,(\mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) - (1-\epsilon))^2\,] = 0$$

Proof:

$$\begin{aligned}
\text{Condition coverage} \;&\Leftrightarrow\; \mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) = 1-\epsilon \text{ a.s} \\
&\Rightarrow\; \mathcal{L}_{CC}(\theta) = \mathbb{E}[(\mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) - (1-\epsilon))^2] \\
&\qquad\qquad\quad = \mathbb{E}[(1-\epsilon-(1-\epsilon))^2] = 0 \\
&\Rightarrow\; (\mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) - (1-\epsilon))^2 = 0 \text{ a.s.} \\
&\Rightarrow\; \mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) = 1-\epsilon \text{ a.s.}
\end{aligned}$$

Conditional coverage objective II

The loss function $\mathcal{L}_{\text{CC}}(\theta) := \mathbb{E}[(\mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) - (1-\epsilon))^2]$ can be approximated using:

$$\mathcal{L}^3(\theta) := \mathbb{E}^{\mathcal{D}}[(g_{\phi^*(\theta)}(\psi) - (1-\epsilon))^2]$$

where $g_{\phi^*(\theta)}(\psi) \approx \mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi)$ is obtained using logistic regression of membership variable $y(\psi, \xi; \theta) := \mathbb{1}\{\xi \in \mathcal{U}_\theta(\psi)\}$ on ψ.

Conditional coverage objective II

The loss function $\mathcal{L}_{\text{CC}}(\theta) := \mathbb{E}[(\mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi) - (1-\epsilon))^2]$ can be approximated using:

$$\mathcal{L}^3(\theta) := \mathbb{E}^{\mathcal{D}}[(g_{\phi^*(\theta)}(\psi) - (1-\epsilon))^2]$$

where $g_{\phi^*(\theta)}(\psi) \approx \mathbb{P}(\xi \in \mathcal{U}_\theta(\psi)|\psi)$ is obtained using logistic regression of membership variable $y(\psi, \xi; \theta) := \mathbb{1}\{\xi \in \mathcal{U}_\theta(\psi)\}$ on ψ.

- I.e., letting the augmented data set

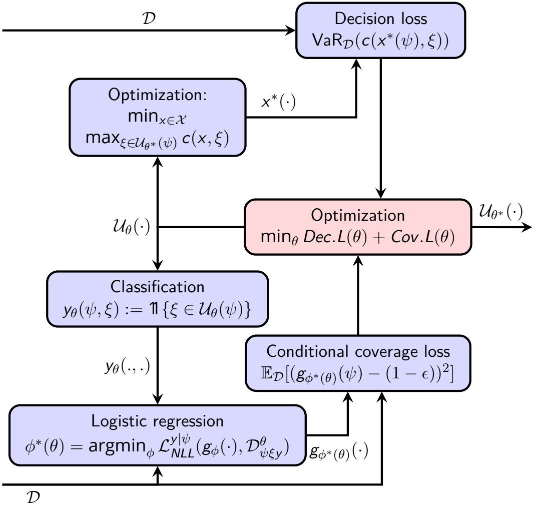
$$\mathcal{D}^\theta_{\psi\xi y} := \{(\psi_1, \xi_1, y(\psi_1, \xi_1; \theta)), \ldots, (\psi_N, \xi_N, y(\psi_N, \xi_N; \theta))\},$$

  one solves $\phi^*(\theta) \in \operatorname{argmin}_\phi \mathcal{L}^{y|\psi}_{NLL}(g_\phi(\cdot), \mathcal{D}^\theta_{\psi\xi y})$ with

$$g_\phi(\psi) := \frac{1}{1 + \exp^{\phi^T\psi + \phi_0}}$$

End-to-end approach with conditional coverage

We train $\mathcal{U}_\theta(\psi)$ using the two tasks: produce good decision + produce good conditional coverage:

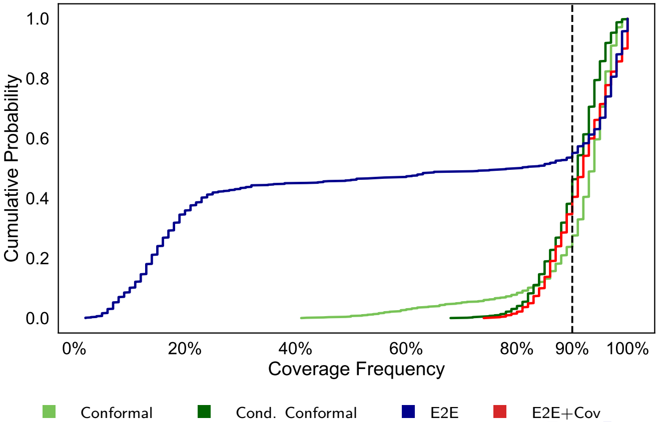


Comparative study

	Conf.	Cond. Conf.	E2E	E2E+cov.
$\text{CVaR}_{0.9}$	1.55	1.47	1.29	**1.24**
$\text{VaR}_{0.9}$	0.91	0.88	0.82	**0.78**
Marginal cov.	**91%**	**90%**	66%	94%

Comparative study

	Conf.	Cond. Conf.	E2E	E2E+cov.
$\text{CVaR}_{0.9}$	1.55	1.47	1.29	**1.24**
$\text{VaR}_{0.9}$	0.91	0.88	0.82	**0.78**
Marginal cov.	**91%**	**90%**	66%	94%



Concluding remarks

- We introduced a new contextual robust optimization approach for solving risk averse contextual optimization problems.
- In Robust-CO, deep neural networks can be used to:
  - Represent richly structured uncertainty sets, e.g. DDDRO, IDCC implementation
  - Adapt uncertainty set continuously to covariates, e.g. E2E+cov. implementation
- Two types of training procedures:
  - IDCC produces sets that are more interpretable but less adaptable
  - E2E+cov. is more obscure but highly adaptive
- Two types of training objectives:
  - Statistical performance: achieving the right marginal/conditional coverage
  - Task-based performance: Producing decisions that achieve low VaR/CVaR

Thank you

Appendix

Bibliography I

Shane Barratt and Stephen Boyd. Covariance prediction via convex optimization, 2021.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

Dimitris Bertsimas, Christopher McCord, and Bradley Sturt. Dynamic optimization with side information. *European Journal of Operational Research*, 2022.

Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. Data-driven conditional robust optimization. In *Advances in Neural Information Processing Systems*, volume 35, pages 9525–9537, 2022.

Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022.

Marc Goerigk and Jannis Kurtz. Data-driven robust optimization using deep neural networks. *Computers and Operational Research*, 151(C), 2023.

Lauren Hannah, Warren Powell, and David Blei. Nonparametric density estimation for stochastic optimization with an observable state variable. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization, 2021.

Bibliography II

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

Christopher George McCord. *Data-driven dynamic optimization with auxiliary covariates*. PhD thesis, Massachusetts Institute of Technology, 2019.

Shunichi Ohmori. A predictive prescription using minimum volume k-nearest neighbor enclosing ellipsoid and robust optimization. *Mathematics*, 9(2):119, 2021.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, Berlin, 2009.

Chunlin Sun, Linyu Liu, and Xiaocheng Li. Predict-then-calibrate: A new perspective of robust contextual lp, 2023.

Kai Wang and Alex Jacquillat. From classification to optimization: A scenario-based robust optimization approach. *Available at SSRN 3734002*, 2020.