# Contextual Scenario Generation for Two-Stage Stochastic Programming

David Islip [1] Sanghyeon Bae [2] Roy H. Kwon [1] Woo Chang Kim [2]

[1]Department of Mechanical and Industrial Engineering, University of Toronto

[2]Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST)

October 28, 2024

# Introduction to Two-Stage Stochastic Programs

Two-stage stochastic programs (2SPs) are a widely adopted decision-making tool that are described by the following setting:

- A 'here-and-now' decision $\boldsymbol{y} \in \mathcal{Y}$ is made, incurring an immediate cost $h(\boldsymbol{y})$
- An uncertain event $\omega \in \Omega$ is then realized, with $\omega$ drawn from distribution $\mathbb{P}$
- Afterward, the decision maker selects a recourse action $\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y}, \omega)$
- The decision maker optimally reacts to $\omega$, given $\boldsymbol{y}$, and chooses a recourse action that minimizes the cost $q(\boldsymbol{z}, \omega)$, where $q$ represents the recourse cost in scenario $\omega$

$$\boldsymbol{y} \longrightarrow \omega \longrightarrow \boldsymbol{z} \quad \text{(2SP ABCs)}$$

- For example, a product manufacturer must allocate capacity to machines ($\boldsymbol{y}$) before knowing product demand ($\omega$)
- Once the product demands are known, they create an optimal production plan ($\boldsymbol{z}$)
- 2SPs are applied in fields such as logistics, portfolio management, and manufacturing (Ntaimo, 2024)

# Introduction to 2SP
## Problem Setup

The decision maker wishes to solve

$$\min_{\boldsymbol{y}} h(\boldsymbol{y}) + \mathbb{Q}(\boldsymbol{y}) \qquad \text{s.t.} \quad \boldsymbol{y} \in \mathcal{Y}, \quad \boldsymbol{y} \in \mathbb{R}^{s_1} \tag{2SP}$$

where

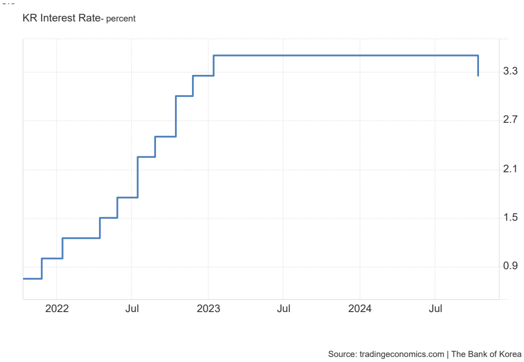
- $\mathbb{Q}(\boldsymbol{y}) = \mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{y}, \omega)]$ is the expected recourse cost

with the optimal recourse cost recourse

$$Q(\boldsymbol{y}, \omega) = \min_{\boldsymbol{z}} q(\boldsymbol{z}, \omega) \quad \text{s.t.} \quad \boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y}, \omega), \quad \boldsymbol{z} \in \mathbb{R}^{s_2} \tag{Stage II}$$

# Contextual Stochastic Optimization

- Suppose the decision maker is manufacturing parts for home builders
- The Bank of Korea has lowered interest rates to 3.25%
- How does this change our decision $\boldsymbol{y}$?
- Information known to the decision maker at decision time is referred to as context $\boldsymbol{x}$
- The decision maker aims to solve 2SP with costs evaluated using $\mathbb{P}_{\omega|\boldsymbol{x}}$

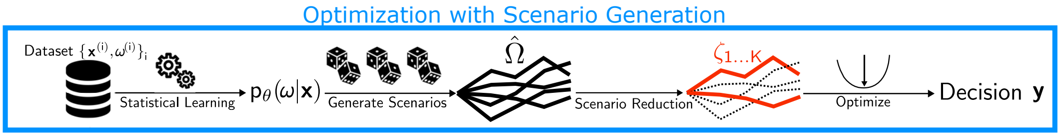


# Contextual Scenario Generation for 2SP

## Problem Setup and an Existing Approach

- They do not have access to $\mathbb{P}_{\omega|\boldsymbol{x}}$, and instead only have access to a joint sample (historical data) $S = \{\boldsymbol{x}^{(i)}, \omega^{(i)}\}_{i=1}^{n}$
- There is a contextual stochastic optimization (CSO) zoo (Estes and Richard, 2023)

This suggests the following approach



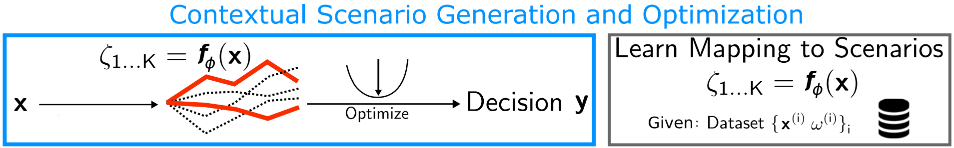

- A dataset of observations is used to estimate a parametric model.
- Sampling from a parametric model of the conditional scenario distribution yields a large 2SP
- The number of scenarios is reduced and the optimization is performed over the reduced set of scenarios

# Contextual Scenario Generation for 2SP
## The main idea

The goal is to directly go from $\boldsymbol{x}$ to a small set of scenarios $\zeta_{1...K} = (\zeta_k)_{k=1}^{K} \in \Omega^K$ by selecting $\boldsymbol{f} : \boldsymbol{x} \mapsto \zeta_{1...K}$ from function class $\mathcal{F}$ (e.g. ReLu networks producing $K$ scenarios)

- By bypassing the scenario generation step, the time between context and decision-making will be decreased while still unlocking the *value of stochastic solutions*
- The proposed approach is situated in the CSO zoo: major pros consisting of feasibility, decision-time tractability, and value of stochastic solutions

The figure below highlights the proposed approach



# Maximum Mean Discrepancy Distance

- A Reproducing Kernel Hilbert Space $\mathcal{H}$ (RKHS) with associated positive semi-definite kernel $k : \Omega \times \Omega \to \mathbb{R}$ is considered
- The squared maximum mean discrepancy allows us to measure the distance between distributions
- Gretton et al. (2012) showed that the squared maximum mean discrepancy between $\mathbb{P}_\eta$ and $\mathbb{P}_\omega$ is given by

$$d^2_{\mathcal{G}_{\text{MMD}}}(\mathbb{P}_\omega, \mathbb{P}_\eta) = \mathbb{E}_{(\omega,\omega')\sim\mathbb{P}_\omega}[k(\omega,\omega')] + \mathbb{E}_{(\eta,\eta')\sim\mathbb{P}_\eta}[k(\eta,\eta')] - 2\mathbb{E}_{\omega\sim\mathbb{P}_\omega,\eta\sim\mathbb{P}_\eta}[k(\omega,\eta)]$$

- There exists a rich theory associated with RKHSs that lead to some desirable properties (concentration inequalities, generalization guarantees, links to stability theory of 2SPs)
- Specification of $k$ determines the distance

# A Distributional Approach

- Select $\boldsymbol{f}$ to minimize a distributional distance between $\boldsymbol{f}(\boldsymbol{x})$ and $\mathbb{P}_{\omega|\boldsymbol{x}}$ ... in expectation over $\mathbb{P}_{\boldsymbol{x}}$
- Distributional Contextual Scenario Generation (DCSG):

$$\min_{\boldsymbol{f}\in\mathcal{F}} \mathcal{L}_{\text{dist}}(\boldsymbol{f}) := \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}}\left[d\left(\mathbb{P}_{\boldsymbol{f}(\boldsymbol{x})}, \mathbb{P}_{\omega|\boldsymbol{x}}\right)\right]$$

- where $\mathbb{P}_{\boldsymbol{f}(\boldsymbol{x})} = \frac{1}{K}\sum_{k=1}^{K}\delta_{f_k(\boldsymbol{x})}$ denotes an empirical measure associated with the $K$ scenarios $\boldsymbol{f}(\boldsymbol{x})$

# A Distributional Approach

- Using the squared MMD distance implies $\mathcal{L}_{\text{dist}}(\boldsymbol{f})$ can be written as

$$\mathcal{L}_{\text{dist}}(\boldsymbol{f}) = C + \mathbb{E}_{(\boldsymbol{x},\omega)\sim\mathbb{P}_{\boldsymbol{x},\omega}} \Bigg[ \underbrace{\frac{1}{K^2}\sum_{i=1}^{K}\sum_{i'=1}^{K} k\big(f_i(\boldsymbol{x}), f_{i'}(\boldsymbol{x})\big) - \frac{2}{K}\sum_{i=1}^{K} k\big(\omega, f_i(\boldsymbol{x})\big)}_{:=\ell_{\text{MMD}}(\boldsymbol{f}(\boldsymbol{x}),\omega)} \Bigg],$$

- where $C = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}}\mathbb{E}_{(\omega,\omega')\sim\mathbb{P}_{\omega|\boldsymbol{x}}}\left[k(\omega,\omega')\right]$ is a constant with respect to $\boldsymbol{f}$
- yielding an empirical risk minimization:

$$\min_{\boldsymbol{f}\in\mathcal{F}} \mathcal{L}_{\text{MMD}}(\boldsymbol{f}) := \mathbb{E}_{(\boldsymbol{x},\omega)\sim\mathbb{P}_{\boldsymbol{x},\omega}}\left[\ell_{\text{MMD}}(\boldsymbol{f}(\boldsymbol{x}),\omega)\right]$$

- We select the energy kernel, $k(\omega,\omega') = -\|\omega-\omega'\|_2$ due to its desirable properties. For example
  - $K = 1$ reduces to least-squares regression (the first ever contextual approach)
  - Considers some of the underlying geometry (Feydy et al., 2019)
  - Results $\mathcal{L}_{\text{dist}}$ being a metric
- Specify the parameterization for $\boldsymbol{f}$ and optimize $\boldsymbol{f}$ via gradients

## A Problem-driven Approach

- We obtain a first-stage solution by solving (2SP) using $\zeta_{1\ldots K} = \boldsymbol{f}(\boldsymbol{x})$

$$\boldsymbol{y}(\zeta_{1\ldots K}) \in \operatorname*{proj}_{\boldsymbol{y}} \operatorname*{argmin}_{\boldsymbol{y},\boldsymbol{z}_1,\ldots\boldsymbol{z}_K} h(\boldsymbol{y}) + \frac{1}{K}\sum_{i=1}^{K} q(\boldsymbol{z}_i, \zeta_i) \qquad (\zeta\text{ -SAA})$$
$$\boldsymbol{y} \in \mathcal{Y},\ \boldsymbol{z}_i \in \mathcal{Z}(\boldsymbol{y}, \zeta_i)\ \forall i \ \in \{1, ..., K\},$$

- which has an optimal solution set $\mathcal{Y}^*(\zeta_{1\ldots K}) :=$

$$\left\{ \boldsymbol{y} \in \mathcal{Y} : h(\boldsymbol{y}) + \frac{1}{K}\sum_{i=1}^{K} q(\boldsymbol{z}_i, \zeta_i) \le v^*(\zeta_{1\ldots K}), \boldsymbol{z}_i \in \mathcal{Z}(\boldsymbol{y}, \zeta_i),\ i \in \{1, \ldots, K\} \right\},$$

where $v^*(\zeta_{1\ldots K})$ is the optimal objective value of ($\zeta$-SAA)

# A Problem-driven Approach

- Select $\boldsymbol{f}$ such that the 2SP solutions on $\boldsymbol{f}(\boldsymbol{x})$ produce the highest quality set of optimal ($\zeta$-SAA) solutions $\mathcal{Y}^*(\boldsymbol{f}(\boldsymbol{x}))$ according to $\mathbb{P}_{\omega|\boldsymbol{x}}$
- We measure the quality of the optimal solution set $\mathcal{Y}^*(\boldsymbol{f}(\boldsymbol{x}))$ by the best possible two-stage performance among solutions in $\mathcal{Y}^*(\boldsymbol{f}(\boldsymbol{x}))$. This corresponds to

$$\boldsymbol{f}(\boldsymbol{x}) \in \underset{\zeta_1,\ldots,\zeta_K}{\operatorname{argmin}} \min_{\boldsymbol{y} \in \mathcal{Y}^*(\zeta_{1\ldots K})} h(\boldsymbol{y}) + \mathbb{E}_{\omega \sim \mathbb{P}_{\omega|\boldsymbol{x}}} \left[ Q\left(\boldsymbol{y}, \omega\right) \right]$$

  holding (*almost surely*) with respect to $\mathbb{P}_{\boldsymbol{x}}$.
- Relaxing the constraint that all $\boldsymbol{y} \in \mathcal{Y}^*(\zeta_{1\ldots K})$ must be the same in every $\omega$, yields the following

$$\min_{\boldsymbol{f} \in \mathcal{F}} \mathbb{E}_{(\boldsymbol{x},\omega) \sim \mathbb{P}_{\boldsymbol{x},\omega}} \left[ \min_{\boldsymbol{y} \in \mathcal{Y}^*(\boldsymbol{f}(\boldsymbol{x}))} h(\boldsymbol{y}) + Q\left(\boldsymbol{y}, \omega\right) \right] \tag{Opt-PCSG'}$$

- Suggesting the following loss function

$$\ell_{\text{opt}}(\zeta_{1\ldots K}, \omega) := \min_{\boldsymbol{y} \in \mathcal{Y}^*(\zeta_{1\ldots K})} h(\boldsymbol{y}) + Q\left(\boldsymbol{y}, \omega\right)$$

# Evaluating the Optimistic Loss

Evaluating $\ell_{\text{opt}}(\zeta_{1\dots K}, \omega)$ is equivalent to solving

$$\ell_{\text{opt}}(\zeta_{1\dots K}, \omega) = \min_{\boldsymbol{y},\boldsymbol{z},\boldsymbol{z}_1,\dots \boldsymbol{z}_K} \quad h(\boldsymbol{y}) + q(\boldsymbol{z}, \omega) \tag{Opt-Search}$$

$$s.t. \quad h(\boldsymbol{y}) + \frac{1}{K}\sum_{i=1}^{K} q(\boldsymbol{z}_i, \zeta_i) \leq v^*(\zeta_{1\dots K}) \tag{1}$$

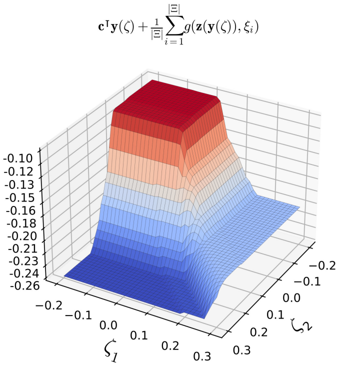
$$\boldsymbol{y} \in \mathcal{Y},\ \boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y}, \omega),\ \boldsymbol{z}_i \in \mathcal{Z}(\boldsymbol{y}, \zeta_i),\ \forall i \ \in \{1, \dots, K\}.$$

- In addition to linear programs and convex programs with multiple solutions, the optimistic approach is generally amenable to mixed-integer programs (MIP) with convex relaxations
- If (Opt-Search) is difficult to solve, we can settle for a feasible solution by solving ($\zeta$-SAA) and (Stage II) (easy for 2SP)
- (Opt-Search) reduces to solving ($\zeta$-SAA) and (Stage II) when ($\zeta$-SAA) exhibits unique solutions

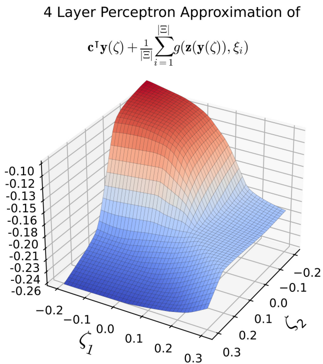
# Gradient Sparsity

Solving Opt-PCSG' via gradient-based methods is not easy since the gradients of the objective with respect to the surrogate scenarios are sparse.

Ex. $K = 1$, unique solutions, $n = 1$, $\boldsymbol{f} = (\zeta_1, \zeta_2)^T$




(a) The loss plotted against two components of the surrogate scenario $\zeta$.




(b) Approximating the loss surface with a neural network has a smoothing effect.

# Neural Network Approximation

We propose a neural architecture inspired by Patel et al. (2022)'s architecture, to model $\ell_{\text{opt}}$

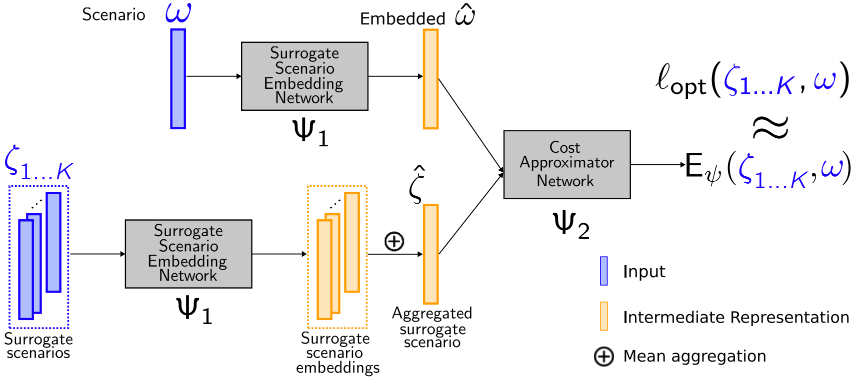



Figure: Loss-Net Architecture

How to train? $\rightarrow$ use $\boldsymbol{f}_{\text{MMD}}$ to generate dataset from contextual observations $\{\boldsymbol{x}^{(i)}, \zeta_{1\ldots K}^{(i)}\}_{i=1}^{n}$ with $\zeta_{1\ldots K}^{(i)} = \boldsymbol{f}_{\text{MMD}}(\boldsymbol{x}) \rightarrow$ fit $E_\psi$ with $\{(\zeta_{1\ldots K}^{(i)}, \omega^{(i)}), \ell_{\text{opt}}(\zeta_{1\ldots K}^{(i)}, \omega^{(i)})\}_{i=1}^{n}$

# Replacing the Loss with the Approximate Loss

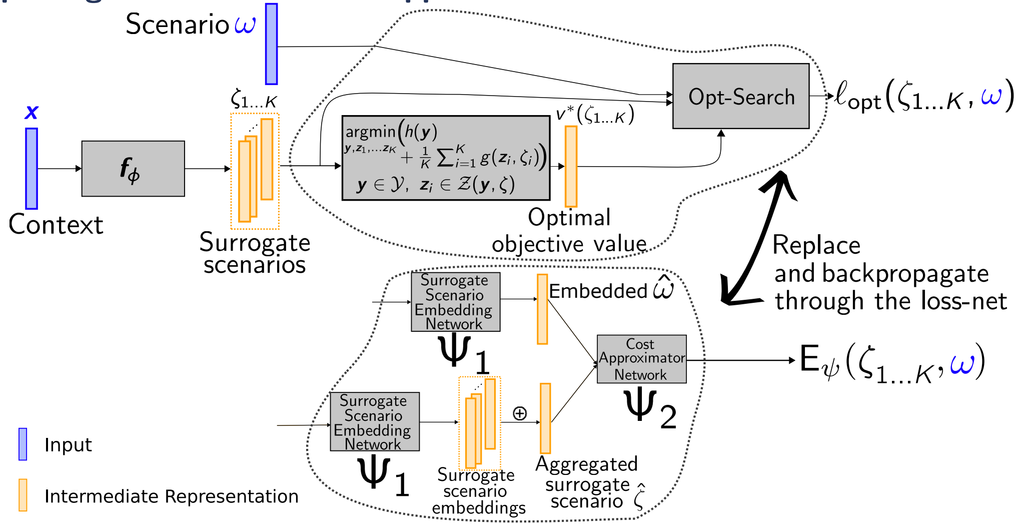



Figure: Replace the true downstream loss with the task-net approximated loss

# With MMD Regularization

- Optimizing the approximate task-loss con produce $\boldsymbol{f}$, such that $\zeta^{(i)}_{1\ldots K} = \boldsymbol{f}(\boldsymbol{x}^{(i)}),\ i = 1, \ldots, n$ are 'far' from $\zeta^{(i)}_{1\ldots K}$ originally used to train the loss approximator
- Idea: regularize the approximate task loss by the MMD loss to select $\boldsymbol{f}_\phi$

$$\min_{\phi} \quad \frac{1}{n}\sum_{i=1}^{n} \hat{E}(\boldsymbol{f}_\phi(\boldsymbol{x}^{(i)}), \omega^{(i)}) + \lambda \ell_{\text{MMD}}(\boldsymbol{f}_\phi(\boldsymbol{x}^{(i)}), \omega^{(i)}), \qquad \text{(Static-PCSG)}$$

- This approach with $K = 1$ is equivalent to Zharmagambetov et al. (2024)'s approach

# Minimizing the Approximate Loss Maximizes the Approximation Error

Fix $\lambda$, $K$, $T$, and define the contextual 2SP to define $\ell_{\text{opt}}$

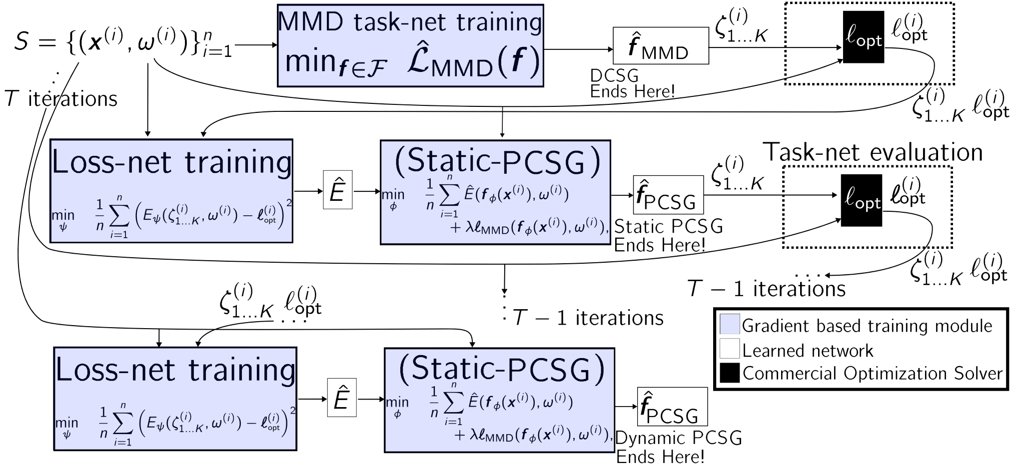



Figure: The relationship between the DCSG, Static-PCSG and Dynamic-PCSG approaches

# Contextual Newsvendor Problem

- Each day, a newsvendor purchases $y$ newspapers at unit-cost $c$ with a budget constraint: $y \leq u$
- The vendor sells as many papers as possible at a unit price $q$ and can return unsold papers at a salvage price $r < c$
- Contextual information $\boldsymbol{x}$, such as day of the week, weather, and previous sales, is observed before purchasing
- Demand $\omega$ is unknown at the time of purchase, after which the newsvendor optimally sells $z$ and salvages $w$ papers
- They wish to solve: $\min_y cy + \mathbb{E}_{\omega \sim \mathbb{P}_{\omega|\boldsymbol{x}}} [Q(y, \omega)] \quad s.t. \quad y \in [0, u]$ where $Q(y, \omega) = \min_{z \geq 0, w \geq 0} \; -qz - rw \quad s.t. \quad z \leq \omega, \; z + w \leq y$

# Experimental Setup

- **Distribution of contextual information:** Context $\boldsymbol{x}$ follows a 2D normal distribution, mapped to demand distribution $\mathbb{P}_{\omega|\boldsymbol{x}}$ over 200 points via a random neural network $\boldsymbol{f}_{\text{Random}}$. Demand has mean 15.1 and std. deviation 0.25
- **Sample:** $n = 500$ samples of $(\boldsymbol{x}, \omega)$; models *MMD* (DCSG), *Static*, and *Dynamic* trained on this set
- **Model Architectures:** $\boldsymbol{f}_\phi$ is a fully connected ReLU network, selected via a random search optimizing $\hat{\mathcal{L}}_{\text{MMD}}$ on a 20% hold-out set. Loss network $E_\psi$ architecture is selected similarly
- **Evaluation:** Models tested on $i = 1, ..., n_{\text{val}} = 100$ out-of-sample $\boldsymbol{x}_{\text{val}}^{(i)}$ and their associated $\mathbb{P}_{\omega|\boldsymbol{x}_{\text{val}}^{(i)}}$. First-stage solution evaluated using 2SP objective, considering surrogate scenarios $\zeta_{1\ldots K}^{(i)} = \boldsymbol{f}(\boldsymbol{x}^{(i)})$ (full process repeated 20 times)
- **Comparison Metric:** True 2SP Objectives $v_{\text{MMD}}^{(i)}$, $v_{\text{Static}}^{(i)}$, $v_{\text{Dynamic}}^{(i)}$, and optimal newsvendor solution $v_{\text{2SP}}^{(i)}$ computed per validation observation

# Two Benchmaks

## The Expected Value Solution

- Determined by considering the newsvendor problem with a single scenario given by $\mathbb{E}[\omega|\boldsymbol{x}]$
- In practice, the conditional mean is not known, so this is a benchmark that cannot be implemented

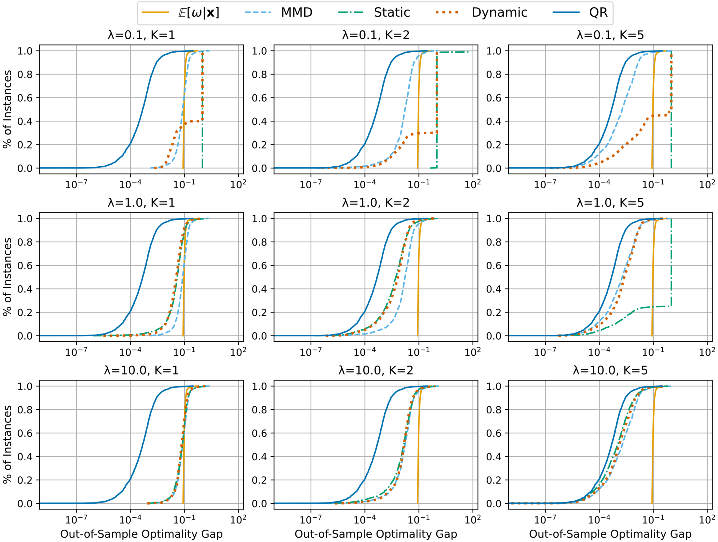
## Analytical Contextual Solution

- The optimal solution to the newsvendor is

$$y^* = \begin{cases} 0 & \text{if } \frac{q-c}{q-r} < F_{\mathbb{P}}(0), \\ u & \text{if } \frac{q-c}{q-r} > F_{\mathbb{P}}(u), \\ F_{\mathbb{P}}^{-1}\left(\frac{q-c}{q-r}\right) & \text{otherwise.} \end{cases}$$

  where $F_{\mathbb{P}}$ is the cumulative density according to $\mathbb{P}$ and $F_{\mathbb{P}}^{-1}(\alpha)$ is the $\alpha$ quantile of $F$.
- suggesting a quantile estimation approach to obtain a linear model $F_{\mathbb{P}_{\omega|\boldsymbol{x}}}^{-1}\left(\frac{q-c}{q-r}\right) = \boldsymbol{\beta}^\intercal \boldsymbol{x}$, from the training sample (Liu et al., 2022)

# Out of Sample Gap – Cumulative Distribution Functions



# Takeaways and Conclusion

- The proposed approach is able to unlock the value of stochastic solutions
- Although, it is not as performant as problem-specific analytic approaches, it is competitive in certain settings
- Regularization is critically important for the problem-driven approach
- Refining the loss approximation tends to improve performance

We perform similar exercises for other 2SPs from manufacturing and finance and demonstrate computational and performance benifets.

**Summary: A generic framework for generating scenarios contextually that is performant and computationally tractable**

Thank you!

I look forward to your questions

# Contextual Scenario Generation for 2SP – Distributional

Select $\boldsymbol{f} : \boldsymbol{x} \mapsto \zeta_{1\ldots K}$ from a vector-valued function class $\mathcal{F}$ such that the distributional distance between the empirical distribution supported on $\boldsymbol{f}(\boldsymbol{x})$ and $\mathbb{P}_{\omega|\boldsymbol{x}}$ is minimized in expectation over $\mathbb{P}_{\boldsymbol{x}}$.

- $\mathcal{F}$: hypothesis set for the task mapping, parametrized via feed-forward fully connected ReLu networks.
- Approach: Distributional Contextual Scenario Generation (DCSG)

$$\min_{\boldsymbol{f} \in \mathcal{F}} \mathcal{L}_{\text{dist}}(\boldsymbol{f}) := \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}}} \left[ d \left( \mathbb{P}_{\boldsymbol{f}(\boldsymbol{x})}, \mathbb{P}_{\omega|\boldsymbol{x}} \right) \right] \tag{DCSG}$$

- $\mathbb{P}_{\boldsymbol{f}(\boldsymbol{x})} = \frac{1}{K} \sum_{\zeta \in \boldsymbol{f}(\boldsymbol{x})}^{K} \delta_{\zeta}$: empirical measure associated with $\boldsymbol{f}(\boldsymbol{x})$.
- $d(\cdot, \cdot)$: measure of distance between the distributions.

# Contextual Scenario Generation for 2SP – Problem Driven

Propose a problem-driven contextual scenario Generation approach (PCSG), that aims to select $\boldsymbol{f}$ such that the 2SP solution defined on $\boldsymbol{f}$ minimizes the 2SP objective evaluated using $\mathbb{P}_{\omega|\boldsymbol{x}}$

$$\begin{aligned} \min_{\boldsymbol{f}\in\mathcal{F}} \quad & \mathbb{E}_{(\boldsymbol{x},\omega)\sim\mathbb{P}_{\boldsymbol{x},\omega}}\left[h(\boldsymbol{y}\left(\boldsymbol{f}(\boldsymbol{x})\right)) + Q\left(\boldsymbol{y}\left(\boldsymbol{f}(\boldsymbol{x})\right),\omega\right)\right] \\ s.t. \quad & \boldsymbol{y}\left(\boldsymbol{f}(\boldsymbol{x})\right) \in \operatorname{proj}_{\boldsymbol{y}} \operatorname*{argmin}_{\boldsymbol{y},\boldsymbol{z}_1,\ldots\boldsymbol{z}_K\in\mathcal{Y}\times\mathcal{Z}(\boldsymbol{y},\boldsymbol{f}(\boldsymbol{x}))} h(\boldsymbol{y}) + \frac{1}{K}\sum_{i=1}^{K} q(\boldsymbol{z}_i,\boldsymbol{f}_i(\boldsymbol{x})) \quad (a.s.) \end{aligned} \tag{PCSG}$$

# Contextual Scenario Generation for 2SP – The Zoo

Estes and Richard (2023) provide an insightful segmentation of contextual methods for stochastic optimization.

**Conditional-density-estimation-then-optimize**

- Estimate $\mathbb{P}_{\omega|\boldsymbol{x}}$ then solve 2SP $\rightarrow$ does not directly address computational considerations

**Direct-solution-prediction**

- Form a mapping $\boldsymbol{g} : \boldsymbol{x} \mapsto \boldsymbol{y}$, from function class $\mathcal{G}$ to solve $\min_{\boldsymbol{g}\in\mathcal{G}} \mathbb{E}_{(\boldsymbol{x},\omega)\sim\mathbb{P}_{\boldsymbol{x},\omega}} [h(\boldsymbol{g}(\boldsymbol{x})) + Q(\boldsymbol{g}(\boldsymbol{x}), \omega)]$ $\rightarrow$ necessitates problem-specific approaches due to feasibility considerations

**Predict-then-optimize**

- Solve $\min_{\boldsymbol{y},\boldsymbol{z}\in\mathbb{R}^{s_1}\times s_2} h(\boldsymbol{y}) + q(\boldsymbol{z}, \tilde{\omega}) \quad s.t \quad \boldsymbol{y} \in \mathcal{Y},\ \boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y}, \tilde{\omega})$ where $\tilde{\omega} = \boldsymbol{h}(\boldsymbol{x}) \in \Omega$ is a scenario predicted by a learned $\boldsymbol{h}$ in some hypothesis class $\mathcal{H}$
- Naive predict-then-optimize: select $\boldsymbol{h}$ via standard statistical approaches $\rightarrow$ does not consider downstream costs
- Smart predict-then-optimize: select $\boldsymbol{h}$ such that the resulting solutions to perform well according to $\mathbb{E}_{(\boldsymbol{x},\omega)\sim\mathbb{P}_{\boldsymbol{x},\omega}} [h(\boldsymbol{y}) + Q(\boldsymbol{y}, \omega)]$ $\rightarrow$ typically problem-specific, suffers from vanishing gradients, resulting solutions do not hedge against uncertainty (by nature of being a point predictor)

# Contextual Scenario Generation for 2SP – The Zoo

The proposed approach is related to the aforementioned areas as follows:

**Conditional-density-estimation-then-optimize**

- Addresses computational concerns by approximating $\mathbb{P}_{\omega|\boldsymbol{x}}$ by a distribution supported on $K$ scenarios

**Direct-solution-prediction**

- (2SP) defined on $K$ discrete scenarios preserves first-stage feasibility. The proposed approach makes minimal assumptions regarding problem structure

**Predict-then-optimize**

- Using a distribution supported on $K$ scenarios results in solutions with more optionality. Furthermore, the loss network, tailored to 2SPs makes addresses the issues of gradients

## Contextual Scenario Generation for 2SP – Experimental Results

- A joint sample $n = 500$ is obtained as the decision maker's historical data
- The proposed methods are trained on the joint sample and evaluated out-of-sample using $n_{\text{eval}}$ true conditional distributions $\mathbb{P}_{\omega|\mathbf{x}}$
- This process is repeated 20 times to ensure reproducibility
- Example: Table 1 shows a contextual version of Higle and Sen (1996)'s CEP1 2SP, comparing methods across $K$

| | | $\mathbb{E}[\omega\|\mathbf{x}]$ | MMD | Static | | | | Dynamic | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\tilde{\lambda}$ | - | - | 0.01 | 0.1 | 1 | 10 | 0.01 | 0.1 | 1 | 10 |
| | 1 | 11.6% | 0.0% | 29.8% | 0.2% | 0.0% | 0.1% | 55.7% | 2.1% | 0.4% | 0.3% |
| $K$ | 2 | 0.2% | 15.1% | 5.4% | 4.9% | 10.0% | 12.6% | 18.9% | 9.2% | 10.1% | 13.8% |
| | 5 | 0.1% | 13.9% | 0.3% | 7.9% | 11.5% | 14.2% | 17.0% | 13.2% | 9.9% | 11.9% |

Table: (CEP1) Fraction of instances where each method has the lowest out-of-sample 2SP cost. The regularizaton is set such that $\lambda = \text{DCSR Objective} * \tilde{\lambda}$

# References I


Estes, A. S. and Richard, J.-P. P. (2023). Smart predict-then-optimize for two-stage linear programs with side information. *INFORMS Journal on Optimization*, 5(3):295–320.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. (2019). Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Higle, J. L. and Sen, S. (1996). *Stochastic Decomposition*. Kluwer Academic Publishers, Norwell, MA, USA.

Liu, C., Letchford, A. N., and Svetunkov, I. (2022). Newsvendor problems: An integrated method for estimation and optimisation. *European Journal of Operational Research*, 300(2):590–601.

Ntaimo, L. (2024). *Example Applications of Stochastic Programming*, pages 111–152. Springer International Publishing, Cham.

Patel, R. M., Dumouchelle, J., Khalil, E., and Bodur, M. (2022). Neur2sp: Neural two-stage stochastic programming. *Advances in Neural Information Processing Systems*, 35:23992–24005.

Zharmagambetov, A., Amos, B., Ferber, A., Huang, T., Dilkina, B., and Tian, Y. (2024). Landscape surrogate: Learning decision losses for mathematical optimization under partial information. *Advances in Neural Information Processing Systems*, 36.