Introduction to Data-driven Contextual Stochastic Optimization

Erick Delage
GERAD & Department of Decision Sciences
HEC MONTRÉAL

(Joint work with Utsav Sadana, Abhilash Chenreddy, Alexandre Forel, Emma Frejinger, Thibaut Vidal)

June 23, 2023

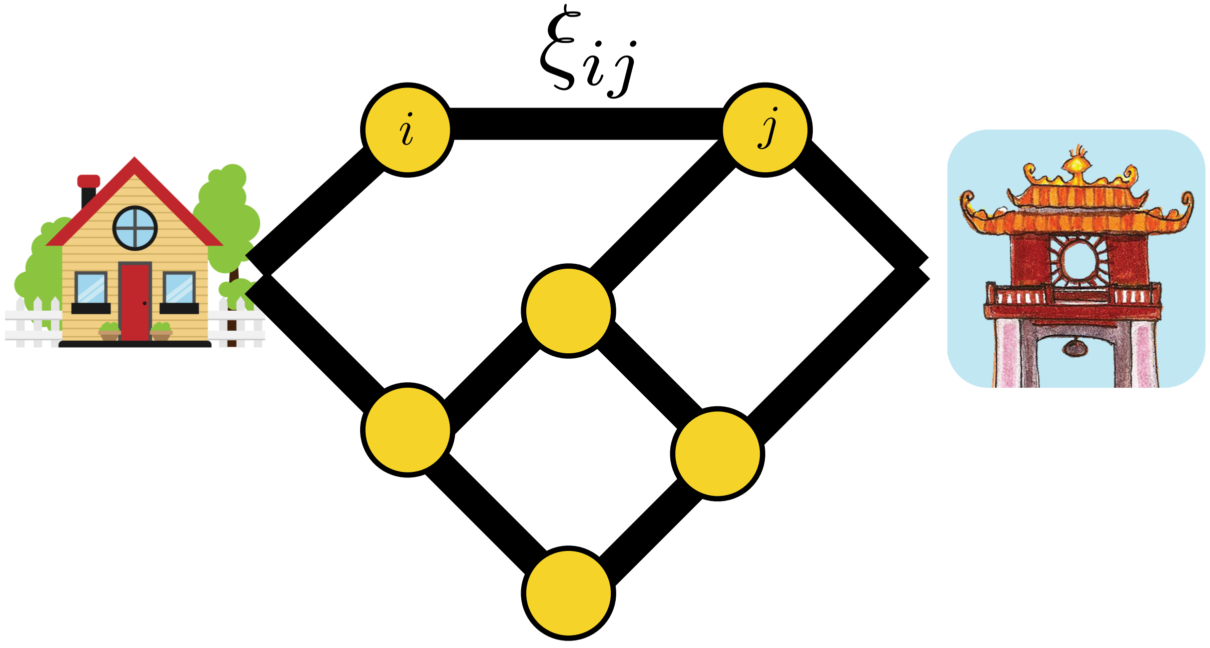
GERAD IVADO

Canada Research Chairs
Chaires de recherche du Canada
Canada

Why contextual stochastic optimization?

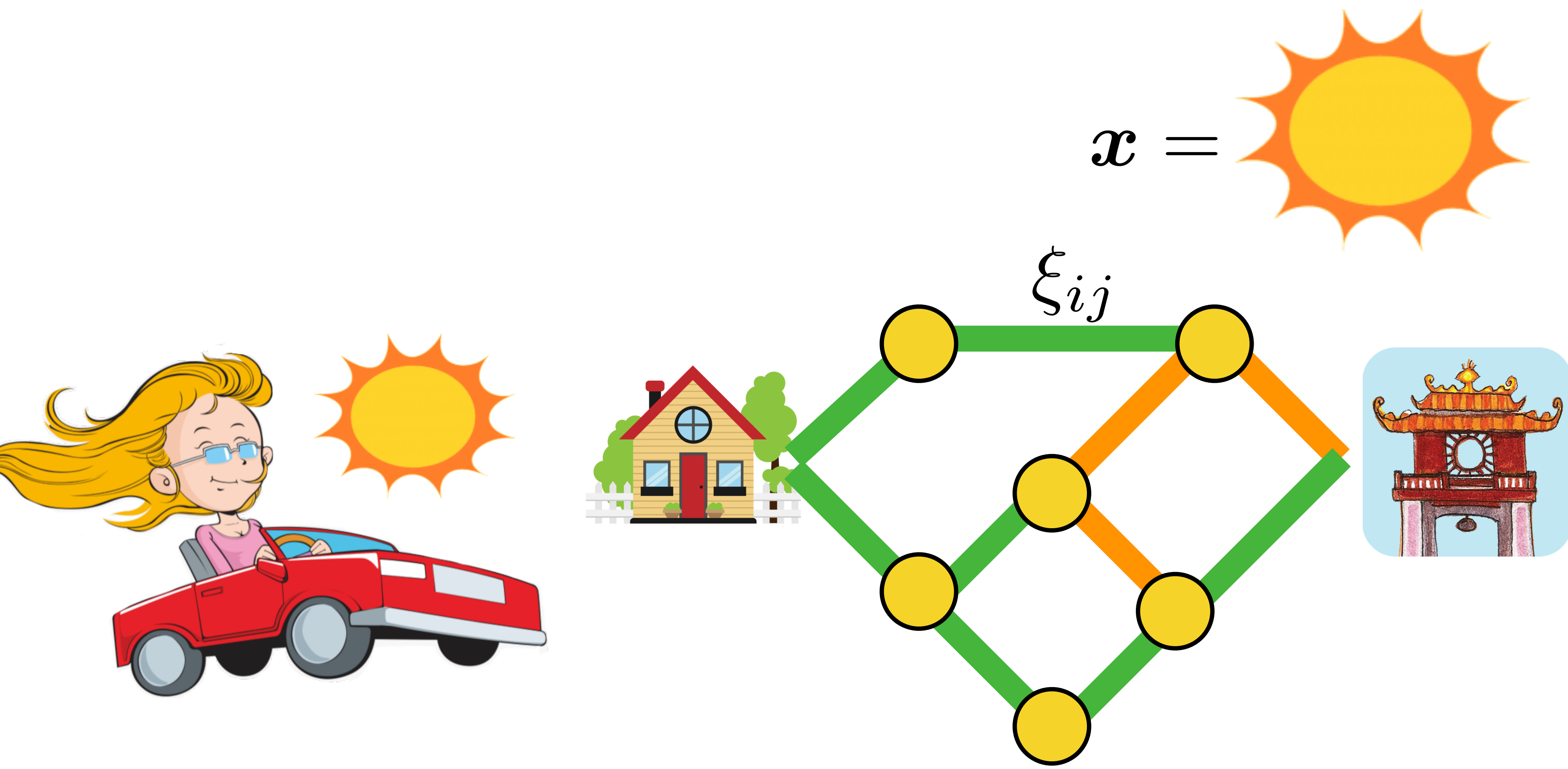
Decision Making with Contextual Information

- Revealed contextual information x
- Hidden random variables ξ



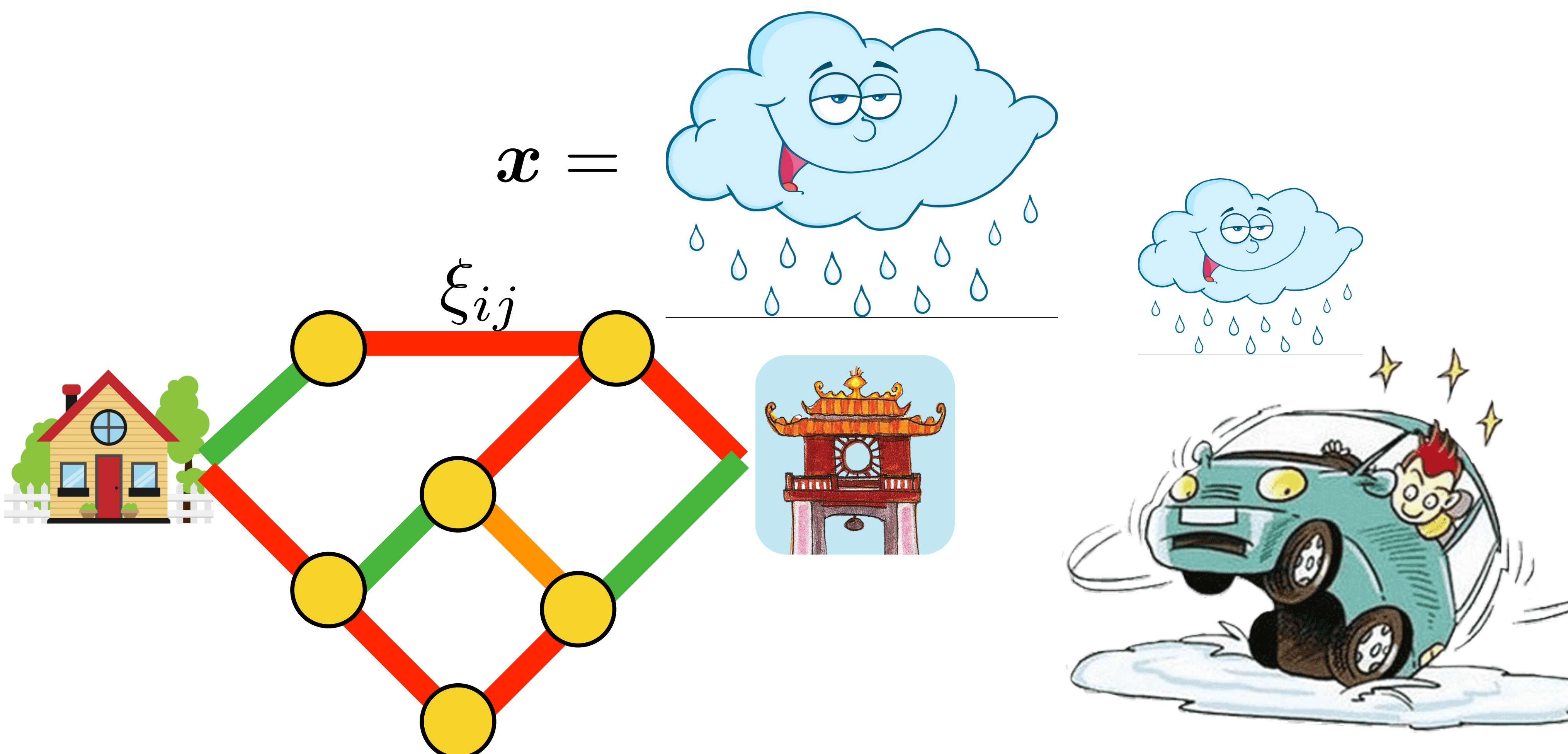
Decision Making with Contextual Information

- Revealed contextual information x
- Hidden random variables ξ



Decision Making with Contextual Information

- Revealed contextual information x
- Hidden random variables ξ

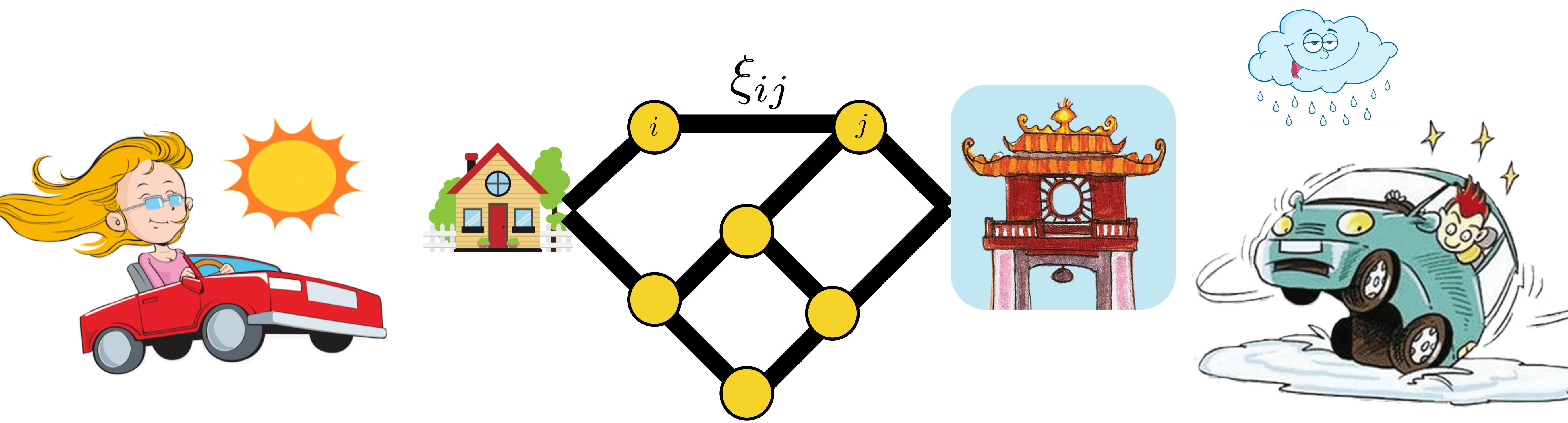


Decision Making with Contextual Information

- Revealed contextual information x
- Hidden random variables ξ

$\mathbb{E}^{\mathbb{P}}[\boldsymbol{\xi}|\boldsymbol{x} =$  $]$

$\mathbb{E}^{\mathbb{P}}[\boldsymbol{\xi}|\boldsymbol{x} =$  $]$

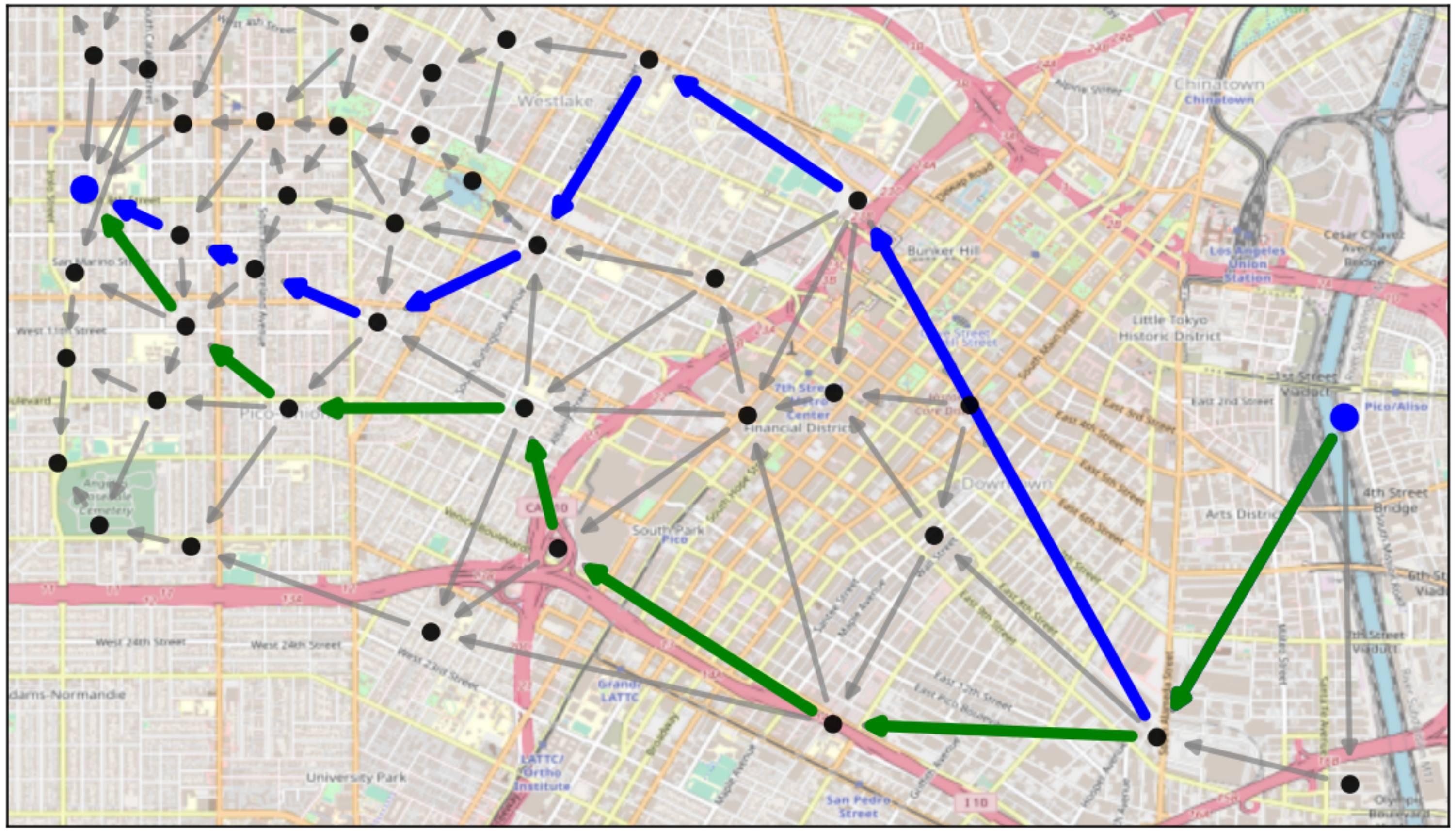


Practical motivation

Example 1:
Shortest path over Los Angeles downtown (Kallus & Mao, 2022)

Problem: find shortest path traversing Los Angeles downtown area from East to West

Travel times over all arcs are uncertain. We have relevant contextual information.

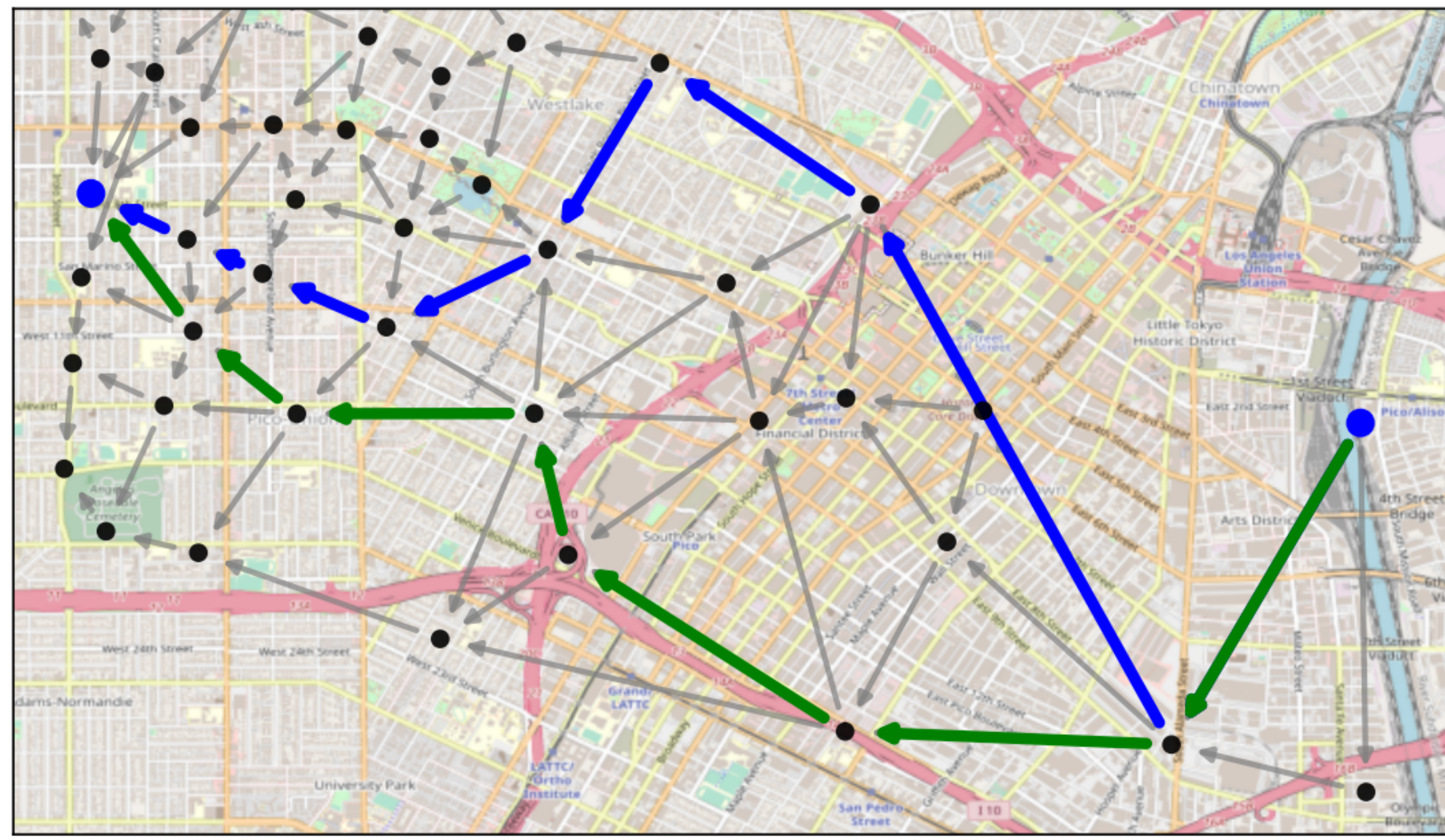


Practical motivation

Example 1:
Shortest path over Los Angeles downtown (Kallus & Mao, 2022)

Problem: find shortest path traversing Los Angeles downtown area from East to West

Travel times over all arcs are uncertain. We have relevant contextual information.



Green path is optimal → Midday
Blue path is optimal → AM

Period	Temp.	Wind speed	Rain	Visibility	Day	Month
Midday	57.17	4	0	6.99	2	11
AM	57.17	4	0	6.99	2	11

What is contextual optimization?

Problem Definition

Stochastic optimization

$$z^* \in \underset{z \in \mathcal{Z}}{\operatorname{argmin}} \, \mathbb{E}_{\mathbb{P}(\xi)} \left[c \left(z, \xi \right) \right]$$

Problem Definition

Stochastic optimization

$$\boldsymbol{z}^* \in \operatorname*{argmin}_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}_{\mathbb{P}(\boldsymbol{\xi})} \left[c\left(\boldsymbol{z}, \boldsymbol{\xi}\right) \right]$$

Context

Conditional stochastic optimization (CSO)

$$\boldsymbol{z}^*(\boldsymbol{x}) \in \operatorname*{argmin}_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}_{\mathbb{P}(\boldsymbol{\xi} | \boldsymbol{x})} \left[c\left(\boldsymbol{z}, \boldsymbol{\xi}\right) \right]$$

Contextual information / covariate / features

Problem Definition

Stochastic optimization

$$z^* \in \underset{z \in \mathcal{Z}}{\operatorname{argmin}} \, \mathbb{E}_{\mathbb{P}(\xi)} \left[c\left(z, \xi \right) \right]$$

Context →

Conditional stochastic optimization (CSO)

$$z^*(x) \in \underset{z \in \mathcal{Z}}{\operatorname{argmin}} \, \mathbb{E}_{\mathbb{P}(\xi | x)} \left[c\left(z, \xi \right) \right]$$

Contextual information / covariate / features

Connection between CSO and policy optimization:

$$\pi^* \in \underset{\pi : \mathcal{X} \to \mathcal{Z}}{\arg\min} \, \mathbb{E}_{\mathbb{P}}[c(\pi(x), \xi)] \quad \Leftrightarrow \quad \pi^*(x) \in \underset{z \in \mathcal{Z}}{\arg\min} \, \mathbb{E}_{\mathbb{P}(\xi|x)}[c(x, \xi)] \text{ a.s.}$$

Problem Definition

Stochastic optimization

$$z^* \in \underset{z \in \mathcal{Z}}{\text{argmin}}\, \mathbb{E}_{\mathbb{P}(\boldsymbol{\xi})}\left[c\left(\boldsymbol{z}, \boldsymbol{\xi}\right)\right]$$

Context

Conditional stochastic optimization (CSO)

$$\boldsymbol{z}^*(\boldsymbol{x}) \in \underset{\boldsymbol{z} \in \mathcal{Z}}{\text{argmin}}\, \underbrace{\mathbb{E}_{\mathbb{P}(\boldsymbol{\xi}|\boldsymbol{x})}\left[c\left(\boldsymbol{z}, \boldsymbol{\xi}\right)\right]}_{h(\boldsymbol{z}, \mathbb{P}(\boldsymbol{\xi}|\boldsymbol{x}))}$$

Contextual information / covariate / features

Conditional expected cost

Connection between CSO and policy optimization:

$$\pi^* \in \underset{\pi:\mathcal{X}\to\mathcal{Z}}{\arg\min}\, \mathbb{E}_{\mathbb{P}}[c(\pi(\boldsymbol{x}), \boldsymbol{\xi})] \quad \Leftrightarrow \quad \pi^*(\boldsymbol{x}) \in \underset{\boldsymbol{z}\in\mathcal{Z}}{\arg\min}\, \mathbb{E}_{\mathbb{P}(\boldsymbol{\xi}|\boldsymbol{x})}[c(\boldsymbol{x}, \boldsymbol{\xi})] \text{ a.s.}$$

Problem Definition

Stochastic optimization

$$z^* \in \underset{z \in \mathcal{Z}}{\operatorname{argmin}} \mathbb{E}_{\mathbb{P}(\xi)} \left[c\left(z, \xi \right) \right]$$

Context

Conditional stochastic optimization (CSO)

$$z^*(x) \in \underset{z \in \mathcal{Z}}{\operatorname{argmin}} \underbrace{\mathbb{E}_{\mathbb{P}(\xi|x)} \left[c\left(z, \xi \right) \right]}_{h(z, \mathbb{P}(\xi|x))}$$

Conditional expected cost

Contextual information / covariate / features

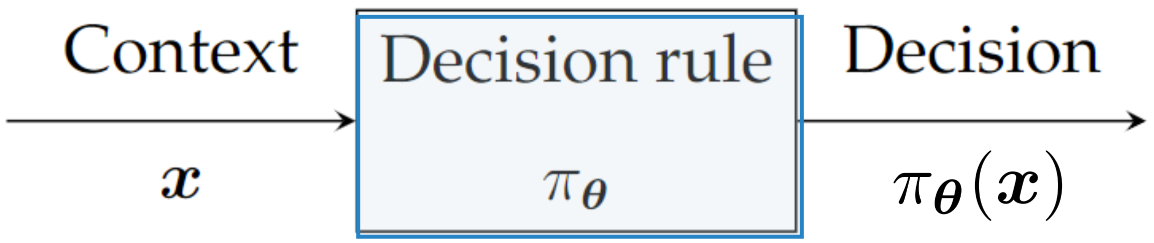
Connection between CSO and policy optimization:

$$\pi^* \in \underset{\pi: \mathcal{X} \to \mathcal{Z}}{\arg\min} \underbrace{\mathbb{E}_{\mathbb{P}}[c(\pi(x), \xi)]}_{H(\pi, \mathbb{P})} \quad \Leftrightarrow \quad \pi^*(x) \in \underset{z \in \mathcal{Z}}{\arg\min} \mathbb{E}_{\mathbb{P}(\xi|x)}[c(x, \xi)] \text{ a.s.}$$

(Unconditional) expected cost

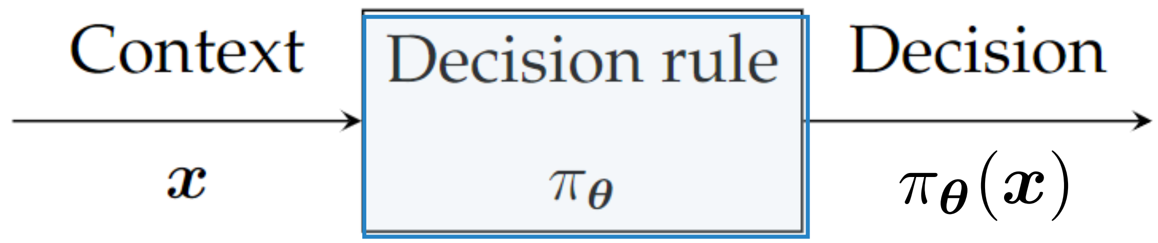
Overview of the three frameworks

Decision rule/Policy optimization

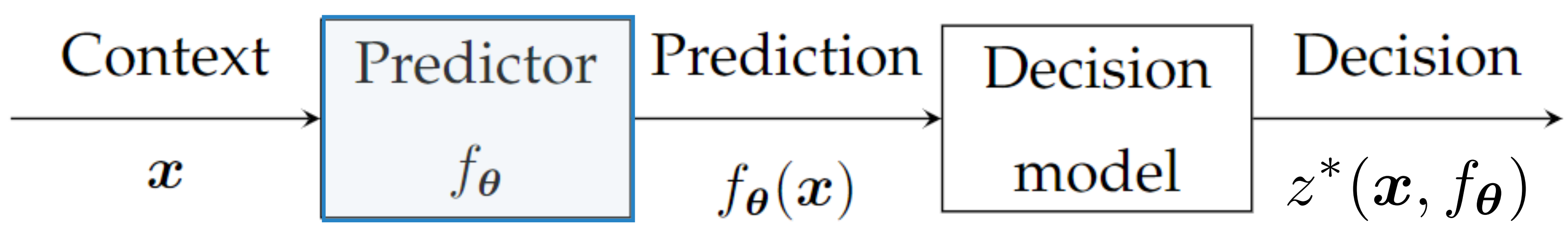


Overview of the three frameworks

Decision rule/Policy optimization

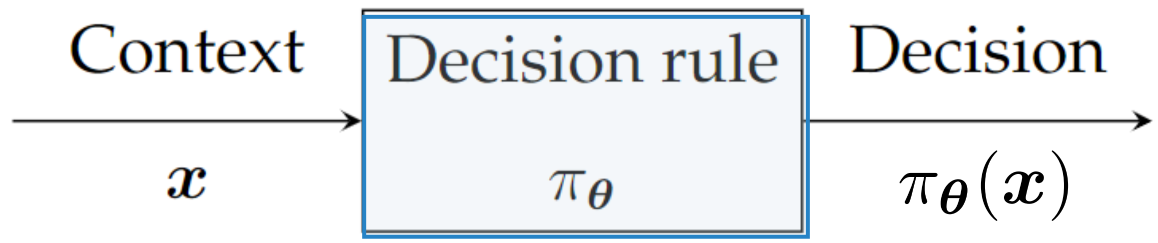


Sequential learning and optimization

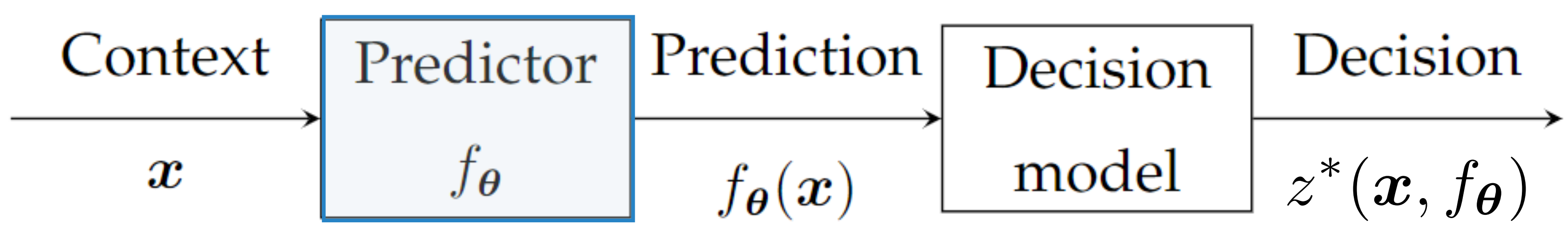


Overview of the three frameworks

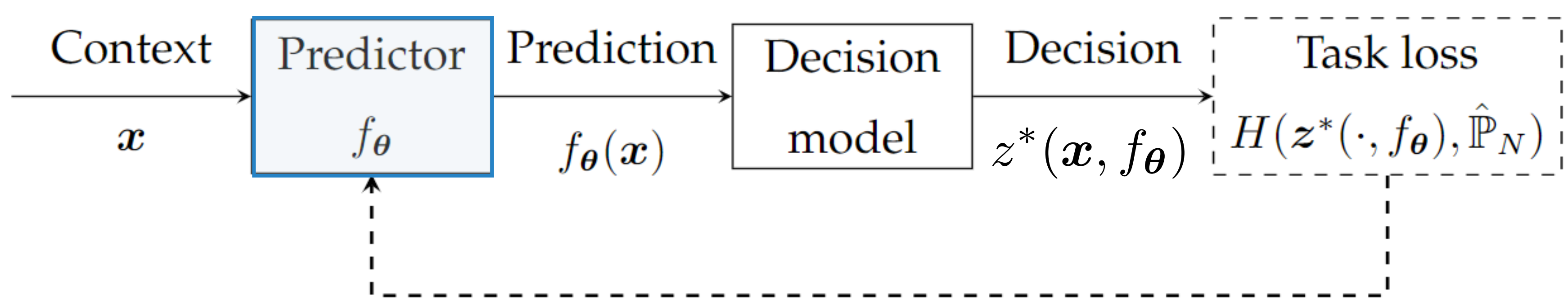
Decision rule/Policy optimization



Sequential learning and optimization



Integrated learning and optimization



Sequential learning and optimization

Learning predictors

Data
$D_N = \{(\boldsymbol{x}^i, \boldsymbol{\xi}^i)\}_{i=1}^N$

→

Predict / Estimate
Minimize *estimation error.*
$\min_{\boldsymbol{\theta}} \rho(f_{\boldsymbol{\theta}}, \hat{\mathbb{P}}_N) + \Omega(\theta)$

→

Optimize

Learning predictors

Data

$$D_N = \{(\boldsymbol{x}^i, \boldsymbol{\xi}^i)\}_{i=1}^N$$

→

Predict / Estimate

Minimize *estimation error.*

$$\min_{\boldsymbol{\theta}} \rho(f_{\boldsymbol{\theta}}, \hat{\mathbb{P}}_N) + \Omega(\theta)$$

→

Optimize

Non-linear cost function

f_θ is a **conditional density estimator**

Maximum Likelihood

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} -\log(\mathbb{P}_{f_\theta(\boldsymbol{x}^i)}(\boldsymbol{\xi}^i)) + \Omega(\theta)$$

Learning predictors

Data

$$D_N = \{(\boldsymbol{x}^i, \boldsymbol{\xi}^i)\}_{i=1}^N$$

→

Predict / Estimate

Minimize *estimation error.*

$$\min_{\boldsymbol{\theta}} \rho(f_{\boldsymbol{\theta}}, \hat{\mathbb{P}}_N) + \Omega(\theta)$$

→

Optimize

Non-linear cost function	**Linear cost function**
f_θ is a **conditional density estimator**	f_θ replaced with **point predictor** (denoted g_θ)
Maximum Likelihood $\hat{\theta} = \arg\min_\theta \frac{1}{N}\sum_{i=1}^N -\log(\mathbb{P}_{f_\theta(\boldsymbol{x}^i)}(\boldsymbol{\xi}^i)) + \Omega(\theta)$	*Mean Square Error* $\hat{\theta} = \arg\min_\theta \frac{1}{N}\sum_{i=1}^N \lVert g_\theta(\boldsymbol{x}^i) - \boldsymbol{\xi}^i\rVert^2 + \Omega(\theta)$

$$\mathbb{E}_{f_\theta(\boldsymbol{x})}[\boldsymbol{\xi}^\top \boldsymbol{z}] = \mathbb{E}_{f_\theta(\boldsymbol{x})}[\boldsymbol{\xi}]^\top \boldsymbol{z} = g_\theta(\boldsymbol{x})^\top \boldsymbol{z}$$

Weighted SAA

Minimizing expected costs w.r.t. a distribution is often done through SAA:

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \mathbb{E}_{f_\theta(\boldsymbol{x})}[c(\boldsymbol{z},\boldsymbol{\xi})] \text{ with } f_\theta(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N}\delta_{\boldsymbol{\xi}^i}$$

Weighted SAA

Minimizing expected costs w.r.t. a distribution is often done through SAA:

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \mathbb{E}_{f_\theta(\boldsymbol{x})}[c(\boldsymbol{z},\boldsymbol{\xi})] \text{ with } f_\theta(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{\xi}^i(\boldsymbol{x})}$$

Residual based
Measure the **error of a trained regression model** on the historical data
$f_\theta(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} \delta_{g_\theta(\boldsymbol{x})+\epsilon_i}$

Weighted SAA

Minimizing expected costs w.r.t. a distribution is often done through SAA:

$$\min_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}_{f_\theta(\boldsymbol{x})}[c(\boldsymbol{z}, \boldsymbol{\xi})] \text{ with } f_\theta(\boldsymbol{x}) := \sum_{i=1}^{N} \delta_{\boldsymbol{\xi}^i} \cdot w_i(\boldsymbol{x})$$

Residual based

Measure the **error of a trained regression model** on the historical data

$$f_\theta(\boldsymbol{x}) := \frac{1}{N} \sum_{i=1}^{N} \delta_{g_\theta(\boldsymbol{x}) + \epsilon^i}$$

Weight based

Measure **proximity in feature space** between x and historical covariates x^i

Weighted SAA

Proximity in feature space

- k-nearest neighbor: $w_i^{\mathrm{kNN}}(\boldsymbol{x}) := (1/k)\mathbb{1}[\boldsymbol{x}^i \in \mathcal{N}_k(\boldsymbol{x})]$
- Kernel density estimation: $w_i^{\mathrm{KDE}}(\boldsymbol{x}) := \dfrac{\mathcal{K}((\boldsymbol{x} - \boldsymbol{x}^i)/\theta)}{\sum_{j=1}^{N} \mathcal{K}((\boldsymbol{x} - \boldsymbol{x}^j)/\theta)}$

Weighted SAA

Proximity in feature space

- k-nearest neighbor: $$w_i^{\mathrm{kNN}}(\boldsymbol{x}) := (1/k)\mathbb{1}[\boldsymbol{x}^i \in \mathcal{N}_k(\boldsymbol{x})]$$
- Kernel density estimation: $$w_i^{\mathrm{KDE}}(\boldsymbol{x}) := \frac{\mathcal{K}((\boldsymbol{x}-\boldsymbol{x}^i)/\theta)}{\sum_{j=1}^N \mathcal{K}((\boldsymbol{x}-\boldsymbol{x}^j)/\theta)}$$

Supervised learning

- Decision tree: $$w_i^{\mathrm{DT}}(\boldsymbol{x}) := \frac{\mathbb{1}[\mathcal{R}(\boldsymbol{x}) = \mathcal{R}(\boldsymbol{x}^i)]}{\sum_{j=1}^N \mathbb{1}[\mathcal{R}(\boldsymbol{x}) = \mathcal{R}(\boldsymbol{x}^j)]}$$
- Random forest: average over set of decision trees.

Why do sequential learning and optimization?

It's fast!

- ➢Train once on historical data:
  no need to solve optimization models during training

It works

- ➢Can perform better than non-contextual approach
- ➢Can be trained using less data when model is well specified

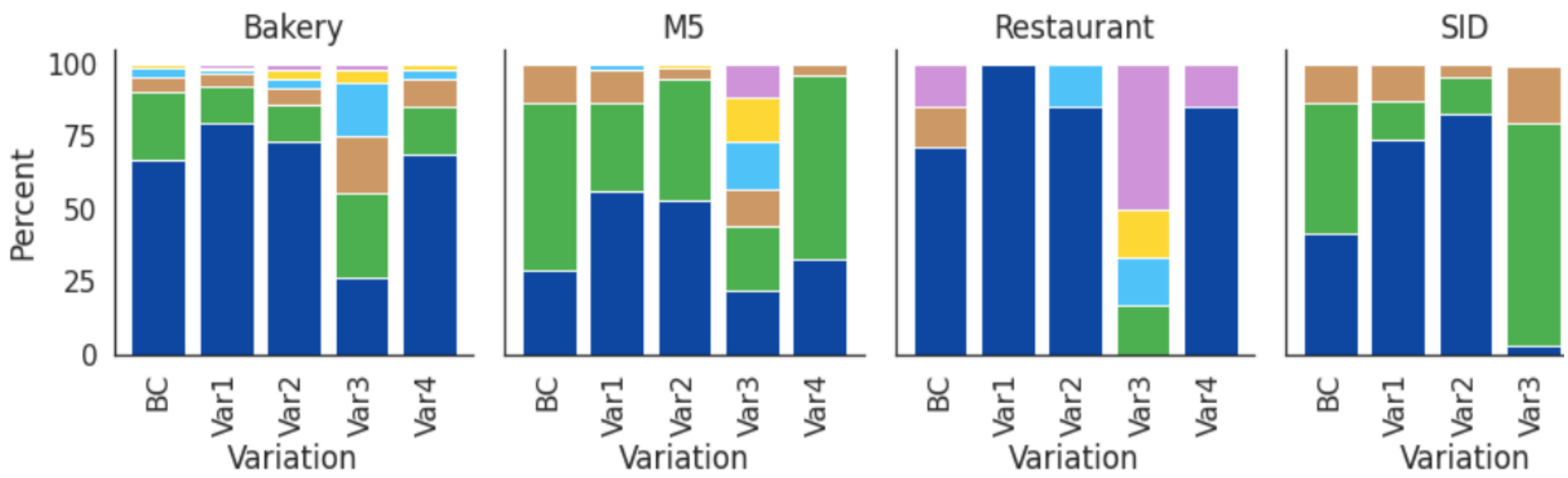
Theoretical guarantees

- ➢Converges to optimal contextual policy as the size of the training set increases **when model is well specified**.

Some benchmark results (Buttler et al., 2023)

Newsvendor Problem	Compare **sequential** L&O and **decision rules** on 4 real-world data sets.

Proportion of instances where methods achieved best performance



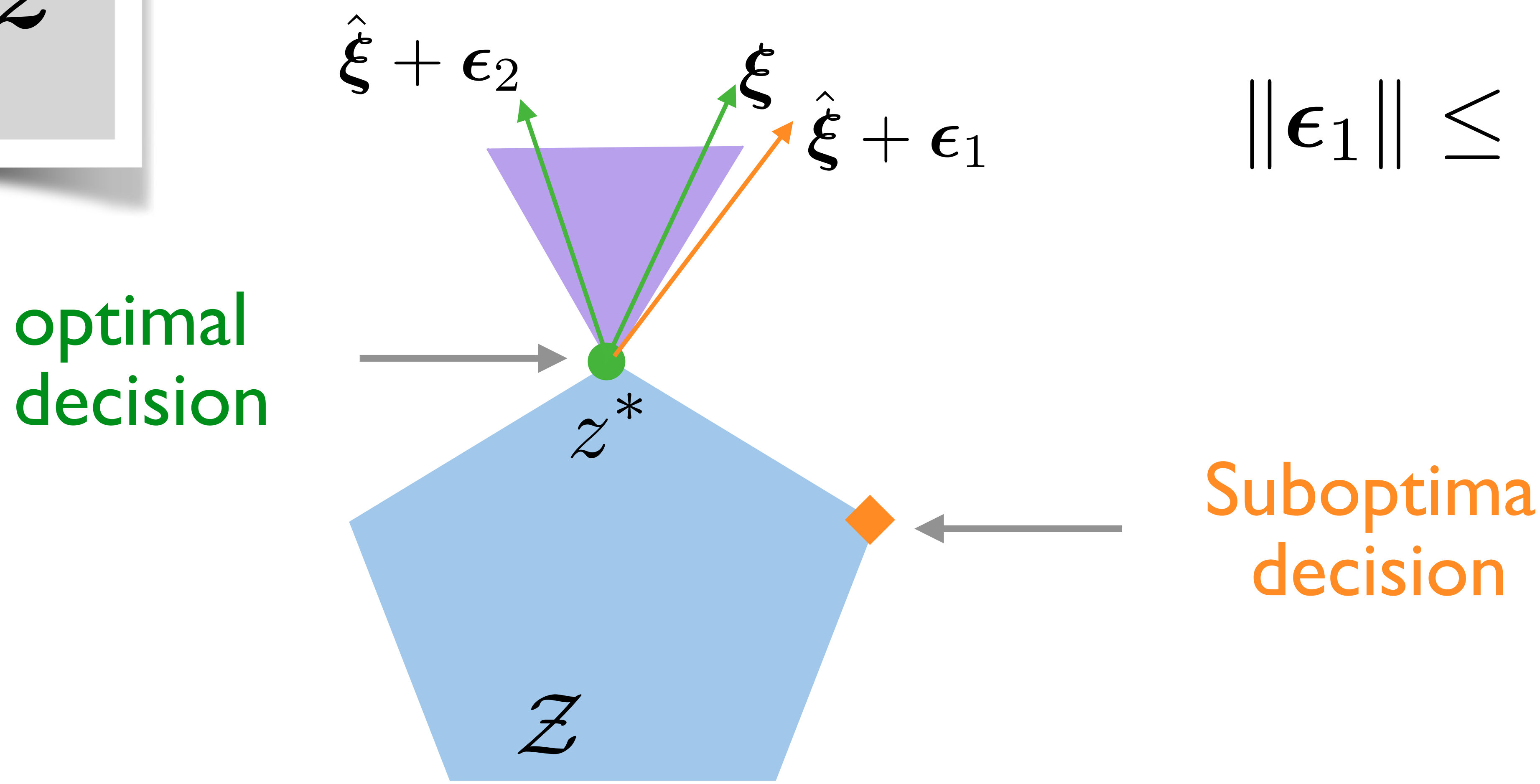
Models:

- Linear rule
- Deep learning
- Kernel weights
- K-nearest neighbour weights
- Decision tree weights
- Random forest weights

Going beyond SLO:
Integrated learning and optimization

Wrong predictions lead to suboptimal decisions

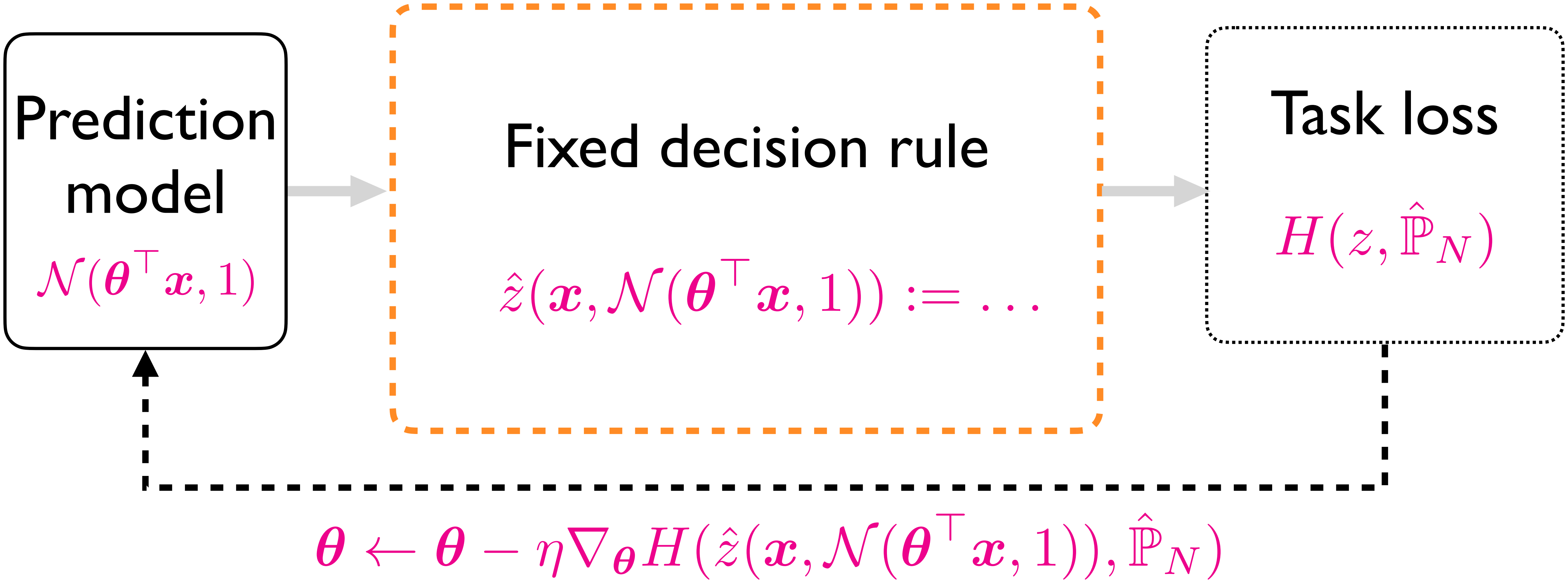
$$\max_{\boldsymbol{z} \in \mathcal{Z}} \boldsymbol{\xi}^\top \boldsymbol{z}$$



$$\|\boldsymbol{\epsilon}_1\| \leq \|\boldsymbol{\epsilon}_2\|$$

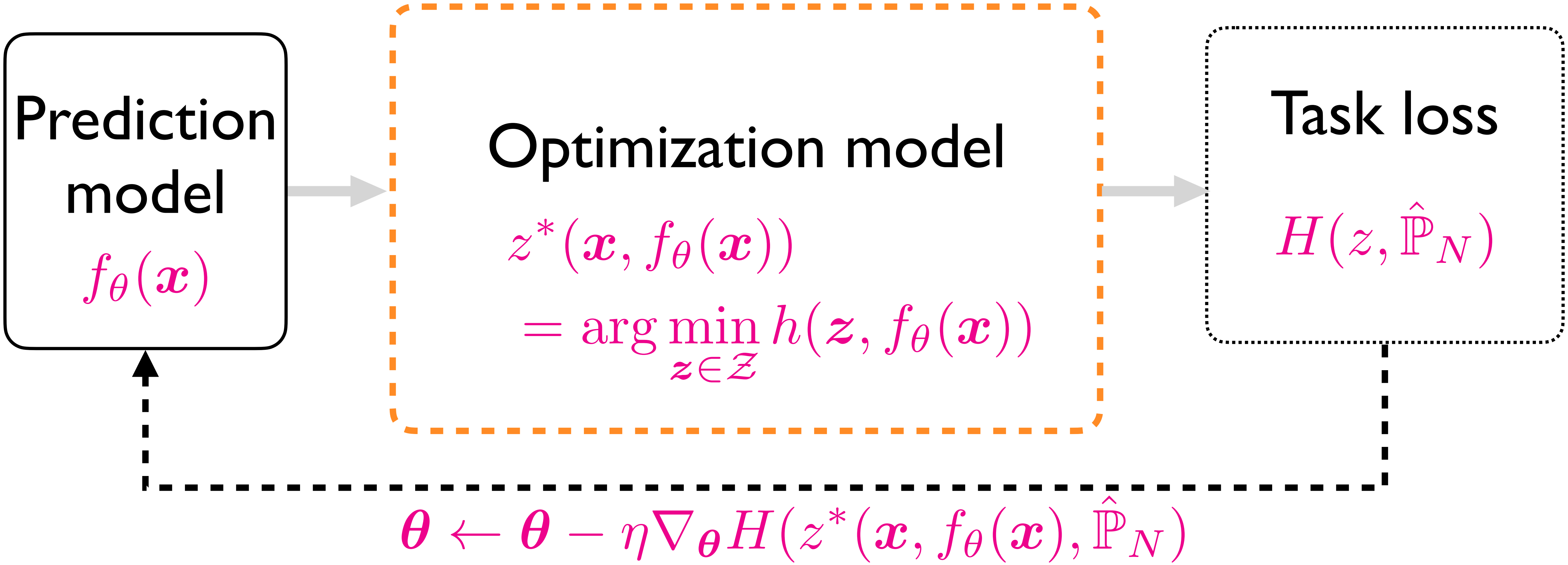
Figure adapted from [Elmachtoub and Grigas 2022]

ILO Training pipeline



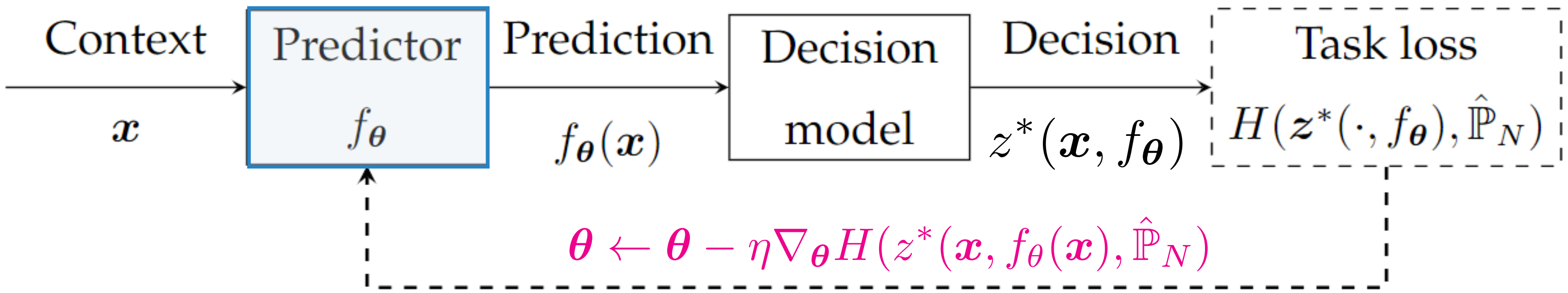
- [Bengio 1997] : Task-aware point prediction under a **fixed decision rule**

ILO Training pipeline



- [Bengio 1997] : Task-aware point prediction under a **fixed decision rule**
- [Donti et al. 2017] : Task-aware conditional density prediction under **CSO model**

How to differentiate through argmin operation



- Implicit differentiation through KKT conditions for convex problems
- Unroll the operations made by the optimization process:
  - Differentiate through its computational graph
  - Implicit differentiation of the fixed point equations at local optimum [Butler and Kwon, 2023] and [Kotary et al. 2023]
- Replace optimizer with a differentiable deep neural network [Grigas et al. 2021]
- Libraries: TorchOpt [Bilevel], CvxpyLayer [Convex], PyEPO [Linear]

Smart "Predict, then optimize"

- Regret minimization [Elmachtoub & Grigas, 2022]:

$$H(z^*(\boldsymbol{x}, f_\theta), \mathbb{P}) \ := \mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi})]$$

Smart "Predict, then optimize"

- Regret minimization [Elmachtoub & Grigas, 2022]:

$$H(z^*(\boldsymbol{x}, f_\theta), \mathbb{P}) \ := \mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi})] \ \ \mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi}) - \min_{\boldsymbol{z} \in \mathcal{Z}} c(z, \boldsymbol{\xi})]$$

Smart "Predict, then optimize"

- Regret minimization [Elmachtoub & Grigas, 2022]:

$$H(z^*(\boldsymbol{x}, f_\theta), \mathbb{P}) \;:= \sout{\mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi})]}\; \mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi}) - \min_{\boldsymbol{z} \in \mathcal{Z}} c(z, \boldsymbol{\xi})]$$

  - Non-convex and discontinuous in θ
  - Replace with SPO+: $\min_{\theta} \mathbb{E}_{\mathbb{P}} \left[\ell_{\mathrm{SPO+}}(g_\theta(\boldsymbol{x}), \boldsymbol{y})\right]$
    with

$$\ell_{\mathrm{SPO+}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) := \sup_{\boldsymbol{z} \in \mathcal{Z}} (\boldsymbol{y} - 2\hat{\boldsymbol{y}})^T \boldsymbol{z} + 2\hat{\boldsymbol{y}}^T \boldsymbol{z}^*(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{y}^T \boldsymbol{z}^*(\boldsymbol{x}, \boldsymbol{y}),$$

  - Solve two optimization problems (MILP) at each iteration
  - SPO+ has slower convergence rate when compared to sequential estimate then optimize model
  - Model misspecification: SPO+ outperforms MSE

Optimal action imitation

- Imitation performance metric:

$$H(z^*(\boldsymbol{x}, f_\theta), \mathbb{P}) \; := \mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi})]$$

Optimal action imitation

- Imitation performance metric:

$$H(z^*(\boldsymbol{x}, f_\theta), \mathbb{P}) := \mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi})] \quad \mathbb{E}_{\hat{\mathbb{P}}_N}[d(z^*(\boldsymbol{x}, f_{\boldsymbol{\theta}}), z^*(\boldsymbol{x}, \boldsymbol{\xi}))]$$

Optimal action imitation

- Imitation performance metric:

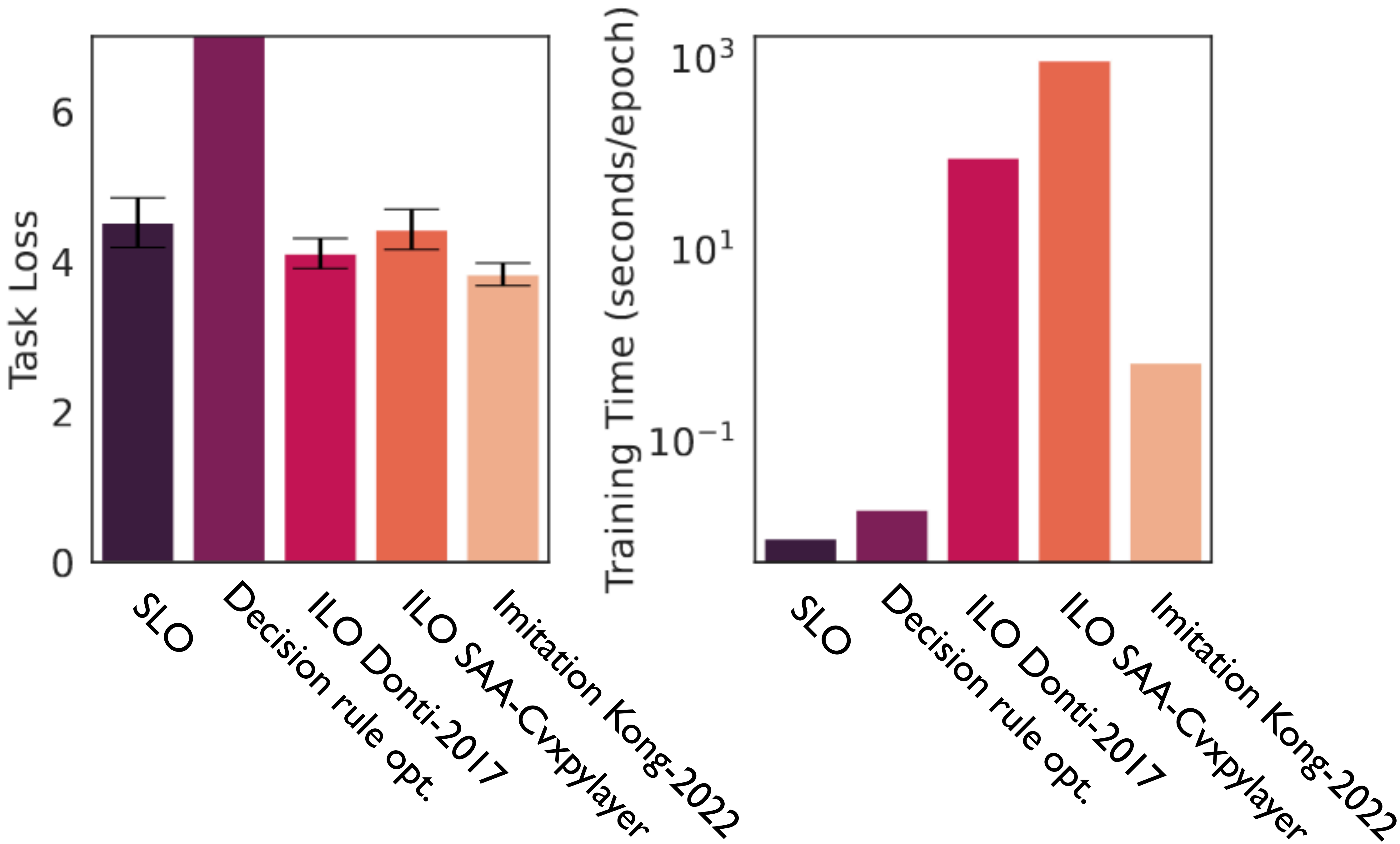
$$H(z^*(\boldsymbol{x}, f_\theta), \mathbb{P}) \;:= \cancel{\mathbb{E}_{\mathbb{P}}[c(z^*(\boldsymbol{x}, f_\theta), \boldsymbol{\xi})]} \quad \mathbb{E}_{\hat{\mathbb{P}}_N}[d(z^*(\boldsymbol{x}, f_{\boldsymbol{\theta}}), z^*(\boldsymbol{x}, \boldsymbol{\xi})]$$

- Training based on perturbed optimizers:
    - [Berthet et al., 2020] uses additive perturbation of point prediction
    - [Dalle et al. , 2022] uses multiplicative perturbations
    - [Mulamba et al., 2021] and [Kong et al., 2022] uses energy based optimizer

$$\tilde{z}(\boldsymbol{x}, f_\theta) \sim \frac{\exp(\alpha h(\boldsymbol{z}, f_\theta(\boldsymbol{x}))}{\int \exp(\alpha h(\boldsymbol{z}, f_\theta(\boldsymbol{x}))d\boldsymbol{z}}$$

Comparison of some models

Load forecasting and generator scheduling problem
(objective similar to newsvendor problem)



Source: [Kong et al. 2022]

Take-away messages

- Contextual stochastic optimization is a rapidly evolving field that provides methods for identifying data-driven decision that exploit most recently available information.
- Three types of approaches:
  - Decision rule/policy optimization
  - Sequential learning and optimization
  - Integrated learning and optimization
- Four types of performance measures:
  - Statistical accuracy of prediction model
  - Task-based expected cost of induced policy
  - Task-based expected regret of induced policy
  - Quality of imitation
- Many potential applications in mining?

(Link to survey paper)