UNIVERSITÄT WÜRZBURG · KU LEUVEN · Loughborough University · NEURAL INFORMATION PROCESSING SYSTEMS

Online Decision-Focused Learning under Semi-Bandit Feedback

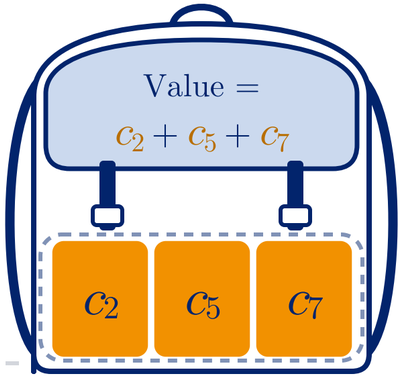
How do you train a model to decide well when it only observes what it picked?

Aabhash Dhakal[1], Tim Lachner[1], Jayanta Mandi[2,3], Marco Foschini[2], Christoph M. Flath[1]

[1] University of Würzburg · [2] KU Leuven · [3] Loughborough University

Pull the arms that fit the knapsack. Only these costs are revealed.



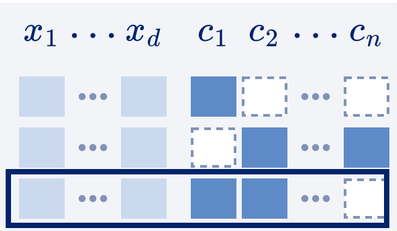


DATA DRIVEN DECISIONS · CAIDAS

Online Decision-Focused Learning Pipeline

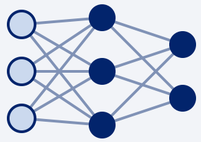
repeated every round $t = 1, \dots, T$

Online rounds



features x · edge costs c

ML model



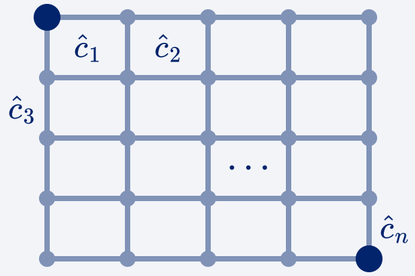
$\hat{c} = f_\omega(x)$

Predictions

$$\hat{c} = \begin{bmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \vdots \\ \hat{c}_n \end{bmatrix}$$

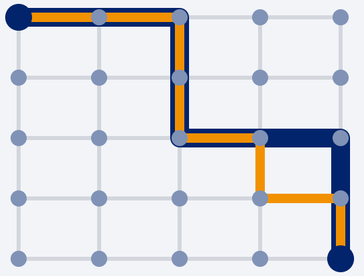
cost per edge

Optimization



$w^*(\hat{c}) = \arg\min_{w \in \mathcal{W}} \hat{c}^\top w$

Decision



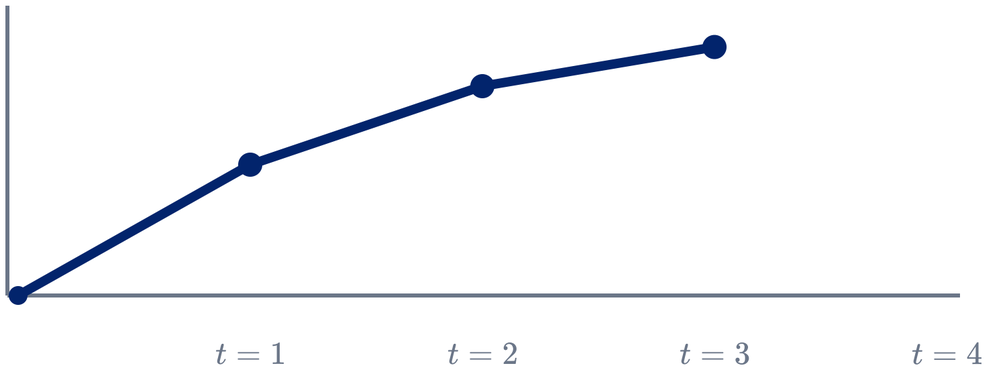
$w^*(\hat{c})$ $w^*(c)$

prediction loss $\|\hat{c} - c\|^2$

decision loss (regret) $c^\top w^*(\hat{c}) - c^\top w^*(c)$

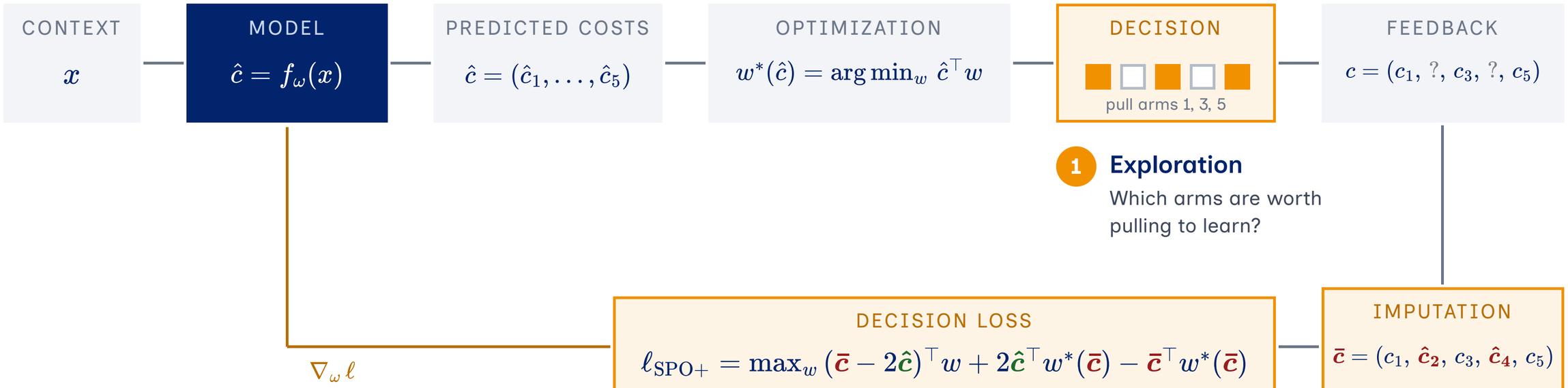
cumulative regret

illustrative



Unknown under semi-bandit feedback:
only chosen edges reveal their cost.

Partial feedback breaks the DFL loop in three places



PRIOR WORK

	Offline	Online
Full	Elmachtoub & Grigas	Capitaine et al.
Semi-bandit	Hu et al.	**This work**

...but filling the gaps with your own predictions sets a trap.

Model's guesses become ground truth

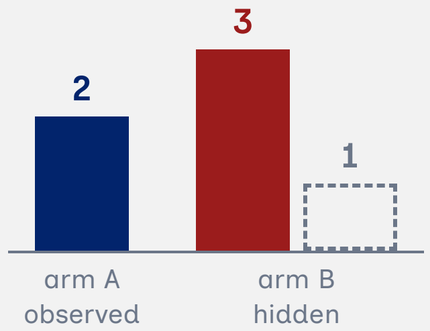
IMPUTATION DEGENERACY

1. Fill gaps with your own predictions
2. The loss rewards matching them
3. Bias flows into the shared features
4. Next imputations get worse

self-reinforcing under partial feedback

BayesianSPO breaks this loop. Next: how.

A TWO-ARM EXAMPLE



imputed
$\bar{c}$ is wrong here

true cost, never observed

$\bar{c} = (2,\ \mathbf{3})$
completed costs

Predicts the truth

$\hat{c} = (2,\ 1)$



picks **arm B**

Training loss vs. $\bar{c}$: **3**

Real loss vs. true c: **0**

Copies the imputation

$\hat{c} = (2,\ \mathbf{3})$



picks **arm A**

Training loss vs. $\bar{c}$: **0**

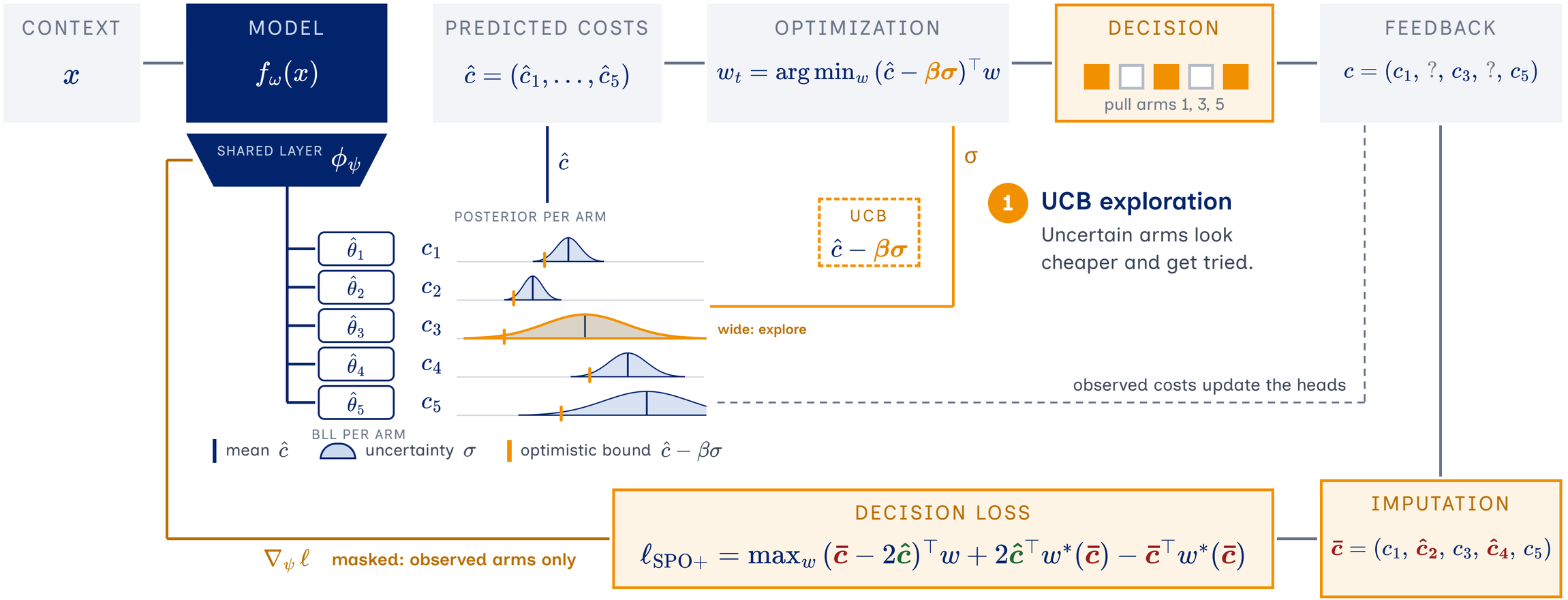
Real loss vs. true c: **3**

The training target is wrong, so the loss rewards the wrong model.

UNIVERSITÄT WÜRZBURG

BayesianSPO answers all three

NEURAL INFORMATION PROCESSING SYSTEMS



Masking and per-arm posteriors keep the imputation grounded.

3 Masked SPO+

Only observed arms send gradients to the features.

2 Posterior mean

Each head fills its own missing cost.

UNIVERSITÄT WÜRZBURG · KU LEUVEN · Loughborough University · DATA DRIVEN DECISIONS · CAIDAS

NEURAL INFORMATION PROCESSING SYSTEMS

BayesianSPO

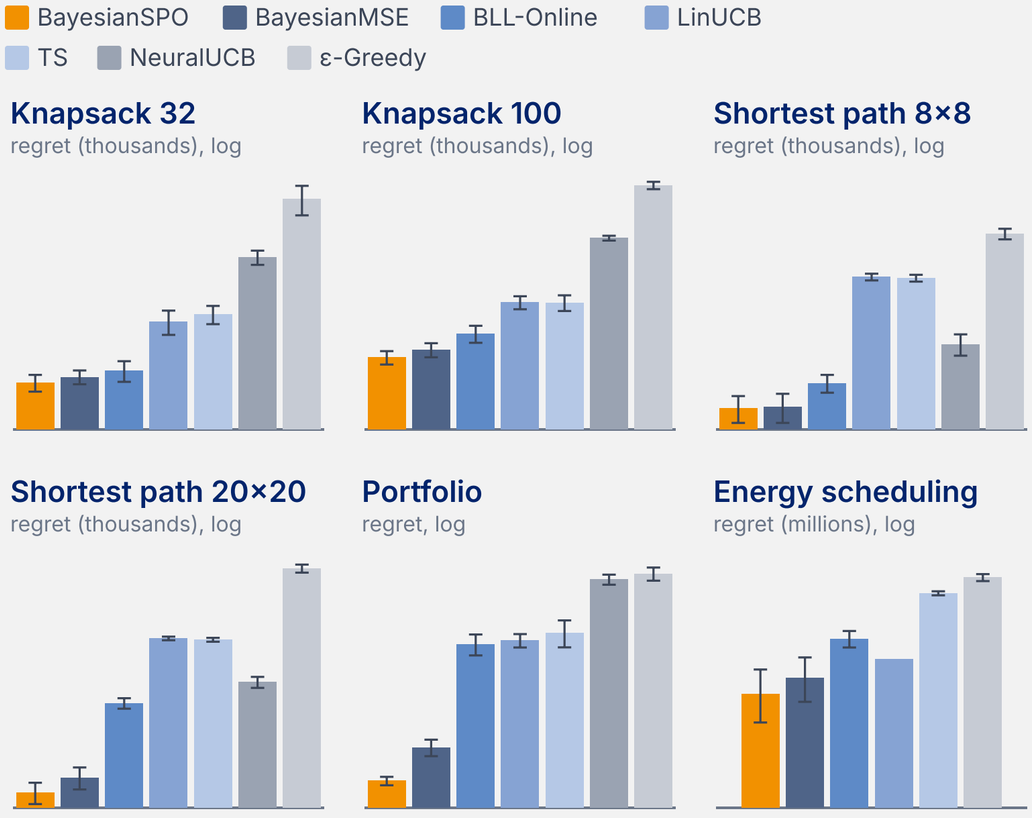
Details and full results at the poster.

POSTER

You saw this paper with semi-bandit feedback: only the parts we selected. We look forward to hearing your full feedback.
Visit us:

Tue, Dec 8, 2026 · 5:00–8:00 PM AEDT · Hall 1-4

RESULTS · LOWEST MEAN REGRET ON 6 OF 6



Final cumulative regret (log scale), mean ± 1 s.d. Knapsack, shortest path, portfolio: T = 10,000, 10 seeds. Energy: 395 days, 30 seeds. BayesianSPO beats BayesianMSE significantly on all but SP 8×8 (one-sided Wilcoxon, $p < 0.05$).