OPTIMAL TREATMENT ASSIGNMENT FROM OBSERVATIONAL DATA: A DECISION-FOCUSED LEARNING APPROACH VIA PSEUDO LABELS

Jiaqi Yang[1], **Zicheng Su**[1,*], **Zhichao Zou**[2], **Peng Zhen**[2], **Wanjing Ma**[1], **Kun An**[1,*]
[1]Tongji University, [2]Didi Chuxing
2410196@tongji.edu.cn, suzicheng@tongji.edu.cn, zouzhichao@didiglobal.com, zhenpeng@didiglobal.com, mawanjing@tongji.edu.cn, kunan@tongji.edu.cn

ABSTRACT

Causal decision-making (CDM) stands as a critical issue in the field of causal inference, as it directly measures the final utility generated by causal effect estimation. In existing literature, the CDM problem typically adopts the predict-then-optimize framework to integrate modules from Machine Learning (ML) and Operations Research. The first step leverages a causal ML model to predict the treatment effect; the second step solves the decision-making problem based on the predictions from the first step. However, due to the propagation of prediction errors from the ML model, the quality of the final decision often remains suboptimal. Decision-Focused Learning (DFL) is an end-to-end modeling paradigm that directly incorporates the ultimate decision loss into the loss function during the prediction model training phase, enabling the ML model to directly maximize the quality of the ultimate decision. Nevertheless, the generalized application of DFL to CDM problems is non-trivial. A core challenge arises from the counterfactual problem: the ML model cannot obtain the ground truth of the treatment effect for each individual. This renders the calculation of decision loss infeasible, thereby impeding the training process of the DFL. In this study, we first define a generalized formulation of the causal treatment assignment problem and theoretically demonstrate the potential advantages of DFL in this context. Furthermore, we propose a Decision-Focused Learning via Pseudo Labels (DFL-PL) approach, which improves the learning process of traditional two-step meta-learner approaches. By enhancing the training pipeline with pseudo-outcomes, our approach enables the calculation of decision loss and the backpropagation of this loss for model training. Finally, we validate the effectiveness of the proposed algorithm on both synthetic datasets and real-world data from Didi Chuxing.

1 INTRODUCTION

Causal Decision Making (CDM) represents an integrated field that combines causal inference and decision theory (Ge et al., 2025). Its primary objective is to address the inadequacy of conventional predictive models in intelligent decision-making systems when tackling counterfactual problems (Feuerriegel et al., 2024). Typically functioning as a framework that integrates causal machine learning (ML) models with decision models, CDM aims to achieve high-quality decision-making outcomes (Athey & Wager, 2021). Moreover, CDM elucidates how causal inference can provide support for decision-making systems (Cavenaghi et al., 2024).

Existing work has mainly focused on the first stage of CDM, namely *causal effect estimation* (Winship & Morgan, 1999). Initially, given historical data, causal machine learning models are trained with the goal of minimizing the fitting error of causal effects, while incorporating covariate information closely related to the problem (Yao et al., 2021). Common implementation methods include causal forests (Wager & Athey, 2018; Athey et al., 2019), neural network-based approaches (Shalit et al., 2017; Johansson et al., 2022), and meta-learners (Künzel et al., 2019; Nie & Wager, 2021; Kennedy, 2023). Regarding the decision-making stage, one major research stream is to directly learn an optimal assignment policy, namely *causal policy learning*, which selects an optimal pol-

Table 1: Comparison of related work on constrained decision-making (CDM) problems

Ref	Method	Data Assumption	Treatment assignment problem
Athey & Wager (2021)	Weighted classification	RCT	Problem without constraint
Zhao et al. (2017)	Regression that maximizes expected response	RCT	Problem without constraint
Fernández-Loría et al. (2023)	Weighted classification	RCT	Problem without constraint
Devriendt et al. (2020)	Learning to rank	RCT	Top-k problem
He et al. (2024)	Learning to rank	RCT	Top-k problem
Betlei et al. (2021)	Learning to rank	RCT	Top-k problem
Fernández-Loría & Provost (2022)	Regression on the proxy target	RCT	Top-k problem
Ai et al. (2022)	Regression that maximizes the heterogeneity	RCT	Cost-efficient problem
Casacuberta & Hardt (2026)	Low-estimation method	RCT	Cost-efficient problem
Zhou et al. (2023)	Learning to rank	RCT	Cost-efficient problem
Kamran et al. (2024)	Learning to rank	Observational	Top-k problem
Vanderschueren et al. (2024)	Learning to rank	Observational	Top-k problem
Zhou et al. (2024)	Decision-Focused Learning	RCT	Cost-efficient problem
Zhang et al. (2025)	Decision-Focused Learning	Both	Cost-efficient problem
Ours	Decision-Focused Learning	Observational	General problem with constraints

icy from a specific policy class (Yadlowsky et al., 2025; Chernozhukov et al., 2022)). In practice, CDM is often subject to constraints such as total resource limits or fairness requirements (Athey, 2017), which can make decision-making complex and potentially NP-hard. Therefore, it is essential to construct and solve an optimization model that aligns with the decision-maker's objectives to achieve optimal decisions. The pipeline that integrates causal machine learning models with the constructed optimization models constitutes the most general CDM framework, which is referred to as the predict-then-optimize (PTO) framework in the literature (Elmachtoub & Grigas, 2022).

However, from the perspective of PTO integration, achieving more accurate causal estimation is not the ultimate objective. Instead, the focus should be on how to optimize the pipeline to best support decision-making—a crucial point that is often overlooked in existing research (Fischer-Abaigar et al., 2024). While the traditional two-stage "predict-then-optimize" framework persists in pursuing the most accurate predictions, it may not necessarily yield optimal decisions due to the propagation of errors (Bengio, 1997). A more appealing approach is to conduct joint modeling of prediction and decision-making, directly optimizing the final decision performance, which is named Decision-Focused Learning (DFL) (Mandi et al., 2022). Due to its end-to-end modeling perspective, the application of DFL to CDM offers two core advantages: it standardizes the evaluation metric of the PTO framework (i.e., decision loss) and provides methodological and theoretical support for learning decision loss during the prediction phase. A key challenge in DFL is converting constrained optimization problems into differentiable loss functions while maintaining stable training, further complicated by non-convexity and discontinuity of the decision loss. Since Amos and Kolter (Amos & Kolter, 2017) first proposed OptNet to incorporate quadratic programming into end-to-end training, scholars have developed various methods, including direct differentiation (Agrawal et al., 2019), perturbation smoothing (Berthet et al., 2020), and surrogate loss (Elmachtoub & Grigas, 2022). Supported by validation on numerous benchmarks, DFL has demonstrated its outstanding performance across a wide range of scenarios (Mandi et al., 2024; Guo et al., 2026).

Nevertheless, despite the availability of methods for achieving DFL, it appears to conflict with CDM. First and foremost, a core challenge addressed by causal ML models in CDM is the estimation of treatment effects when the ground truth of treatment effects in historical data remains unobservable. In contrast, DFL training computes the decision loss using optimal decisions derived from the ground truth of treatment effects. This logic is intuitive: since DFL aims to learn methods for achieving superior decisions, it naturally requires superior decisions as supervisory signals for training. However, in the pipeline of CDM, neither the ground truth of treatment effects nor the optimal decisions are available, creating a fundamental mismatch (Zhou et al., 2024). Second, integrating DFL into CDM necessitates addressing the compatibility between decision loss and causal loss. Causal learning aims to eliminate biases induced by confounding factors that may impair predictive accuracy, whereas DFL deliberately introduces certain biases to better align predictions with downstream decision tasks (Yang et al., 2025b). Consequently, within the CDM framework, DFL and causal learning must be combined with care so as to preserve the advantages of both.

In Table 1, we summarize related work that seeks to directly maximize downstream decision performance at the prediction stage. The primary differences among existing approaches stem from the specific treatment assignment problems they consider and the underlying data assumptions. First, prior studies typically formulate specific treatment assignment problems and design customized

learning method to match the corresponding decision tasks (e.g., weighted classification for unconstrained problem and learning-to-rank for Top-k selection). This specificity may lead to a lack of transferability, as such methods are often difficult to generalize across different decision settings. In contrast, a key advantage of DFL is its ability to adaptively accommodate diverse decision tasks under arbitrary constraints while maintaining precise alignment with the target decision objective, without requiring expert-crafted, task-specific loss functions. As a result, DFL provides a more general framework for enhancing CDM pipelines. Prior work has provided empirical validation of DFL in large-scale subsidy allocation problems, notably in Zhou et al. (2024) and Zhang et al. (2025).

Regarding data assumptions, most existing works rely on the availability of randomized controlled trial (RCT) data. The central advantage of RCT data lies in the independence between treatments and covariates, which enables unbiased estimation of policy performance via expected outcome metrics (Zhao et al., 2017; Ai et al., 2022). When sufficient RCT data are available, this setting approximately resolves the counterfactual inference problem, thereby enabling the computation of decision loss and offline policy evaluation—both of which are critical for DFL. However, the most prevalent and challenging CDM setting is one in which only historical observational data are available. First, RCT data are often difficult to obtain, as conducting randomized experiments typically incurs substantial costs or entails policy risks. Second, selection bias inherent in observational data induces a cascade of adverse effects: although existing methods can debias observational data to recover unbiased estimates of treatment effects, they often suffer from high prediction variance. Third, the dependence between covariates and treatments in observational data prevents straightforward evaluation of policy performance like RCT data, further limiting the applicability of DFL.

In this paper, we first define a general treatment assignment problem, then quantify the decision quality for this problem and analyze the differences between learning the most accurate treatment effect and pursuing the optimal decision in this problem. Concurrently, We propose a DFL-based causal learning method, termed *Decision-Focused Learning via Pseudo-Labels* (DFL-PL), which constructs a decision loss using pseudo-labels to enhance the decision performance of causal learning. Specifically, our key contributions are outlined as follows:

- We first define a class of CDM problems based on the PTO framework. In the first stage, historical observational data can be used to estimate the treatment effects. In the second stage, based on the estimated values, our goal is to maximize decision performance in a general treatment assignment problem.
- We quantify the evaluation metrics for CDM problems and discuss the differences between learning optimal decisions and learning treatment effects. This allows us to explore the potential sources of gains from decision-focused learning in this scenario and guide improvements to existing CDM method.
- We design a pipeline that integrates causal learning and decision-focused learning for observational data. The proposed method enhances causal meta-learners by constructing pseudo-outcomes and incorporating decision loss, and trains the entire pipeline using a cross-fitting strategy. This enables the algorithm to directly optimize decision performance, leading to improved decision quality.

We conduct model testing using both synthetic data and real-world data. For the synthetic data experiments, we include both simple and complex decision scenarios; for the real-world data experiments, we select city-level subsidy assignment data from Didi Chuxing. Across all tested problems, our framework demonstrates higher decision quality, which verifies the superiority of our approach.

2 Problem Formulation

2.1 Problem Definition

We first define the CDM problem studied in this paper. Our goal is to derive a treatment assignment policy for a batch of individuals $\mathcal{I} = \{1, \ldots, i, \ldots, I\}$, where I denotes the total number of individuals. The treatment policy is denoted by $\boldsymbol{t} = [t_1, \ldots, t_i, \ldots, t_I]$, where the treatment decision for each individual i is binary, i.e., $t_i \in \{0, 1\}$. Accordingly, each individual i has two potential outcomes, $y_i(0)$ and $y_i(1) \in \mathbb{R}$, corresponding to $t_i = 0$ and $t_i = 1$, respectively. We denote the

covariates of all individuals as $\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_I]$, where $x_i \in \mathbb{R}^d$. Given both potential outcomes, the individual treatment effect is defined as the difference $\tau_i = y_i(1) - y_i(0)$. Throughout this paper, we adopt the standard assumptions (Yao et al., 2021), summarized in Appendix A.

Due to practical limitations such as budget constraints and fairness considerations, the treatment assignment policy is required to satisfy a set of constraints, i.e., the decision must lie within a predefined feasible region $\mathcal{S}$. Our objective is to maximize the total treatment effect across all individuals i who receive the treatment. This objective can be formulated as the following constrained optimization problem (Wager, 2024):

$$o(\boldsymbol{\tau}) = \max_{\boldsymbol{t} \in \mathcal{S}} \sum_{i \in \mathcal{I}} \tau_i \cdot t_i \tag{1}$$

Here, the vector $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_i, \ldots, \tau_I]$ represents the realized individual treatment effects. Since $\boldsymbol{\tau}$ is generally unknown and unobservable prior to decision making, conditioning on covariates x refines Equation 1 into a formulation that maximizes the conditional expectation of the treatment effect as follows:

$$o(\boldsymbol{\tau}) = \max_{\boldsymbol{t} \in \mathcal{S}} \sum_{i \in \mathcal{I}} \mathbb{E}\big[\tau_i \,|\, x_i\big] \cdot t_i \tag{2}$$

where $\mathbb{E}\big[\tau_i \,|\, x_i\big]$ represents the conditional average treatment effect (CATE) for an individual i such that (Rubin, 1974):

$$\mathbb{E}[\tau_i \,|\, x_i] = \mathbb{E}[y_i(1) - y_i(0) \mid x_i] \tag{3}$$

Consequently, we can define the treatment assignment policy as: $\pi : \boldsymbol{x} \to \boldsymbol{t} \in \mathcal{S}$, which denotes the mapping from the covariates $\boldsymbol{x}$ to the assignment decisions $\boldsymbol{t}$. Furthermore, the optimal policy π^* is defined as the mapping that maximizes the objective function in Equation 2 while satisfying the constraints. Existing studies related to CDM typically focus on unconstrained optimization tasks (see, e.g., Athey & Wager (2021); Fernández-Loría & Provost (2022)) or on settings where constraints can be relaxed into unconstrained ones (see, e.g., Zhou et al. (2024); Zhang et al. (2025)). Under such settings, obtaining π^* reduces to a threshold-based solution with independent determination for each individual i, where the condition $\mathbb{E}[\tau_i \,|\, x_i] > 0$ yields the optimal policy $\pi(x) = \mathbf{1}_{\{\tau(x) > 0\}}$. In contrast, this paper studies constrained optimization tasks in which a nontrivial constraint set $\mathcal{S}$ with hard feasibility requirements must be strictly satisfied, thereby shifting the CDM solution from the individual level to the group level. In this setting, treatment assignment for an individual depends not only on their own attributes, but also on their relative standing among other individuals.

2.2 Predict-then-Optimize Framework

As the ground truth of the CATE is unobservable prior to decision-making, the common practice is to predict τ in advance based on historical observational data. This methodology lies within the well-established Predict-then-Optimize (PTO) framework, which concatenates a prediction mapping ϕ with learnable parameters and an optimization mapping ψ such that:

$$\pi = \psi \circ \phi_\theta \tag{4}$$

The PTO pipeline is promising for two reasons: on the one hand, this framework enables the direct leverage of existing well-established CATE estimation methods, such as double machine learning and causal forests (Wager & Athey, 2018); on the other hand, alternative frameworks such as end-to-end learning struggle to address decision-making problems with constraints (Fernández-Loría et al., 2023). Given the historical observational data $\mathcal{D} = \{x_j, t_j, y_j\}_{j=1}^{M}$, where j is the sample index and M denotes the number of data samples, the PTO pipeline can be formulated as follows:

Step 1 (CATE prediction with $\phi_\theta : \boldsymbol{x} \to \hat{\boldsymbol{\tau}}$): Employ a causal machine learning model to estimate the CATE from the covariates. The predicted treatment effect for each individual is defined as:

$$\hat{\tau}_i = \phi_\theta(x_i), \quad \forall i \in \mathcal{I} \tag{5}$$

Step 2 (Treatment optimization with $\psi : \hat{\boldsymbol{\tau}} \to \boldsymbol{t}$): Given the predictions $\hat{\tau}_i$ obtained in Step 1, replace the unobservable expectation $\mathbb{E}[\tau_i \mid x_i]$ in Equation 2 with the predictions. Solving the resulting optimization problem yields the final treatment assignment. This mapping is defined as:

$$\boldsymbol{t} = \psi(\hat{\boldsymbol{\tau}}) = \arg\max_{\boldsymbol{t} \in \mathcal{S}} \sum_{i \in \mathcal{I}} \hat{\tau}_i t_i \tag{6}$$

It is worth noting that the optimization mapping ψ is implemented by solving a fixed surrogate optimization model that relies on the predictive results of ϕ_θ. When ϕ_θ achieves high predictive accuracy, the resulting policy π can yield near-optimal decisions; conversely, low predictive accuracy may leads to suboptimal outcomes. Consequently, the core of the PTO-based treatment assignment problem lies in the mapping ϕ_θ, as it directly determines the quality of the treatment assignment $\boldsymbol{t}$. As the fitting error of the CATE is inevitable, this two-stage PTO framework may lead to suboptimal solutions. The suboptimality of each solution can be quantified using the *regret* metric, defined as follows (Elmachtoub & Grigas, 2022):

$$regret = \boldsymbol{\tau}^\top \boldsymbol{t}^*(\boldsymbol{\tau}) - \boldsymbol{\tau}^\top \boldsymbol{t}^*(\hat{\boldsymbol{\tau}}) \tag{7}$$

where $\boldsymbol{\tau} = [\tau_1, ..., \tau_I]$ is the vector of the realized value of treatment effect for all individuals in a batch. And $\boldsymbol{t}^*(\cdot)$ is the optimal decision based on the input:

$$\boldsymbol{t}^*(\boldsymbol{\tau}) = \arg\max_{\boldsymbol{t}\in\mathcal{S}} \sum_{i\in\mathcal{I}} \tau_i t_i \tag{8}$$

$$\boldsymbol{t}^*(\hat{\boldsymbol{\tau}}) = \arg\max_{\boldsymbol{t}\in\mathcal{S}} \sum_{i\in\mathcal{I}} \hat{\tau}_i t_i \tag{9}$$

Although the *regret* cannot be computed exactly due to the absence of ground-truth treatment effects in observational data, it remains a precise metric for CDM, quantifying the gap between a given policy and the theoretically optimal one (Athey & Wager, 2021; Fernández-Loría et al., 2023).

Therefore, the treatment assignment problem we study can be summarized as: minimizing the *regret* of treatment assignment decisions by optimizing the mapping ϕ_θ.

2.3 Open Issues in PTO-based Treatment Assignment Problems

In PTO framework, an excellent mapping ϕ_θ requires two key factors: on the one hand, ϕ_θ needs to learn the causal effects implicit in historical data; on the other hand, it must align with our ultimate goal: minimizing the regret. This poses strict challenges to the construction of ϕ_θ:

1. **No ground truth for realized causal effect**: We aim to learn a mapping from the individual covariate vector x to the CATE τ. However, accurately estimating the CATE from observational data is challenging for two main reasons. First, only one historical outcome y is observed for each individual, while the counterfactual outcome remains unobserved, making it impossible to directly obtain $\tau_{\text{realized}} = y(1) - y(0)$. As a result, learning $\mathbb{E}\big[\tau \,|\, x\big]$ without access to ground-truth treatment effects is inherently challenging. Second, observational data suffer from non-negligible selection bias, which further increases the variance and instability of CATE estimation. Since the policy learning process follows a two-stage framework, estimation errors in the first stage can propagate to the decision-making stage, ultimately leading to suboptimal decisions (Loke et al., 2022).

2. **No ground truth for optimal decision**: The key method to address the cumulative error in the two-stage framework is to perform end-to-end modeling of the prediction mapping ϕ_θ, incorporating downstream decision information during the training of the prediction model to maximize decision quality, which is commonly referred to as Decision-focused Learning (DFL) (Mandi et al., 2024). DFL requires optimal decisions from historical data as supervision, which in turn demands either knowledge of the historical optimal decisions or the realized treatment effects τ_{realized}; neither condition holds in purely observational settings.

Due to the particularity of this scenario, the effectiveness of treatment assignment decisions is difficult to guarantee. Therefore, our aim to implement a decision-focused mapping ϕ_θ to maximize the decision quality.

2.4 Qualitative Insights

In the PTO framework, the true value vector $\boldsymbol{\tau}$ is unobservable, and decisions are obtained by solving a surrogate optimization problem whose objective is defined by a predicted value vector $\hat{\boldsymbol{\tau}}$. Formally, the decision is: $\boldsymbol{t}^*(\hat{\boldsymbol{\tau}}) = \arg\max_{\boldsymbol{t}\in S} \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}$. Intuitively, if the prediction is sufficiently accurate, the surrogate problem should reproduce the optimal decision of the original model. However, a fundamental observation is that **perfect prediction is not required** for achieving the optimal

decision Casacuberta & Hardt (2026). Even when $\hat{\boldsymbol{\tau}}$ differs from $\boldsymbol{\tau}$ in magnitude or direction, the PTO framework may still return the optimal decision.We formalize this observation as follows, with the proof and an illustrative toy example deferred to Appendix A.

Theorem 1 (Tolerance to prediction errors)**.** *Let $\boldsymbol{t}^* = \boldsymbol{t}^*(\boldsymbol{\tau})$ be the optimal solution to the problem equation 2. If the predicted vector $\hat{\boldsymbol{\tau}}$ satisfies*

$$\hat{\boldsymbol{\tau}}^\top \boldsymbol{t}^* \geq \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}, \qquad \forall \boldsymbol{t} \in S,$$

then the surrogate problem $\max_{\boldsymbol{t}\in S} \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}$ returns $\boldsymbol{t}^*$.

Even when prediction errors exist, some errors do not change the optimal decision while others significantly deteriorate it. The precise condition for preserving optimality is given below.

Theorem 2 (Necessary and sufficient condition for optimality)**.** *Let $\mathrm{ext}(S)$ denote the extreme points of S. Then $\boldsymbol{t}^*(\hat{\boldsymbol{\tau}}) = \boldsymbol{t}^*$ if and only if*

$$\hat{\boldsymbol{\tau}}^\top \boldsymbol{t}^* \geq \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}, \quad \forall \boldsymbol{t} \in \mathrm{ext}(S).$$

Based on Theorem 2, we understand that achieving optimal decisions does not require accurately predicting the $\boldsymbol{\tau}$; rather, it suffices to ensure that the predicted vector lies within the normal cone defined by a finite set of linear constraints, as shown in Figure 1. Consequently, the learning objective is fundamentally to satisfy a set-based constraint, rather than to approximate a single scalar or vector. In sharp contrast, conventional supervised learning models—typically rely on a single point as the supervision signal, requiring the model to fit it as precisely as possible. Even deviations that have no impact on the final decision are penalized, thereby increasing the difficulty of the learning task. By contrast, leveraging structured supervision via the normal cone constraint allows the model to focus on ensuring decision optimality, reducing unnecessary overfitting while improving the decision quality.

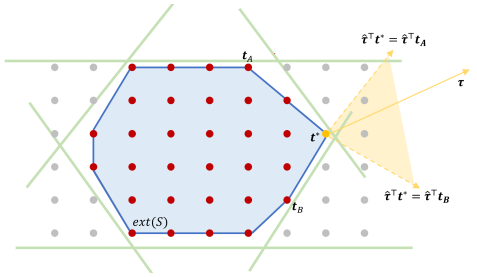


Figure 1: Illustration of the normal cone for optimal decision. Red points denote feasible solutions, the blue region is the convex hull, and the normal cone inducing the optimal solution is determined by a subset of extreme points (i.e., A and B). When the predicted vector lies in the normal cone, the resulting decision remains optimal.

3 Decision-focused Learning via Pseudo Labels

3.1 Causal meta-learner

We have paid attention to the causal meta-learner, which typically follow a two-step paradigm: in the first step, some base models (e.g., propensity score models) are fitted; in the second step, the fitted base models are combined for building the final predictive model. The meta-learner defines a flexible learning framework, not a specific model, which can be naturally combined with DFL. In this section, we introduce a DFL approach based the DR-learner (Kennedy, 2023), and our method can be extended to some other meta-learners, such as X-learner, R-learner (Künzel et al., 2019). The DR-learner is well suited for observational data because it explicitly accounts for selection bias by modeling both the outcome and the propensity score (Kennedy, 2023); its standard procedure is described in Appendix B.

3.2 Decision-focused learning

In standard DFL training, both the decision loss and evaluation metric adopt the regret as shown in Equation 7. This requires us to solve the CDM problem using an arbitrary solver to obtain the prediction-based optimal decision and the theoretically optimal decision respectively. Then we need to calculate the regret based on the realized value of treatment effects, i.e., $\boldsymbol{\tau}$. In our scenario, since $\boldsymbol{\tau}$ is unavailable, we use the pseudo labels $\tilde{\boldsymbol{\tau}}$ to replace $\boldsymbol{\tau}$ as follows:

$$\mathcal{L}(\tilde{\boldsymbol{\tau}}, \hat{\boldsymbol{\tau}}) = \tilde{\boldsymbol{\tau}}^\top \boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}) - \tilde{\boldsymbol{\tau}}^\top \boldsymbol{t}^*(\hat{\boldsymbol{\tau}}) \tag{10}$$

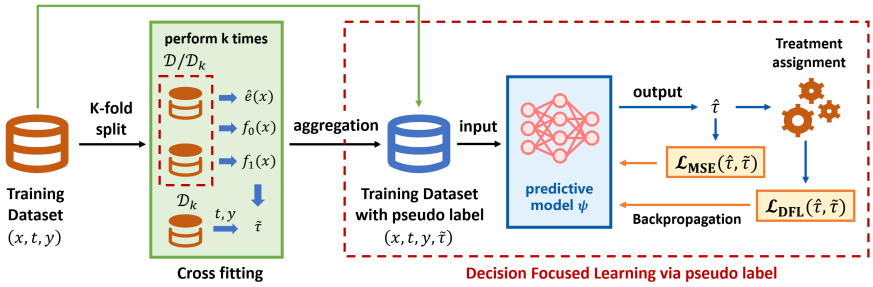


Figure 2: Framework of the decision-focused learning via pseudo labels

where $t^*(\cdot)$ is the optimal decision based on the input. The idea behind this approach is to find the most suitable substitute when τ is unknown. In the DR-learner, $\tilde{\tau}$ is the most appropriate as an unbiased estimator of τ (Kennedy, 2023; Kamran et al., 2024).

With the decision loss $\mathcal{L}$ defined above, the key step in applying this loss is to backpropagate $\mathcal{L}(\tilde{\boldsymbol{\tau}}, \hat{\boldsymbol{\tau}})$ to the prediction model ϕ_θ, which enables the gradient to be decomposed via the chain rule:

$$\frac{\partial \mathcal{L}(\tilde{\boldsymbol{\tau}}, \hat{\boldsymbol{\tau}})}{\partial \theta} = \frac{\partial \mathcal{L}(\tilde{\boldsymbol{\tau}}, \hat{\boldsymbol{\tau}})}{\partial \boldsymbol{t}^*(\hat{\boldsymbol{\tau}})} \cdot \frac{\partial \boldsymbol{t}^*(\hat{\boldsymbol{\tau}})}{\partial \hat{\boldsymbol{\tau}}} \cdot \frac{\partial \hat{\boldsymbol{\tau}}}{\partial \theta} \tag{11}$$

where the first and third terms can be computed via standard gradient backpropagation. However, computing the second term $\frac{\partial \boldsymbol{t}^*(\hat{\boldsymbol{\tau}})}{\partial \hat{\boldsymbol{\tau}}}$ is more challenging, as there is usually no closed form formula for the optimization mapping $\hat{\boldsymbol{\tau}} \to \boldsymbol{t}^*(\hat{\boldsymbol{\tau}})$ (Mandi et al., 2024).

To address this, we compute gradients via differentiable approximations of $\mathcal{L}(\tilde{\boldsymbol{\tau}}, \hat{\boldsymbol{\tau}})$ that preserve its essential information. This framework is flexible and allows for problem-specific approximate gradient formulations. In this paper, we adopt two methods to compute the gradient: a surrogate loss method (SPO+) and a perturbation-based smoothing method (PFY). The corresponding loss functions and training algorithms are detailed in Appendix B.

3.3 Training preocess via cross fitting

A crucial step in the DR-learner is the regression on pseudo-labels. To avoid the "own observation" bias when estimating the nuisance functions $f_t(x)$ and $e(x)$, we adopt the cross-fitting technique, which has become standard in double machine learning (Robins et al., 2008; Chernozhukov et al., 2018). Specifically, the dataset is partitioned into K folds. For each fold k, the nuisance functions are estimated using all samples outside the fold, and the resulting estimates are then used to construct pseudo-labels for the samples in fold k. These pseudo-labels are subsequently pooled across all folds to train a single DR estimator ϕ_{DR} on the full dataset. This joint training strategy follows the recommended practice in (Chernozhukov et al., 2018), as it yields more stable empirical behavior by aggregating information across folds, rather than fitting fold-specific estimators. Moreover, since DFL requires a complete set of decision instances $\mathcal{I}$ during training, this cross-fitting scheme naturally integrates with the DFL framework and avoids complications induced by fold-wise data separation. The detailed implementation of the DFL via cross-fitting is summarized in algorithm 1.

4 Empirical studies

In this section, we first consider a range of optimization problems on generated data, including both simple assignment problems that have been the focus of prior work Fischer-Abaigar et al. (2024) and more complex combinatorial optimization problems Hans et al. (2004). We then evaluate the method in a real-world setting using a city-level subsidy allocation problem.

4.1 Baseline

In the experiments, all methods employed the same DR-learner pipeline, while using different loss functions during the regression on the pseudo-labels, thereby enabling a systematic comparison.

Algorithm 1 Decision-Focused DR-Learner via Cross-Fitting

Input: Dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^n$; decision loss $\mathcal{L}$
Output: Final estimator $\psi_\theta^\star$

Randomly split index set $[n]$ into K disjoint folds $\{I_k\}_{k=1}^K$ of equal size
for $k = 1, \ldots, K$ **do**
  Construct complement index set $I_k^c = \{1, \ldots, n\} \setminus I_k$
  Fit nuisance models on I_k^c

$$\begin{aligned} \hat{e}^k(x) &= \hat{e}\big((x_i)_{i \in I_k^c}\big), \\ f_0^k(x) &= f_0\Big((x_i)_{i \in J_k^0}\Big), & J_k^0 &= \{i \in I_k^c : t_i = 0\}, \\ f_1^k(x) &= f_1\Big((x_i)_{i \in J_k^1}\Big), & J_k^1 &= \{i \in I_k^c : t_i = 1\} \end{aligned}$$

  foreach $i \in I_k$ **do**
    Construct the DR pseudo label

$$\tilde{\tau}_{k,i} = \frac{t_i - \hat{e}^k(x_i)}{\hat{e}^k(x_i)\big(1 - \hat{e}^k(x_i)\big)}\big(y_i - f_{t_i}^k(x_i)\big) + f_1^k(x_i) - f_0^k(x_i)$$

Fit the optimal estimator by minimizing the decision loss

$$\theta^\star = \arg\min_\theta \sum_{k=1}^K \sum_{i \in I_k} \mathcal{L}\big(\tilde{\tau}_{k,i}, \phi(x_i)\big)$$

return $\phi_{\theta^\star}^\star$

Given the limited baselines for observational data, we extended two ranking-based learning methods to further evaluate the effectiveness of our approach. The following provides brief introductions to the baseline models:

- **MSE**: The loss function is solely MSE, which corresponds to the standard DR-learner.
- **LTR(Pair)**: A sample-based pairwise learning-to-rank method uses a cross-entropy loss Vanderschueren et al. (2024). We extend it to scenarios aimed at cost-effectiveness.
- **LTR(List)**: An extended version of the pairwise ranking, which uses the normalized discounted cumulative gain as the weight for each pairwise objective Vanderschueren et al. (2024). We extend it to scenarios aimed at cost-effectiveness.
- **SPO+(w/o)**: The method based solely on the SPO+ loss, without incorporating the MSE.
- **PFY(w/o)**: The method based solely on the PFY loss, without incorporating the MSE.
- **SPO+(w)**: The method based on a weighted combination of the SPO+ loss and the MSE.
- **PFY(w)**: The method based on a weighted combination of the PFY loss and the MSE.

4.2 Experiments on synthetic data

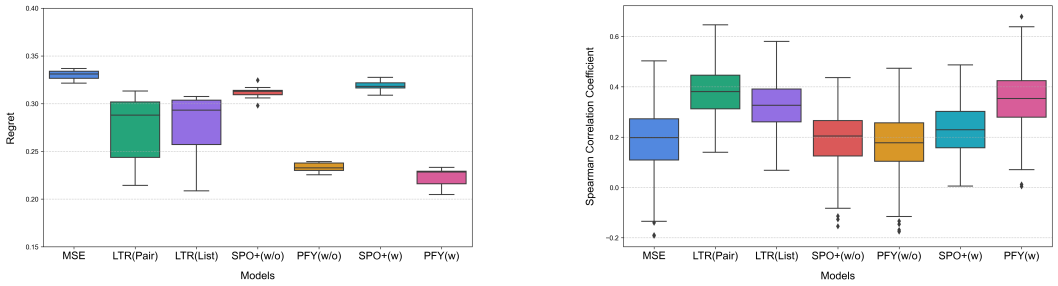
Following the common optimization problems in Fischer-Abaigar et al. (2024) and the combinatorial optimization problems discussed in Hans et al. (2004), we consider four distinct decision-making scenarios. We further follow the data generation procedures in Kamran et al. (2024) and Athey & Wager (2021) to simulate two sets of observational data. Additional details and experimental settings are provided in Appendix C.

We report both the normalized regret and MSE of each method, as shown in Table 2. Our proposed DFL-PL method achieves the lowest two regret metrics across nearly all datasets and tasks. DFL-PL's advantage is further amplified on Dataset 2, where the more complex data generation makes accurate prediction more challenging. LTR(Pair) remains highly competitive in the two simpler scenarios, but its performance becomes unstable in the two more complex settings. For SPO+-based methods, adding the MSE loss minimally affects decision quality but substantially reduces prediction error. PFY also demonstrates strong decision performance, though it is less stable than SPO+; incorporating the MSE loss improves both its prediction accuracy and decision performance.

Table 2: Results of different methods on four well-established treatment assignment problems

			PTO	Learning to Rank		DFL-PL			
Task	Dataset	Metric	MSE	LTR(Pair)	LTR(List)	SPO+(w/o)	PFY(w/o)	SPO+(w)	PFY(w)
Top-k	Dataset1	Regret	4.68 ± 0.70	4.90 ± 0.98	12.47 ± 3.38	4.39 ± 0.78	6.35 ± 1.38	**4.02 ± 0.63**	4.78 ± 1.23
		MSE	32.02 ± 4.54	69.05 ± 17.67	75.29 ± 17.79	40.35 ± 26.37	107.64 ± 12.54	24.35 ± 2.13	24.91 ± 0.85
	Dataset2	Regret	10.89 ± 1.29	5.88 ± 1.05	14.45 ± 7.57	7.48 ± 0.58	**5.59 ± 0.26**	6.97 ± 0.83	5.87 ± 0.25
		MSE	51.92 ± 4.20	196.03 ± 42.17	243.05 ± 85.03	112.68 ± 85.56	288.65 ± 41.55	41.36 ± 3.88	27.22 ± 1.99
CE	Dataset1	Regret	5.95 ± 0.61	5.33 ± 1.61	19.25 ± 21.15	**3.66 ± 0.13**	6.76 ± 2.05	3.84 ± 0.22	5.19 ± 0.29
		MSE	36.68 ± 4.10	117.22 ± 21.43	148.37 ± 91.12	27.86 ± 4.10	161.87 ± 4.78	27.04 ± 1.42	47.28 ± 3.63
	Dataset2	Regret	24.09 ± 1.59	6.43 ± 1.82	14.99 ± 14.35	6.60 ± 0.50	**5.65 ± 0.72**	6.99 ± 0.60	9.53 ± 1.36
		MSE	126.73 ± 11.33	249.54 ± 53.75	263.89 ± 108.42	81.35 ± 10.52	312.94 ± 5.29	57.26 ± 3.84	204.14 ± 27.16
PCKP	Dataset1	Regret	4.91 ± 0.46	15.59 ± 14.09	25.43 ± 15.84	3.59 ± 0.18	5.47 ± 0.68	**3.53 ± 0.22**	4.32 ± 0.30
		MSE	34.58 ± 3.46	146.01 ± 42.23	160.75 ± 50.55	30.47 ± 10.89	164.80 ± 1.44	27.46 ± 2.14	42.61 ± 2.52
	Dataset2	Regret	20.85 ± 1.72	10.35 ± 15.51	10.06 ± 4.07	5.85 ± 0.38	**5.30 ± 0.34**	5.95 ± 0.52	8.00 ± 1.18
		MSE	129.94 ± 15.04	251.75 ± 87.36	258.91 ± 69.78	207.12 ± 37.13	273.36 ± 18.61	68.79 ± 32.82	170.93 ± 31.17
CKP	Dataset1	Regret	5.37 ± 0.54	5.49 ± 2.67	11.09 ± 6.99	3.95 ± 0.54	6.01 ± 1.67	**3.89 ± 0.26**	4.68 ± 0.34
		MSE	35.27 ± 3.02	119.25 ± 21.16	106.50 ± 27.19	50.33 ± 36.91	164.08 ± 3.11	28.20 ± 2.37	46.44 ± 3.33
	Dataset2	Regret	19.84 ± 2.31	11.06 ± 15.94	9.91 ± 4.15	**5.91 ± 0.48**	6.01 ± 1.80	5.98 ± 0.50	8.18 ± 0.70
		MSE	133.64 ± 15.96	271.40 ± 85.05	220.17 ± 81.75	175.10 ± 55.10	288.58 ± 18.55	58.50 ± 3.67	189.83 ± 47.35

4.3 Experiments on real-world data



(a) Regret

(b) Spearman's rank correlation coefficient

Figure 3: Performance on the SA problem

We evaluate the performance of our algorithm using real-world data from DiDi Chuxing (Yang et al., 2025a). The results of the repeated experiments are presented in Figure 3a. As a result, the DFL-PL methods consistently dominate the PTO approach. The LTR method are also effective, with decision regret lying between those of SPO+ and PFY. We further compute the Spearman rank correlation between each predicted vector and the corresponding ground-truth vector, as shown in Figure 3b. Notably, stronger ranking performance does not necessarily translate into better decision outcomes. While the DFL-PL methods and the PTO method exhibit comparable rank correlations, the DFL-PL methods consistently achieve superior decision performance. In contrast, LTR methods attain the highest rank correlation yet fail to deliver the best decisions. This discrepancy suggests that expert-crafted objectives may capture information that is irrelevant to the downstream decision task.

5 Conclusion

In this paper, we propose a CDM framework based on DFL. CDM problems are formulated as a class of constrained optimization problems, where the treatment effect is unknown prior to decision-making and is predicted by causal ML models. When traditional causal ML models are applied to causal decision-making, the quality of decisions may be suboptimal due to the disconnection between the two stages of prediction and optimization. To address this issue, we propose an algorithm that integrates the DFL approach into causal meta-learners to improve the decision quality of causal ML models. The proposed DFL algorithm is validated on generated data and real world data. The results demonstrate that DFL can effectively improve the quality of causal decisions, leading to a significant reduction in decision regret. This work demonstrates the potential of implementing DFL using surrogate information, paving the way for future research in DFL and causal learning.

6 ACKNOWLEDGMENTS

Financial support from National Key R&D Program of China (2024YFB4303100), National Natural Science Foundation of China (72361137005, 52302411), CCF-DiDi GAIA Collaborative Research Funds (No. 202518) and Shanghai Academic Research Leader Program (23XD1404200) are gratefully acknowledged.

REFERENCES

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

Meng Ai, Biao Li, Heyang Gong, Qingwei Yu, Shengjie Xue, Yuan Zhang, Yunzhou Zhang, and Peng Jiang. Lbcf: A large-scale budget-constrained causal forest algorithm. In *Proceedings of the ACM Web Conference 2022*, pp. 2310–2319, 2022.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pp. 136–145. PMLR, 2017.

Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. 2019.

Yoshua Bengio. Using a financial training criterion rather than a prediction criterion. *International journal of neural systems*, 8(04):433–443, 1997.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift modeling with generalization guarantees. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 55–65, 2021.

Sílvia Casacuberta and Moritz Hardt. Good allocations from bad estimates. *arXiv preprint arXiv:2601.05597*, 2026.

Emanuele Cavenaghi, Alessio Zanga, Fabio Stella, and Markus Zanker. Towards a causal decision-making framework for recommender systems. *ACM Transactions on Recommender Systems*, 2 (2):1–34, 2024.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022.

Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4888–4904, 2020.

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

Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same… and why it matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022.

Carlos Fernández-Loría, Foster Provost, Jesse Anderton, Benjamin Carterette, and Praveen Chandar. A comparison of methods for treatment assignment with an application to playlist generation. *Information Systems Research*, 34(2):786–803, 2023.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Unai Fischer-Abaigar, Christoph Kern, and Frauke Kreuter. The missing link: Allocation performance in causal machine learning. *arXiv preprint arXiv:2407.10779*, 2024.

Lin Ge, Hengrui Cai, Runzhe Wan, Yang Xu, and Rui Song. A review of causal decision making. *arXiv preprint arXiv:2502.16156*, 2025.

Yuhang Guo, Zicheng Su, Hai Yang, Enming Liang, Chen Zhong, and Wanjing Ma. A smart predict-then-optimize framework for vehicle rebalancing problem. *Transportation Research Part B: Methodological*, 206:103411, 2026.

Kellerer Hans, Pferschy Ulrich, and Pisinger David. Knapsack problems, 2004.

Bowei He, Yunpeng Weng, Xing Tang, Ziqiang Cui, Zexu Sun, Liang Chen, Xiuqiang He, and Chen Ma. Rankability-enhanced revenue uplift modeling framework for online marketing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5093–5104, 2024.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.

Fahad Kamran, Maggie Makar, and Jenna Wiens. Learning to rank for optimal treatment allocation under resource constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 3727–3735. PMLR, 2024.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Gar Goei Loke, Qinshen Tang, and Yangge Xiao. Decision-driven regularization: A blended model for predict-then-optimize. *Available at SSRN 3623006*, 2022.

Jayanta Mandi, Vıctor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In *International conference on machine learning*, pp. 14935–14947. PMLR, 2022.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pp. 335–422. Institute of Mathematical Statistics, 2008.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Toon Vanderschueren, Wouter Verbeke, Felipe Moraes, and Hugo Manuel Proença. Metalearners for ranking treatment effects. *arXiv preprint arXiv:2405.02183*, 2024.

Stefan Wager. Causal inference: A statistical learning approach, 2024.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Christopher Winship and Stephen L Morgan. The estimation of causal effects from observational data. *Annual review of sociology*, 25(1):659–706, 1999.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51, 2025.

Jiaqi Yang, Lexiao Chen, Zicheng Su, Wanjing Ma, Zhichao Zou, and Kun An. Decision-focused learning for optimal subsidy allocation in ride-hailing services. *Transportation Research Part C: Emerging Technologies*, 180:105301, 2025a.

Jiaqi Yang, Enming Liang, Zicheng Su, Zhichao Zou, Peng Zhen, Jiecheng Guo, Wanjing Ma, and Kun An. Dff: Decision-focused fine-tuning for smarter predict-then-optimize with limited data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26868–26876, 2025b.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

Shuli Zhang, Hao Zhou, Jiaqi Zheng, Guibin Jiang, Bing Cheng, Wei Lin, and Guihai Chen. Bi-level decision-focused causal learning for large-scale marketing optimization: Bridging observational and experimental data. *arXiv preprint arXiv:2510.19517*, 2025.

Yan Zhao, Xiao Fang, and David Simchi-Levi. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 588–596. SIAM, 2017.

Hao Zhou, Shaoming Li, Guibin Jiang, Jiaqi Zheng, and Dong Wang. Direct heterogeneous causal learning for resource allocation problems in marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5446–5454, 2023.

Hao Zhou, Rongxiao Huang, Shaoming Li, Guibin Jiang, Jiaqi Zheng, Bing Cheng, and Wei Lin. Decision focused causal learning for direct counterfactual marketing optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6368–6379, 2024.

A Problem Analysis

A.1 Assumptions

Throughout this paper, we adopt the standard assumptions as follows:

Assumption 1 (SUTVA)**.** *The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

Assumption 2 (Consistency)**.** *The potential outcome of treatment t equals to the observed outcome if the actual treatment received is t.*

Assumption 3 (Ignorability)**.** *Given pretreatment covariates x, the outcome variables $y(0)$ and $y(1)$ is independent of treatment assignment, i.e., $y(0), y(1) \perp\!\!\!\perp t|x$.*

Assumption 4 (Positivity)**.** *For any set of covariates $\boldsymbol{x}$, the probability to receive treatment 0 or 1 is positive, i.e., $0 < P(t_i = 1|x_i = x) < 1$ for all $x \in x_i$.*

A.2 QUALITATIVE INSIGHTS

A.2.1 NECESSARY AND SUFFICIENT CONDITIONS FOR ACHIEVING THE OPTIMAL DECISION

Theorem 1 (Tolerance to prediction errors)**.** *Let $\boldsymbol{t}^* = \boldsymbol{t}^*(\boldsymbol{\tau})$ be the optimal solution*[1] *to the problem. If the predicted vector $\hat{\boldsymbol{\tau}}$ satisfies*

$$\hat{\boldsymbol{\tau}}^\top \boldsymbol{t}^* \geq \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}, \qquad \forall \boldsymbol{t} \in S,$$

then the surrogate problem $\max_{\boldsymbol{t}\in S} \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}$ returns $\boldsymbol{t}^*$. In particular,

1. *(Magnitude) if $\hat{\boldsymbol{\tau}} = \alpha\boldsymbol{\tau}^*$ with $\alpha > 0$, the optimizer is unchanged;*
2. *(Direction) any $\hat{\boldsymbol{\tau}}$ lying in the normal cone*

$$N_{\mathrm{conv}(S)}(\boldsymbol{t}^*) = \{\boldsymbol{\tau} : \ \boldsymbol{\tau}^\top(\boldsymbol{t}^* - \boldsymbol{t}) \geq 0, \ \forall \boldsymbol{t} \in \mathrm{conv}(S)\}$$

preserves optimality; for a 0–1 feasible set $\mathrm{conv}(S)$ is a polytope and this cone has non-trivial angular width.

Proof. First, the inequality $\hat{\boldsymbol{\tau}}^\top \boldsymbol{t}^* \geq \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}$ for all $\boldsymbol{t} \in S$ implies that no feasible point attains a strictly larger objective value under $\hat{\boldsymbol{\tau}}$, hence t^* is an optimal solution of the surrogate problem.

If $\hat{\boldsymbol{\tau}} = \alpha\boldsymbol{\tau}$ with $\alpha > 0$, then $\hat{\boldsymbol{\tau}}^\top \boldsymbol{t} = \alpha\boldsymbol{\tau}^\top \boldsymbol{t}$, so the ordering of objective values is preserved and the optimizer remains t^*.

To analyze directional tolerance, replace S by $\mathrm{conv}(S)$; linear objectives attain maxima at extreme points, hence this substitution does not affect the maximizer. For binary decision variables $\mathrm{conv}(S)$ is a polytope, and the set of vectors that keep $\boldsymbol{t}^*$ optimal is the normal cone:

$$N_{\mathrm{conv}(S)}(t^*) = \{\boldsymbol{\tau} : \ \boldsymbol{\tau}^\top(\boldsymbol{t}^* - \boldsymbol{t}) \geq 0, \ \forall t \in \mathrm{conv}(S)\}.$$

For a polytope vertex this cone is polyhedral and typically has nonzero solid angle; therefore any $\hat{\boldsymbol{\tau}}$ whose direction lies in this cone preserves optimality. This completes the proof.

∎ □

Theorem 2 (Necessary and sufficient condition for optimality)**.** *Let $\mathrm{ext}(S)$ denote the extreme points of S. Then*

$$\boldsymbol{t}^*(\hat{\boldsymbol{\tau}}) = \boldsymbol{t}^*$$

if and only if

$$\hat{\boldsymbol{\tau}}^\top \boldsymbol{t}^* \geq \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}, \quad \forall \boldsymbol{t} \in \mathrm{ext}(S).$$

Proof. A linear function attains its maximum over a convex set at its extreme points. Because S consists of binary vectors, $\mathrm{conv}(S)$ is a polytope with finitely many extreme points. Therefore,

$$\max_{\boldsymbol{t}\in S} \hat{\boldsymbol{\tau}}^\top \boldsymbol{t} = \max_{\boldsymbol{t}\in \mathrm{ext}(S)} \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}.$$

If $\hat{\boldsymbol{\tau}}^\top \boldsymbol{t}^* \geq \hat{\boldsymbol{\tau}}^\top \boldsymbol{t}$ holds for all extreme points, then $\boldsymbol{t}^*$ is the optimal solution of the surrogate problem. Conversely, if the inequality fails for some extreme point, that point produces a strictly better value under $\hat{\boldsymbol{\tau}}$, contradicting optimality.

∎ □

[1] Without loss of generality, we assume that the optimization problem has a unique optimal solution.

A.2.2 A TOY CASE FOR COMPARISON

In this section, we consider a simple treatment assignment problem to illustrate that the conditions for optimal decisions can be further simplified in many scenarios, which significantly reduces the learning difficulty of ϕ_θ. Suppose we need to provide treatment for the top 5 out of 10 individuals. This constitutes a classical Top-k problem where $k = 5$, and ψ can be formulated as the following optimization problem:

$$\max_{\boldsymbol{t}} \sum_{i=1}^{10} \hat{\tau}_i t_i \tag{12}$$

$$\text{s.t.} \sum_{i=1}^{10} t_i = 5, \tag{13}$$

$$t_i \in 0, 1, \forall i \in 1, \ldots, 10, \tag{14}$$

where $\hat{\tau}_i = \phi_\theta(x_i)$ denotes the prediction of the CATE for individual i.

For this problem, we assume that the optimal solution is $\boldsymbol{t}^* = [t_1^*, \ldots, t_{10}^*]$, where the set $\mathcal{P}$ denotes the indices of all individuals receiving treatment (i.e., $t_p^* = 1$ for $p \in \mathcal{P}$), and the set $\mathcal{Q}$ denotes the indices of all individuals not receiving treatment (i.e., $t_q^* = 0$ for $q \in \mathcal{Q}$). Then, regardless of the true CATE of each individual i, the optimality condition for this problem can be expressed as:

$$\hat{\tau}_p \geq \hat{\tau}_q, \quad \forall p \in \mathcal{P}, q \in \mathcal{Q} \tag{15}$$

This is a necessary and sufficient condition for achieving the optimal decision. First, it is a necessary condition for achieving the optimal decision. When for any two individuals $p \in \mathcal{P}$ and $q \in \mathcal{Q}$, we have $\hat{\tau}_p \leq \hat{\tau}_q$, if we set $t_p^* = 0$ and $t_q^* = 1$ while keeping other elements of $\boldsymbol{t}^*$ unchanged, the objective function value of the optimization problem will increase, indicating that $\boldsymbol{t}^*$ is not the optimal solution. Furthermore, this is a sufficient condition for achieving the optimal decision. For each $p \in \mathcal{P}$, $\hat{\tau}_p \geq \hat{\tau}_q$ holds for any $q \in \mathcal{Q}$, which means that the ranking of p (i.e., rank(p)) among the 10 individuals satisfies rank$(p) \leq k$. Therefore, any $p \in \mathcal{P}$ satisfies rank$(p) \leq k$, and since the set $\mathcal{P}$ exactly contains k elements, the optimal solution $\boldsymbol{t}^*$ is necessarily obtained.

In this context, we find that the key to achieving optimal decision-making lies solely in separating individuals in set $\mathcal{P}$ from those in set $\mathcal{Q}$ via ϕ_θ. This is equivalent to transforming a regression problem into a binary classification problem, where we only need to learn which individuals belong to set $\mathcal{P}$ and which belong to set $\mathcal{Q}$. In contrast, we neither need to accurately predict the CATE for each individual nor determine the precise ranking of each individual. The intuition behind this learning approach is straightforward: under the same data conditions, learning a binary classification problem is likely less difficult than learning a regression problem.

Another critical perspective is that conventional methods employ point-wise learning to fit the value of CATE, whereas learning optimal decisions requires list-wise learning. Since decision-relevant data varies, the same individual may receive treatment in some scenarios but not in others, depending on whether they fall within the Top-k. Consequently, the same individual cannot have a fixed learning label (treatment or not), but instead requires a dynamic label derived from comparisons with other samples. From this perspective, our problem resembles classical contrastive learning, which aims to separate positive and negative samples as much as possible. For example, in our problem, elements in set $\mathcal{P}$ can be treated as positive samples, and elements in set $\mathcal{Q}$ as negative samples. The key is not to estimate elements in $\mathcal{P}$ to be arbitrarily large or those in $\mathcal{Q}$ to be arbitrarily small; rather, the goal is to ensure that elements in $\mathcal{P}$ are always greater than those in $\mathcal{Q}$.

Through this simple case, we anticipate that methods for directly learning optimal decisions can enhance the decision-making performance of this pipeline. However, the second key challenge is that we do not know the optimal decision, which is particularly difficult for observational data. In the next section, we propose a solution that leverages decision-relevant information during training to enhance decision-making performance.

B METHOD

B.1 STANDARD DOUBLY ROBUST LEARNER

First, we present the standard procedure of the DR-Learner.

In the first step, we need to fit an propensity score model and two outcome model for different treatment group:

$$\hat{e}(x) = E(T|X = x) \tag{16}$$

$$f_0(x) = E(Y|X = x, T = 0) \tag{17}$$

$$f_1(x) = E(Y|X = x, T = 1) \tag{18}$$

In the second step, based on the doubly robust formula (Robins et al., 1994), we can obtain unbiased estimates of τ for all training data (Kennedy, 2023):

$$\tilde{\tau}_i = \frac{t_i - \hat{e}(x_i)}{\hat{e}(x_i)(1 - \hat{e}(x_i))}(y_i - f_{t_i}(x_i)) + f_1(x_i) - f_0(x_i) \tag{19}$$

To estimate individual treatment effects τ from covariates x, we use the unbiased estimates from Equation 19 on the training data as pseudo-labels for fitting a regression model:

$$\hat{\tau} = \phi_\theta^{DR}(\tilde{\tau}_i|x) \tag{20}$$

For a new individual, we can use ϕ_θ^{DR} alone for estimating the treatment effect only based on covariates x.

B.2 DECISION LOSS

In this paper, we adopt two methods to compute the gradient: a surrogate loss method (SPO+) and a perturbation smoothing method (PFY)[2], as follows.

Smart Predict-then-Optimize Loss (SPO+). The SPO+ loss constructs a differentiable convex upper bound on the regret (Elmachtoub & Grigas, 2022):

$$\mathcal{L}_{\text{SPO+}}(\hat{\boldsymbol{\tau}}, \tilde{\boldsymbol{\tau}}) = \min_{\boldsymbol{t}\in\mathcal{S}}\{(2\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\tau}})^\top \boldsymbol{t}\} + 2\hat{\boldsymbol{\tau}}^\top \boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}) - \tilde{\boldsymbol{\tau}}^\top \boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}) \tag{21}$$

A useful subgradient of the SPO+ loss is given as follows:

$$2(\boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}) - \boldsymbol{t}^*(2\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\tau}})) \in \frac{\partial \mathcal{L}_{\text{SPO+}}(\hat{\boldsymbol{\tau}}, \tilde{\boldsymbol{\tau}})}{\partial \hat{\boldsymbol{\tau}}} \tag{22}$$

Perturbed Fenchel-Young Loss (PFY). In PFY loss, the predictions are perturbed with Gaussian noise, and the expected function of the perturbed minimizer is (Berthet et al., 2020):

$$F(\hat{\boldsymbol{\tau}}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\min_{\boldsymbol{t}\in\mathcal{S}}\{(\hat{\boldsymbol{\tau}} + \sigma\boldsymbol{\epsilon})^\top \boldsymbol{t}\}\right] \tag{23}$$

With $\Omega(\boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}))$, the Fenchel-Young dual of $F(\tilde{\boldsymbol{\tau}})$, the PFY loss is defined as:

$$\mathcal{L}_{\text{PFY}}(\hat{\boldsymbol{\tau}}, \tilde{\boldsymbol{\tau}}) = \hat{\boldsymbol{\tau}}^T \boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}) - F(\hat{\boldsymbol{\tau}}) - \Omega(\boldsymbol{t^*(\tau)}) \tag{24}$$

Although we cannot actually compute $\Omega(\boldsymbol{t^*(\tau)})$, it does not depend on the predicted values $\hat{\boldsymbol{\tau}}$. Thus, the gradient is:

$$\frac{\partial \mathcal{L}_{\text{PFY}}(\hat{\boldsymbol{\tau}}, \tilde{\boldsymbol{\tau}})}{\partial \hat{\boldsymbol{\tau}}} = \boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}) - \mathbb{E}_{\boldsymbol{\epsilon}}\left[\arg\min_{\boldsymbol{t}\in\mathcal{S}}\{(\hat{\boldsymbol{\tau}} + \sigma\boldsymbol{\epsilon})^T \boldsymbol{t}\}\right] \tag{25}$$

[2]We present the general forms of the loss functions for minimization problems, which can be adapted to our maximization problem by simple negation.

B.3 DECISION-FOCUSED LEARNING VIA PSEUDO LABEL

The above decision functions are incorporated into a standard machine-learning training pipeline, as illustrated in algorithm 2.

Algorithm 2 Decision-Focused Learning via Pseudo Labels

Input: Dataset $\mathcal{D}$; estimator ψ_θ; decision loss $\mathcal{L}$
Output: Learned parameters θ

Initialize parameters θ for estimator ψ_θ
Solve the optimal solution via pseudo labels: $\boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}) \leftarrow \tilde{\boldsymbol{\tau}}$
for *each training epoch* **do**
  for *each mini-batch* $(\boldsymbol{x}, \tilde{\boldsymbol{\tau}})$ **do**
    Predict treatment effects: $\hat{\boldsymbol{\tau}} \leftarrow \psi_\theta(\boldsymbol{x})$
    Solve the optimal solution $\boldsymbol{t}^*(\hat{\boldsymbol{\tau}}) \leftarrow \hat{\boldsymbol{\tau}}$
    Compute decision loss $\mathcal{L}_{\text{DFL}} \leftarrow \tilde{\boldsymbol{\tau}}, \boldsymbol{t}^*(\tilde{\boldsymbol{\tau}}), \boldsymbol{t}^*(\hat{\boldsymbol{\tau}})$
    Backpropagation to update parameters

$$\theta \leftarrow \theta - \alpha \cdot \frac{\partial \mathcal{L}_{\text{DFL}}}{\partial \hat{\boldsymbol{\tau}}} \cdot \frac{\partial \hat{\boldsymbol{\tau}}}{\partial \boldsymbol{\theta}}$$

return θ

C CASE DETAILS

C.0.1 TREATMENT ASSIGNMENT PROBLEM

Following the classification of common optimization problems in Fischer-Abaigar et al. (2024) and the combinational optimization problems discussed in Hans et al. (2004), we select four distinct decision-making scenarios such that:

Top-K Allocation (TOP-K) In treatment assignment problems, a common policy is to allocate treatment to the k individuals with the largest positive estimated CATE values:

$$\text{(Top-k)} \quad \max_{\boldsymbol{t} \in \mathcal{S}} \quad \sum_{i \in \mathcal{I}} \tau_i(x_i) t_i \tag{26}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} t_i \leq k \tag{27}$$

Cost Efficient Allocation (CE) In practice, treatment costs may vary across individuals. Accordingly, each treatment is associated with an individual-specific cost $c_i \in \mathbb{R}$, and the policy-maker seeks to maximize the aggregated benefit subject to a budget constraint B. This yields the following integer programming formulation for the optimal policy:

$$\text{(CE)} \quad \max_{\boldsymbol{t} \in \mathcal{S}} \quad \sum_{i \in \mathcal{I}} \tau_i(x_i) t_i \tag{28}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} c_i t_i \leq B \tag{29}$$

Precedence Constraint Knapsack Problem (PCKP). The PCKP, also known as the partially ordered knapsack problem, extends the classical knapsack problem by incorporating precedence relations among items. Specifically, a precedence relation "item m precedes item n" requires that item n can be selected only if item m is also selected. For a given instance of PCKP, the precedence relations are represented by a directed acyclic graph $G_{\mathcal{I}} = (V_{\mathcal{I}}, A_{\mathcal{I}})$, where the vertex set $V_{\mathcal{I}}$ corresponds to the set of items $\mathcal{I}$. An arc $(m, n) \in A_{\mathcal{I}}$ indicates that item m must be selected before

item n. Formally, PCKP is obtained by augmenting the CE problem with the precedence constraint in Equation 32 as follows:

$$\begin{aligned}
\text{(PCKP)} \quad \max_{t \in S} \quad & \sum_{i \in \mathbf{I}} \tau_i(x_i) t_i && (30)\\
\text{s.t.} \quad & \sum_{i \in \mathcal{I}} c_i t_i \leq B, && (31)\\
& t_m \geq t_n, \quad (m,n) \in A_{\mathcal{I}} && (32)\\
& t_i \in \{0,1\}, \quad \forall i \in \mathcal{I} && (33)
\end{aligned}$$

The Collapsing Knapsack Problem (CKP). The collapsing knapsack problem (CKP) is a variant of the classical knapsack problem in which the effective capacity of the knapsack decreases as items are selected (denotes as $g(\cdot)$). This collapsing effect models interdependencies among items, whereby the inclusion of certain items reduces the remaining capacity available for subsequent selections:

$$\begin{aligned}
\text{(CKP)} \quad \max_{t \in S} \quad & \sum_{i \in \mathbf{I}} \tau_i(x_i) t_i && (34)\\
\text{s.t.} \quad & \sum_{i \in \mathcal{I}} c_i t_i \leq g(\sum_{i \in \mathcal{I}} t_i), && (35)\\
& t_i \in \{0,1\}, \quad \forall i \in \mathcal{I} && (36)
\end{aligned}$$

C.0.2 DATASETS

We follow the data generation methods of Kamran et al. (2024) and Athey & Wager (2021) to simulate two sets of observational data. In Dataset 1, the τ values for each sample have significant variation. Dataset 2 involves more complex nonlinear functional relationships. The datasets is shown as follows:

Dataset 1

$$\begin{aligned}
\mathbf{x}_i &\sim \mathcal{N}(0, \mathbf{I}_{10\times 10}),\\
t_i \mid \mathbf{x}_i &\sim \text{Bern}\left(\frac{1}{1+e^{-x_{i,3}}}\right),\\
\epsilon_i \mid \mathbf{x}_i, t_i &\sim \mathcal{N}(0,1),\\
\tau_i \mid \mathbf{x}_i &= \frac{\max(x_{i,1},0) + \max(x_{i,2},0) + x_{i,4}^2 + |x_{i,6}|^3}{2},\\
y_i \mid \mathbf{x}_i, \tau_i, \epsilon_i, t_i &= t_i \tau_i + \epsilon_i\\
&+ \max(0, x_{i,3} + x_{i,4}) + |x_{i,5}| + x_{i,6} x_{i,7}
\end{aligned} \tag{37}$$

Dataset 2

$$\begin{aligned}
\mathbf{x}_i &\sim \mathcal{N}(0, \mathbf{I}_{10\times 10}),\\
t_i \mid \mathbf{x}_i &\sim \text{Bern}\left(\frac{1}{1+e^{-x_{i,3}}}\right),\\
\epsilon_i \mid \mathbf{x}_i, t_i &\sim \mathcal{N}(0,1),\\
\tau \mid \mathbf{x}_i &= 1 + 2\,|x_{i,4}| + x_{i,10}^2,\\
y_i \mid \mathbf{x}_i, \tau_i, \epsilon_i, t_i &= t_i \tau_i + \epsilon_i\\
&+5\,(2 + 0.5\sin(\pi x_{i,1}) - 0.5 x_{i,2} + 0.75 x_{i,3} x_{i,9})
\end{aligned} \tag{38}$$

C.0.3 SET UP

For the decision problem with $I = 20$, we gathered a total dataset of 3000 samples, split into a training set:validation set:test set ratio of 2:1:3. Due to the requirement of computing decision regret every N samples, we opted for a larger proportion of the test set to ensure more accurate final evaluation metrics. To construct the pseudo-labels for DR-Learner, we used logistic regression to estimate the propensity score $e(x)$ and CatBoost to estimate the two regressors $f_1(x)$ and $f_0(x)$. In the final regression stage, we employed a three-layer neural network with 32 hidden units and a learning rate of 5×10^{-4}, trained for up to 500 epochs with early stopping based on the minimum validation loss. Additionally, shuffle-based data augmentation was applied to increase scenario diversity.

C.0.4 REAL-WORLD SUBSIDY ALLOCATION PROBLEM

For a specific ride-hailing category across multiple cities in China, we have a certain budget B each day. We consider two subsidy schemes among the population in each city $i \in \mathcal{I}$: basic subsidy ($t_i = 0$) and enhanced subsidy ($t_i = 1$), where the enhanced subsidy increases the subsidy rate by q on top of the basic subsidy rate (relative to the city's total Gross Merchandise Volume (GMV) G_i) . Our goal is to maximize the transaction value growth for the company using the additional subsidy. The subsdiy allocation model is as follows:

$$\text{(SA)} \quad \max_{t \in S} \quad \sum_{i \in \mathcal{I}} \tau_i(x_i) t_i \tag{39a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} c_i t_i \leq B \tag{39b}$$

$$c_i = gmv_i * q, \quad \forall i \in \mathcal{I} \tag{39c}$$

$$t_i \in \{0, 1\}, \quad \forall i \in \mathcal{I} \tag{39d}$$

Here, τ represents the additional GMV generated by using the additional subsidy compared to the basic subsidy in that city. Although the ground truth of τ is unavailable, it can be approximately estimated through randomized experiments within the city.

We selected data from 62 cities in China from March 1, 2024, to May 31, 2024. These cities conducted intra-city experiments daily, giving each city a value of τ each day. We performed 10 experiments, randomly selecting 60 days as the training set and 32 days of data as the test set for each experiment.