Contextual Distributionally Robust Optimization with Causal and Continuous Structure: An Interpretable and Tractable Approach

Fenglin Zhang, Jie Wang*

January 15, 2026

Abstract

In this paper, we introduce a framework for contextual distributionally robust optimization (DRO) that considers the causal and continuous structure of the underlying distribution by developing interpretable and tractable decision rules that prescribe decisions using covariates. We first introduce the causal Sinkhorn discrepancy (CSD), an entropy-regularized causal Wasserstein distance that encourages continuous transport plans while preserving the causal consistency. We then formulate a contextual DRO model with a CSD-based ambiguity set, termed Causal Sinkhorn DRO (Causal-SDRO), and derive its strong dual reformulation where the worst-case distribution is characterized as a mixture of Gibbs distributions. To solve the corresponding infinite-dimensional policy optimization, we propose the Soft Regression Forest (SRF) decision rule, which approximates optimal policies within arbitrary measurable function spaces. The SRF preserves the interpretability of classical decision trees while being fully parametric, differentiable, and Lipschitz smooth, enabling intrinsic interpretation from both global and local perspectives. To solve the Causal-SDRO with parametric decision rules, we develop an efficient stochastic compositional gradient algorithm that converges to an ε-stationary point at a rate of $\mathcal{O}(\varepsilon^{-4})$, matching the convergence rate of standard stochastic gradient descent. Finally, we validate our method through numerical experiments on synthetic and real-world datasets, demonstrating its superior performance and interpretability.

Keywords: Contextual distributionally robust optimization, Causal Sinkhorn discrepancy, Soft regression forest, Stochastic compositional optimization

1 Introduction

Contextual stochastic optimization (CSO) is a widely applied method in real-world engineering and business decision-making (Sadana et al., 2025). It assumes that decision-makers have access to historical data, including uncertain parameters and associated contextual information (called

*Jie Wang is affiliated with School of Artificial Intelligence and School of Data Science at The Chinese University of Hong Kong, Shenzhen; Fenglin Zhang is affiliated with School of Artificial Intelligence at The Chinese University of Hong Kong, Shenzhen; Email: zhangfl819@outlook.com, jwang@cuhk.edu.cn; Corresponding author: Jie Wang.

covariates or features, Chenreddy et al. (2022)). Such contextual information enables a more precise characterization of uncertain parameters, leading to better decisions (Ban and Rudin, 2019; Bertsimas and Kallus, 2020). The goal of CSO is to seek an optimal policy that maps covariates to decisions, thereby avoiding the need to solve optimization models repeatedly for every new covariate (Sadana et al., 2025).

In practice, the CSO model may suffer from misspecification. This issue arises from statistical errors due to limited sample sizes and distributional shifts between training and testing environments, resulting in suboptimal out-of-sample performance and even fragility (Bennouna and Van Parys, 2022; F. Liu et al., 2024). To hedge against this uncertainty, the contextual distributionally robust optimization (DRO) method has received much attention in recent literature (R. Chen and Paschalidis, 2019; Esteban-Pérez and Morales, 2022; Kannan et al., 2024; Nguyen et al., 2025; Sim et al., 2025; Srivastava et al., 2021; T. Wang et al., 2021). This approach seeks the optimal robust decision that minimizes the worst-case risk over an *ambiguity set* containing all plausible joint distributions of covariates and uncertain parameters. Unlike traditional DRO, which takes into account the ambiguity of uncertain parameters but ignores that of covariates, contextual DRO considers both to avoid suboptimal, overly conservative, or even infeasible solutions (Ban and Rudin, 2019; Zhu et al., 2022).

For contextual DRO, extensive literature constructs ambiguity sets based on optimal transport (Kannan et al., 2024; Nguyen et al., 2020, 2025; T. Wang et al., 2021; J. Yang et al., 2022). A critical yet often overlooked aspect in contextual DRO is the causal information structure: future covariates are conditionally independent of historical uncertain parameters given the history. For instance, in a newsvendor setting, while daily temperature (covariate) influences demand (uncertain parameter), if the historical temperature is known, the historical demand and the future temperature are conditionally independent, as they cannot affect each other. To characterize this structure, J. Yang et al. (2022) propose a contextual DRO model with an ambiguity set constructed using causal Wasserstein distance that takes into account this causal relation. This model hedges against a discrete worst-case distribution that remains causally consistent, thereby avoiding causally implausible robustness scenarios.

Another critical observation is that the underlying distribution of CSO is typically continuous in practical applications (e.g., continuous temperature and demand distributions), and a discrete worst-case distribution from the aforementioned DRO framework (J. Yang et al., 2022) may lead to overly conservative decisions. Therefore, it remains an open question to *develop a contextual DRO model that simultaneously captures the causal structure and absolute continuity of the worst-case distribution*. J. Wang et al. (2025) recently develop a new DRO framework based on the Sinkhorn discrepancy to characterize the continuity of underlying distributions. This framework, referred to as Sinkhorn DRO, incorporates entropic regularization into the Wasserstein distance, thereby excluding all discrete distributions in the ambiguity set. The variants of Sinkhorn DRO have also been explored in literature (Azizian et al., 2023a, 2023b; Birrell and Ebrahimi, 2025; Blanchet et al., 2023) and has wide applications in hypothesis testing (J. Wang and Xie, 2022; J. Wang et al., 2024; S.-B. Yang and Li, 2023), experimental design (Dapogny et al., 2023; Jiang and Mao, 2025), machine learning (Cescon et al., 2025; Ouasfi et al., 2025; Shen et al., 2023; Song et al., 2024), etc.

While the Sinkhorn DRO is capable of providing continuous worst-case distributions, the infinite-dimensional nature of policy optimization imposes a computational challenge. To address this, recent literature has developed both parametric and non-parametric decision rule approaches to seek effective approximate policies. For parametric rules, Ban and Rudin (2019) consider affine decision rules for the CSO. Although computationally efficient, it may lead to suboptimal decisions because it often fails to capture complex and nonlinear relations between covariates and decisions. Later Bertsimas and Koduri (2022), Qi et al. (2023), Han et al. (2025), and Z. Liu et al. (2025) propose kernel-based and deep-learning-based decision rules, respectively, to approximate the complicated function space. These methods have superior empirical performance, but are often difficult to interpret. Alternatively, non-parametric rules developed in Zhang et al. (2024) and Nguyen et al. (2025) offer better interpretability, but they are computationally inefficient as the sample sizes increase and are only applicable to special contextual DRO problems. Inspired by literature, we aim to answer the following question:

> *How to develop a decision rule approach with interpretability and computational tractability for solving general contextual DRO problems?*

The tree-based family, including decision tree and random forest, has been developed for general CSO models for estimating conditional distributions of uncertain parameters to improve the interpretability (Bertsimas and Kallus, 2020; Elmachtoub et al., 2020; Kallus and Mao, 2023). However, few studies employ tree-based models for decision rule optimization. This is primarily because constructing an optimal tree is NP-Hard, and its non-differentiable structure precludes efficient end-to-end policy optimization (Aghaei et al., 2025; Notz and Pibernik, 2024). It is desirable to explore a computationally tractable tree-based model for decision rule optimization that preserves interpretability.

In this paper, we develop a new contextual DRO model with an ambiguity set based on causal and entropy-regularized Wasserstein distance, which retains the causal and continuous structure of the underlying distribution, referred to as Causal Sinkhorn DRO (Causal-SDRO). To efficiently approximate optimal policies, we propose a parametric decision rule based on the Soft Regression Forest (SRF), which ensembles several differentiable soft decision trees to prescribe end-to-end and interpretable decisions. Our main contributions are summarized as follows.

(I) To model the ambiguity set of Causal-SDRO, we introduce the causal Sinkhorn discrepancy, which is a variant of Wasserstein distance that combines the causal property from (J. Yang et al., 2022) and the continuous property of transport plans from (J. Wang et al., 2025). We further derive the strong dual reformulation and the expression of the worst-case distribution for the inner problem of Causal-SDRO under general assumptions.

(II) We propose a Soft Regression Forest (SRF) decision rule, which approximates optimal policies within arbitrary measurable function spaces. The proposed SRF retains the intrinsic interpretability of traditional non-differentiable decision trees while offering a parametric, differentiable, and Lipschitz smooth decision rule. We demonstrate the interpretability of this decision rule, grounded in its structural transparency, stability, and robustness, and introduce intrinsic interpretation measures from both global and local perspectives.

(III) We reformulate Causal-SDRO with parametric decision rules as a multi-level stochastic compositional optimization. We first consider the sample average approximation (SAA) approach, which approximates the target problem as deterministic optimization. Its theoretical sample complexity is $\mathcal{O}(\delta^{-2})$ to control the approximation error within δ with high probability. However, due to the multi-level compositional structure of the problem, it is still computationally challenging to solve the SAA problem. Instead, we develop a stochastic compositional gradient algorithm that converges to an ε-stationary point at a rate of $\mathcal{O}(\varepsilon^{-4})$, which is at the same order as standard stochastic gradient descent (Ghadimi et al., 2016).

(IV) We validate the proposed approach through numerical experiments on three applications: a feature-based newsvendor problem, a feature-based inventory substitution problem (representing a two-stage contextual DRO setting), and a real-world data-driven portfolio selection problem. Throughout the experiments, our methods are both computationally efficient and exhibit interpretability compared with existing baselines.

The remainder of this paper is organized as follows. The next two subsections in this section review the related literature and introduce conventions and notations throughout this paper. Section 2 presents the necessary definitions and formulates the Causal-SDRO model. Section 3 derives the strong dual formulation and the worst-case distribution of the Causal-SDRO model given a fixed decision rule. Section 4 proposes the interpretable soft regression forest decision rule. Section 5 discusses several methods for solving the Causal-SDRO model with parametric decision rules. Section 6 reports the numerical results of our methods on three contextual DRO applications. Section 7 concludes the paper. Proofs and additional analyses are provided in the E-companion to this paper.

1.1 Related Literature

In this subsection, we review existing literature related to our work.

On data-driven prescriptive analytics. Our study is rooted in the data-driven decision-making paradigm, which leverages data to prescribe decisions in optimization problems under uncertain (Bertsimas and Kallus, 2020; Sim et al., 2025). This approach, leveraging rich covariates to improve decision-making with uncertain parameters, is called Contextual Stochastic Optimization (Sadana et al., 2025). To address this problem, existing studies have developed various frameworks. Bertsimas and Kallus (2020) propose a data-driven framework based on weighted sample average approximation. This method estimates the conditional distribution by generating data-driven weights via machine learning models (e.g., k-nearest neighbor method and decision trees) and prescribes decisions by minimizing the reweighted empirical cost. Building on this, Kallus and Mao (2023) propose a stochastic optimization forest method, which calculates weights by optimizing the downstream decision quality rather than prediction accuracy. Elmachtoub and Grigas (2022) develop a smart predict-then-optimize framework, which integrates learning and optimization by introducing a decision-oriented loss function. Qi et al. (2025) present an integrated conditional estimation-optimization framework based on the downstream objective. Dis-

tinct from the aforementioned approaches that still involve an intermediate estimation process, our work focuses on the decision rule approach, which directly maps covariates to final decisions (Liyanage and Shanthikumar, 2005). Notable examples include the linear and kernel-based decision rules for newsvendor problems by Ban and Rudin (2019) and the deep-learning-based rules for inventory management by Qi et al. (2023).

On contextual Distributionally Robust Optimization (DRO). To hedge against the distributional shift issue in CSO, contextual DRO has received much attention with various types of ambiguity sets. Early approaches focused on ϕ-divergence (Poursoltani et al., 2023; Srivastava et al., 2021) and moment-based (Perakis et al., 2023) ambiguity sets, primarily due to their computational tractability. In recent literature, the optimal transport-based (i.e., Wasserstein distance-based) ambiguity sets are widely used for contextual DRO modeling (Esteban-Pérez and Morales, 2022; Kannan et al., 2024; Nguyen et al., 2025; Qi et al., 2022; Zhang et al., 2024), due to their data-driven nature and satisfactory out-of-sample guarantees. Most relevant to our work is J. Yang et al. (2022), which provide an ambiguity set based on the causal transport distance to preserve the conditional independence structure between historical uncertain parameters and newly observed covariates. However, as a variant of the Wasserstein contextual DRO, this ambiguity set hedges against discrete worst-case distributions, which may lead to overly conservative contextual decisions, as the true distribution is often continuous in many applications. We bridge this gap by extending the causal transport distance to a causal and entropy-regularized discrepancy (J. Wang et al., 2025), ensuring that the resulting DRO model hedges against continuous distributions while preserving causal structure.

On decision rule approach for contextual DRO. Decision rule approaches seek an optimal policy within a pre-specified function class that makes end-to-end decisions based on covariates. Existing research has developed several parametric and non-parametric decision rule approaches for solving contextual DRO. J. Yang et al. (2022) reformulate the causal optimal transport-based DRO as a conic program for affine decision rules. However, it may not be tractable for more general parametric decision rules. Under such a case, Hu et al. (2023) reformulate this problem as a large-scale bilevel program, whereas it is still computationally challenging. For non-parametric decision rules, finite-dimensional convex reformulations of contextual DRO have been provided (Esteban-Pérez and Morales, 2022; Fu et al., 2024; Nguyen et al., 2025; J. Yang et al., 2022; Zhang et al., 2024) for special problem structures or with special choices of the Wasserstein distance and its variants. In this paper, we propose a Soft Regression Forest (SRF) decision rule for general contextual DRO, aiming to balance the trade-off between interpretability and computational tractability.

On trustworthy and interpretable decision-making. In machine learning, interpretability generally refers to the ability to explain or to present in understandable terms to humans (Bertsimas et al., 2019; Doshi-Velez and Kim, 2017). As mentioned above, kernel or deep-learning-based methods for contextual optimization perform well but lack trustworthiness and interpretability (Bertsimas and Koduri, 2022; Oroojlooyjadid et al., 2020). Therefore, in high-stakes applications, such as healthcare and finance, those methods may pose risks (Forel et al., 2023; Rudin, 2019). Instead of explaining deep learning models (Lundberg et al., 2020), extensive literature is dedi-

cated to applying or developing inherently trustworthy models for decision-making (Forel et al., 2023), such as decision tree and forest methods, which have been considered interpretable and used in some decision-making processes due to their explicit "if-then" logical structures (Aghaei et al., 2025; Bertsimas and Dunn, 2017; Bertsimas and Stellato, 2021; Rudin, 2019). These methods have also been applied in CSO (Bertsimas and Kallus, 2020; Elmachtoub et al., 2020; Kallus and Mao, 2023; Notz and Pibernik, 2024). Distinct from existing tree-based methods, where trees are constructed via greedy heuristics due to non-differentiability and NP-hardness, the proposed SRF decision rule is differentiable and can be efficiently trained by gradient-based methods, while maintaining the intrinsic interpretability.

1.2 Conventions and Notations

For integer $K \in \mathbb{Z}_+$, define $[K] \triangleq \{1, \cdots, K\}$. We denote random vectors by bold upper case letters (e.g., $\boldsymbol{X}, \boldsymbol{Y}$) and their realizations by bold lower case letters (e.g., $\boldsymbol{x}, \boldsymbol{y}$). For a measurable set $\mathcal{Y}$, denote $\mathcal{M}(\mathcal{Y})$ as the set of measures on $\mathcal{Y}$, and $\mathcal{P}(\mathcal{Y})$ as the set of probability measures on $\mathcal{Y}$. Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}}), (\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be subsets of normed vector spaces. For simplicity, the subscripts in $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ will be omitted if no ambiguity exists. Denote $\mathbb{P} \otimes \mathbb{Q}$ as the product measure of two probability measures $\mathbb{P}$ and $\mathbb{Q}$. Given a probability distribution $\mathbb{P}$ and a measure μ, we denote by $\operatorname{supp} \mathbb{P}$ the support of $\mathbb{P}$, and write $\mathbb{P} \ll \mu$ if $\mathbb{P}$ is absolutely continuous with respect to μ. Let the logarithm function log take with base e. A function $f : \mathcal{X} \to \mathbb{R}^n$ is said to be L-Lipschitz continuous, if there exists a constant $L > 0$ such that $\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_2 \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be S-Lipschitz smooth if it is continuously differentiable and its gradient is S-Lipschitz continuous. A function $f : \mathcal{X} \to \mathbb{R}^n (n > 1)$ is said to be S-Lipschitz smooth if it is continuously differentiable and its Jacobian $J_{\boldsymbol{f}}(\boldsymbol{x})$ is S-Lipschitz continuous. Let $\mathbb{V}_{\boldsymbol{\xi}}(t(\cdot; \boldsymbol{\xi}))$ denote the variance of the random variable (or random vector) $t(\cdot; \boldsymbol{\xi})$. Let $(\cdot)^+$ denote the element-wise positive part operator. That is, for any vector $\boldsymbol{x} = [x_1, \ldots, x_{d_x}]^\top \in \mathbb{R}^{d_x}$, we have $(\boldsymbol{x})^+ = [\max\{x_1, 0\}, \ldots, \max\{x_{d_x}, 0\}]^\top$. For a vector $\boldsymbol{w} \in \mathbb{R}^d$, we denote $[\boldsymbol{w}]_k$ as its k-th element for any $k \in [d]$. We use $|\cdot|$ to represent the cardinality of a set.

2 The Causal-SDRO Model

In this section, we introduce the Causal-SDRO model. Consider a CSO model where a decision rule $f : \mathcal{X} \to \mathcal{Z}$ maps a covariate vector $\boldsymbol{x}$ for a compact covariate space $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ to a decision $\boldsymbol{z} \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$. The goal is to $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ to minimize the expectation of the measurable loss function $\Psi : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ with respect to uncertain parameters $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Consequently, the CSO is formulated as

$$\inf_{f \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \widehat{\mathbb{P}}} \Big[\Psi(f(\boldsymbol{x}), \boldsymbol{y}) \Big],$$

where $\mathcal{F}$ is a space of measurable functions, and $\widehat{\mathbb{P}}$ represents the empirical joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$. Based on CSO, contextual DRO assumes the unknown true distribution $\mathbb{P}$ lies within an ambiguity set $\mathfrak{P}(\widehat{\mathbb{P}})$ constructed based on $\widehat{\mathbb{P}}$. The objective is to identify a robust decision rule that

minimizes the worst-case expected loss

$$\inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathfrak{P}(\widehat{\mathbb{P}})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\Big].$$

To construct our ambiguity set, we first recall the causal transport distance (J. Yang et al., 2022), which incorporates the causal structure into the optimal transport framework.

Definition 1. (Causal Transport Distance, J. Yang et al., 2022). *Let distributions $\mathbb{P},\mathbb{Q}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$. A joint distribution $\gamma\in\Gamma(\mathbb{P},\mathbb{Q})$ is termed a causal transport plan if, for $((\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}}),(\boldsymbol{X},\boldsymbol{Y}))\sim\gamma$, the random variable $\boldsymbol{X}$ is conditionally independent of $\widehat{\boldsymbol{Y}}$ given $\widehat{\boldsymbol{X}}$, denoted as*

$$\boldsymbol{X}\perp\widehat{\boldsymbol{Y}}\mid\widehat{\boldsymbol{X}}.$$

Let $\Gamma_c(\mathbb{P},\mathbb{Q})$ be the set of all causal transport plans within $\Gamma(\mathbb{P},\mathbb{Q})$. For $p\in[1,\infty)$, the p-causal transport distance between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$C_p(\mathbb{P},\mathbb{Q}):=\left(\inf_{\gamma\in\Gamma_c(\mathbb{P},\mathbb{Q})}\mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\Big[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\Big]\right)^{1/p},$$

where $c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))=\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|^p+\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^p$ is a transport cost function. ◇

The ambiguity set of causal transport distance-based DRO includes both discrete and continuous distributions, whereas it typically hedges against a discrete one, which may not be realistic in practice, as the true distribution is often continuous in many applications. In this paper, we introduce the *Causal Sinkhorn Discrepancy (CSD)* by incorporating entropy regularization, which excludes all discrete distributions in the ambiguity set. The CSD is defined as follows.

Definition 2. (Causal Sinkhorn Discrepancy, CSD). *Consider distributions $\mathbb{P},\mathbb{Q}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$, and let measures $\mu,\nu\in\mathcal{M}(\mathcal{X}\times\mathcal{Y}),\nu_{\mathcal{X}}\in\mathcal{M}(\mathcal{X}),\nu_{\mathcal{Y}}\in\mathcal{M}(\mathcal{Y})$ be reference measures such that $\mathbb{P}\ll\mu$ and $\mathbb{Q}\ll\nu$. For regularization parameter $\epsilon\geq 0$ and $p\in[1,\infty)$, the p-CSD between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

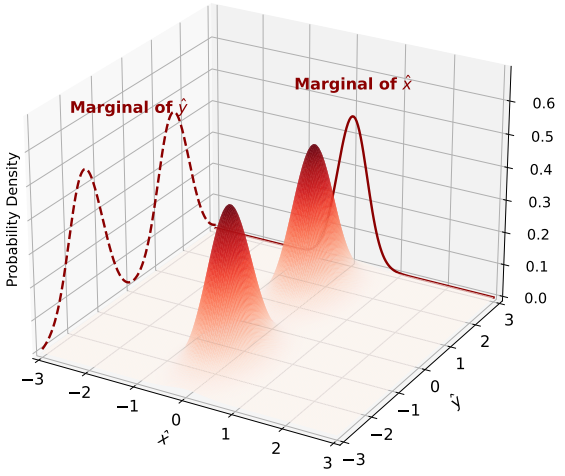
$$R_p(\mathbb{P},\mathbb{Q}):=\left(\inf_{\gamma\in\Gamma_c(\mathbb{P},\mathbb{Q})}\mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\Big[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\Big]+\epsilon\cdot H\Big(\gamma\mid\mu\otimes\big(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}\big)\Big)\right)^{1/p},$$

where $\Gamma_c(\mathbb{P},\mathbb{Q})$ is the causal transport plan set, and the relative entropy of γ with respect to the measure $\mu\otimes\big(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}\big)$ is given by

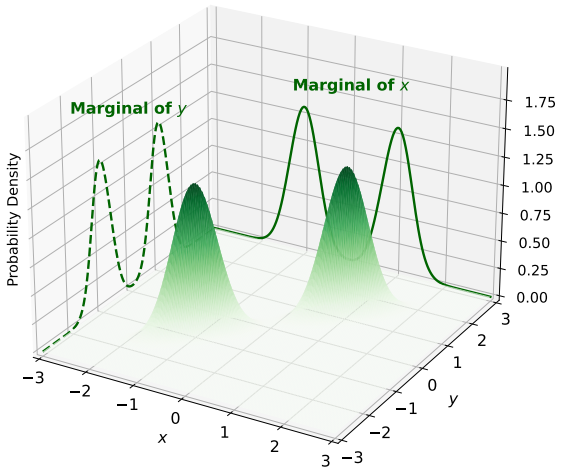
$$H\Big(\gamma\mid\mu\otimes\big(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}\big)\Big)=\mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\left[\log\left(\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right)\right],$$

where $\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}$ stands for the density ratio of γ with respect to $\mu\otimes\big(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}\big)$ evaluated at $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})$. ◇

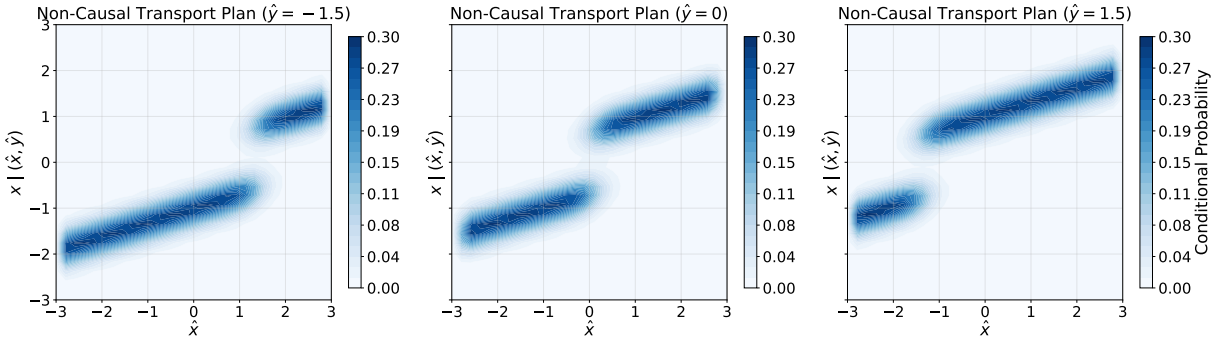
Unlike deterministic transport plans derived from the Wasserstein distance, Sinkhorn transport plans are probabilistic. Specifically, entropy regularization ($\epsilon>0$) penalizes deterministic



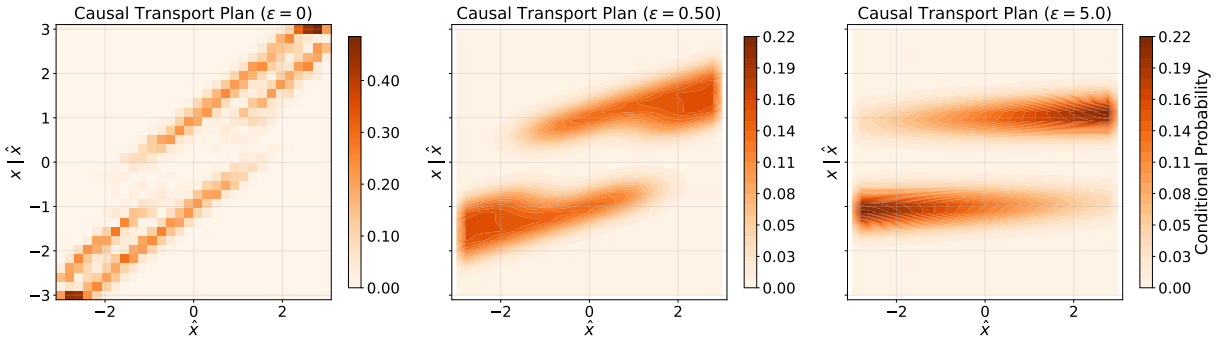
(a) Source Distribution $(\widehat{x}, \widehat{y})$



(b) Target Distribution (x, y)



(c) Non-causal Sinkhorn transport plans with $\widehat{y} \in \{-1.5, 0, 1.5\}$ when $\epsilon = 0.5$



(d) Causal Sinkhorn transport plans with $\epsilon \in \{0, 0.5, 5.0\}$

Figure 1. Visualization for causal and non-causal Sinkhorn transport plans.

transport plans, yielding smooth plans that map each source point to a probability distribution over the target space rather than a single point.

Based on CSD, we study the following contextual DRO model, where the outer minimization seeks optimal decision rules and the inner maximization seeks the worst-case distribution in the ambiguity set constructed by CSD around the empirical distribution $\widehat{\mathbb{P}}$:

$$\inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathfrak{P}(\widehat{\mathbb{P}})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\Big], \tag{Causal-SDRO}$$

where

$$\mathfrak{P}(\widehat{\mathbb{P}}) = \Big\{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y}) : R_p(\widehat{\mathbb{P}},\mathbb{P})^p \le \rho^p\Big\}.$$

This Causal Sinkhorn Distributionally Robust Optimization (Causal-SDRO) model hedges against a continuous worst-case distribution (see the discussion in Section 3.2) while preserving the causal information structure, and thereby avoiding overly conservative and causally inconsistent decisions.

Figure 1 visualizes and compares the causal and non-causal transport plans for illustrative source $(\widehat{x},\widehat{y})$ and target (x,y) distributions supported on $[-3,3]^2$. The source shown in Figure 1(a) comprises $\widehat{x}\sim\mathcal{N}(0,0.1)$ and $\widehat{y}$ from an equiprobable mixture of $\mathcal{N}(\pm1.5,0.3)$. The target shown in Figure 1(b) is a mixture of two bivariate Gaussians centered at $(-1,-1)$ and $(1,1)$ with positive correlation. Figure 1(c) illustrates non-causal Sinkhorn transport plans, where the mapping from $\widehat{x}$ to x explicitly depends on $\widehat{y}$. Specifically, when $\widehat{y}=-1.5$ (left panel), the source mode at $\widehat{x}=0$ is transported primarily to the target's bottom-left peak ($x\approx-1$) to minimize transport cost. Similarly, when $\widehat{y}=1.5$ (right panel), the mass is directed toward the closer upper-right peak ($x\approx1$). In contrast, their corresponding causal Sinkhorn transport plan (Figure 1(d), middle panel) enforces conditional independence, and consequently, it appears as an aggregate of the non-causal plans. Figure 1(d) further demonstrates the difference between the causal Wasserstein ($\epsilon=0$) and the causal Sinkhorn transport plans ($\epsilon=0.5$ and 5). As ϵ increases, the causal Sinkhorn transport plans converge towards the product of marginal distributions of $\widehat{x}$ and x.

In the following, we introduce several practical applications of (Causal-SDRO).

Example 1. (Feature-based Newsvendor Problem). Consider a newsvendor who sells d_z kinds of products. Let $\boldsymbol{h}\in\mathbb{R}^{d_z}$ and $\boldsymbol{b}\in\mathbb{R}^{d_z}$ to represent the holding and stock-out cost. The newsvendor loss function $\Psi_{\mathrm{N}}:\mathcal{Z}\times\mathcal{Y}\to\mathbb{R}$ for given uncertain demand $\boldsymbol{y}\in\mathcal{Y}\subseteq\mathbb{R}^{d_y}(d_y=d_z)$ is defined as

$$\Psi_{\mathrm{N}}(\boldsymbol{z},\boldsymbol{y}) := \boldsymbol{h}^\top\Big(\boldsymbol{z}-\boldsymbol{y}\Big)^+ + \boldsymbol{b}^\top\Big(\boldsymbol{y}-\boldsymbol{z}\Big)^+.$$

Consider features $\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^{d_x}$ (e.g., season and weather), the problem is given by

$$\inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathfrak{P}(\widehat{\mathbb{P}})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\boldsymbol{h}^\top\Big(f(\boldsymbol{x})-\boldsymbol{y}\Big)^+ + \boldsymbol{b}^\top\Big(\boldsymbol{y}-f(\boldsymbol{x})\Big)^+\Big].$$

This problem will be revisited in Sections 3.2 and 6.1. ♣

Example 2. (Feature-based Inventory Substitution Problem). This problem, as a variant of the supply chain substitution problem in X. Chen and Gao (2019), is a two-stage optimization

problem. Consider a firm selling d_z types of products to satisfy customer demands, the products are indexed by $i \in [d_z]$ where a lower index implies a higher quality. There is a demand class corresponding to each product, indexed by $j \in [d_y]$ ($d_z = d_y$). If any demand class j cannot be satisfied, products with higher quality $i < j$ can substitute for demand j at an extra cost $s_{i,j}$. Before the real demand is observed, the "wait-and-see" decision for the firm is to decide the prepared inventory level $\boldsymbol{z} \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ (suppose that the initial inventory level is zero) with cost $\boldsymbol{c} \in \mathbb{R}^{d_z}$. After knowing the demand $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, the "here-and-now" decision is to allocate the inventory to each demand class, targeting the lowest total cost. Let decision variable $w_{i,j} \geq 0$ denote the quality of product i that substitutes j, while $h_i \geq 0$ and $b_j \geq 0$ denote the unit holding cost of product i and shortage cost of demand j, respectively. Taking features $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ (e.g., demographic data) into account, this two-stage contextual DRO problem is given by

$$\inf_{f \in \mathcal{F}} \max_{\mathbb{P} \in \mathfrak{P}(\widehat{\mathbb{P}})} \quad \boldsymbol{c}^\top f(\boldsymbol{x}) + \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi_{\mathrm{I}}(f(\boldsymbol{x}), \boldsymbol{y})\Big]$$

where the inventory substitution loss function $\Psi_{\mathrm{I}} : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ is defined as

$$\begin{aligned}
\Psi_{\mathrm{I}}(\boldsymbol{z}, \boldsymbol{y}) := \min \quad & \sum_{j=1}^{d_y} \sum_{i=1}^{j} s_{i,j} w_{i,j} + \sum_{i=1}^{d_z} h_i u_i + \sum_{j=1}^{d_y} b_j u'_j & \\
\text{s.t.} \quad & \sum_{j=i}^{d_y} w_{i,j} + u_i = z_i, & \forall i \in [d_z], \\
& \sum_{i=1}^{j} w_{i,j} + u'_j = y_j, & \forall j \in \Big[d_y\Big], \\
& u_i, u'_j, w_{i,j} \geq 0, & \forall i \in [d_z], j \in \Big[d_y\Big],
\end{aligned}$$

where u_i represents the leftover inventory of product i, u'_j represents the shortage of demand j, and the three terms in $\Psi_{\mathrm{I}}(\boldsymbol{z}, \boldsymbol{y})$ represent the total purchasing cost, total holding cost, and total shortage cost, respectively. We conduct numerical experiments for this problem in Section 6.2 ♣

Example 3. (Data-driven Portfolio Selection Problem). Conditioned on a covariate $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ (e.g., macroeconomic indicators), the portfolio manager determines a portfolio allocation strategy across various assets that minimizes the worst-case conditional risk-return tradeoff (Nguyen et al., 2025). In this data-driven portfolio selection problem, the random vector $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denotes the assets' future return, and the risk can be described by variance, conditional value-at-risk, etc. We choose the variance of return as the measure of risk in this example, leading to the following model

$$\inf_{f \in \mathcal{F}} \max_{\mathbb{P} \in \mathfrak{P}(\widehat{\mathbb{P}})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi_{\mathrm{P}}(f(\boldsymbol{x}), \boldsymbol{y})\Big],$$

where the portfolio loss function $\Psi_{\mathrm{P}} : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ is defined as

$$\Psi_{\mathrm{P}}(\boldsymbol{z}, \boldsymbol{y}) := \left\{ -\omega \cdot \sum_{i=1}^{d_y} y_i z_i + \left(\sum_{i=1}^{d_y} y_i z_i - z_0 \right)^2 \quad \middle| \quad \begin{array}{l} \sum_{i=1}^{d_y} z_i = 1, \\ \min_{i\in[d_y]}\{y_i z_i\} \leq z_0 \leq \max_{i\in[d_y]}\{y_i z_i\}, \\ z_i \geq 0, \qquad \forall i \in [d_y], \end{array} \right\},$$

where the parameter ω balances the trade-off between the portfolio return (the first term) and the associated risk (the second term), and for decision variables $\boldsymbol{z} = \left(z_0, z_1, \cdots, z_{d_y}\right)^\top \in \mathcal{Z} \subseteq \mathbb{R}^{d_y+1}$, z_0 represents the expected portfolio return while z_i represents the portfolio on the asset i for each $i \in [d_y]$. We will solve this problem on real data in Section 6.3. ♣

3 Duality Reformulation for Causal-SDRO

In this section, we establish the strong duality and characterize the worst-case distribution for the inner maximization problem of (Causal-SDRO), assuming a fixed decision rule $f \in \mathcal{F}$. The primal problem is defined as

$$v_{\mathrm{P}} := \max_{\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \left\{ \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\Big] : R_p(\widehat{\mathbb{P}},\mathbb{P})^p \le \rho^p \right\}. \tag{1}$$

We derive the corresponding dual problem v_{D} as

$$v_{\mathrm{D}} := \inf_{\lambda \ge 0} \left\{ \lambda\rho^p + \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\lambda\epsilon \log \int_{\mathcal{X}} \exp\left(\frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)}{\lambda\epsilon} \right) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \right] \right\}, \tag{2a}$$

where

$$g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}} \left[\lambda\epsilon \log \int_{\mathcal{Y}} \exp\left(\frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon} \right) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y}) \right]. \tag{2b}$$

Next, we reformulate the dual problem v_{D} as a stochastic optimization with nested expectation structure such that, except for $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}$, all random vectors are mutually independent. Define the kernel probability distributions (e.g., they are Laplace or Gaussian distributions when $p \in \{1,2\}$) for the random vectors $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ as

$$\mathrm{d}Q_\epsilon(\boldsymbol{\xi}_1) := \frac{e^{-\|\boldsymbol{\xi}_1\|^p/\epsilon}}{\int_{\mathbb{R}^{d_x}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{\xi}_1), \tag{3a}$$

$$\mathrm{d}W_\epsilon(\boldsymbol{\xi}_2) := \frac{e^{-\|\boldsymbol{\xi}_2\|^p/\epsilon}}{\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})} \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{\xi}_2), \tag{3b}$$

and a constant

$$\bar{\rho} := \rho^p + \epsilon \cdot \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\log \int_{\mathbb{R}^{d_x}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u}) \right] + \epsilon \cdot \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}} \left[\log \int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u}) \right]. \tag{4}$$

Then, Problem (2) can be reformulated as

$$v_{\mathrm{D}} = \inf_{\lambda \ge 0} \left\{ \lambda\bar{\rho} + \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon} \left[\exp\left(\frac{g'(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon} \right) \right] \right] \right\}, \tag{5a}$$

where

$$g'(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}} \left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon} \left[\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon} \right) \right] \right]. \tag{5b}$$

In the following, Section 3.1 presents the main result of strong duality and related discussions. Section 3.2 presents the worst-case distribution of (1) and compares it with existing DRO models.

3.1 Main Results for the Dual Formulation

In this part, we first present the strong duality theorem that $v_{\mathrm{P}} = v_{\mathrm{D}}$, and next provide related discussions. We consider the following assumptions.

Assumption 1. (I) *Both $\mathcal{X}$ and $\mathcal{Z}$ are measurable sets, and the loss function $\Psi : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ and decision rule $f : \mathcal{X} \to \mathcal{Z}$ are measurable.*

(II) *For every joint distribution γ on $(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$ with first marginal distribution $\widehat{\mathbb{P}}$, it has a regular conditional distribution $\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}$ given the value of the first marginal equals $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})$.*

(III) *The transport cost function $c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))$ is measurable, and for $\widehat{\mathbb{P}} \otimes \nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}}$-almost every $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})$, it holds that $0 \le c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) < \infty$.*

(IV) *The normalization constants for the kernel distributions are positive $\int_{\mathbb{R}^{d_x}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u}) < \infty$ and $\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u}) < \infty$.*

Assumption 1(I) ensures that the expectation over $\Psi(f(\boldsymbol{x}),\boldsymbol{y})$ is well-defined. Assumption 1(II) ensures that each optimal transport plan can be decomposed into many conditional optimal transport plans. Assumptions 1(III) and 1(IV) ensure that the optimal value of Causal-SDRO is well-defined. Based on Assumption 1(IV), we introduce the light-tail condition on functions Ψ and g' in the following Condition 1 to distinguish the cases $v_{\mathrm{D}} < \infty$ and $v_{\mathrm{D}} = \infty$. We provide sufficient conditions to easily verify whether Condition 1 holds in E-companion EC.1.1.

Condition 1. *There exists $\lambda > 0$ such that $\mathbb{E}_{\boldsymbol{\xi}_1 \sim Q_\epsilon}\left[\exp\left(\frac{g'(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon}\right)\right] < \infty$ for $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$-almost every $\widehat{\boldsymbol{x}}$ and $\mathbb{E}_{\boldsymbol{\xi}_2 \sim W_\epsilon}\left[\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right] < \infty$ for $\widehat{\mathbb{P}} \otimes \nu_{\mathcal{X}}$-almost every $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})$.*

We call the constraint $R_p(\widehat{\mathbb{P}},\mathbb{P})^p \le \rho^p$ in primal problem (1) the CSD constraint in the following. Based on Assumption 1 and Condition 1, the following strong duality theorem holds.

Theorem 1. (Strong Duality). *Under Assumption 1, the following results hold.*

(I) *The primal problem v_{P} is feasible if and only if $\bar{\rho} \ge 0$.*

(II) *Additionally, assume $\bar{\rho} \ge 0$ is bounded above such that the CSD constraint is binding, then:*

- *If Condition 1 holds, then $v_{\mathrm{P}} = v_{\mathrm{D}} < \infty$;*
- *Otherwise, $v_{\mathrm{P}} = v_{\mathrm{D}} = \infty$.*

The proof of Theorem 1 is provided in E-companion EC.1.2. We present several remarks regarding Theorem 1.

Remark 1. (Comparison with Causal-WDRO). If $\epsilon \to 0$, then the dual objective of the problem (2) converges to (see E-companion EC.1.3 for detailed proof)

$$\lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\sup_{\boldsymbol{x}\in\operatorname{supp}\nu_{\mathcal{X}}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\left[\sup_{\boldsymbol{y}\in\operatorname{supp}\nu_{\mathcal{Y}}}\left\{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\right\} \mid \widehat{\boldsymbol{X}}\right]\right\}\right], \tag{6}$$

which is the same as the dual formulation of the Causal-WDRO problem in J. Yang et al. (2022). The optimization for Causal-WDRO is computationally challenging due to the nested inner supremums within expectations in (6), which leads to a non-smooth stochastic min-max structure. Existing algorithms for solving Causal-WDRO typically focus on special cases. For example, J. Yang et al. (2022) assume a specific structure for the loss function, and Hu et al. (2023) assume that λ is sufficiently large and convert it to a contextual stochastic bilevel optimization with a strongly convex lower level problem. ♣

With a proper level of entropy regularization, the supremum operators in (6) are replaced by smooth log-sum-exp type operators, which implies that the original dual problem is replaced by a special stochastic program, as discussed in Remark 2.

Remark 2. (Stochastic Optimization Formulation). Problem (5) can be rewritten as a stochastic multi-level compositional optimization problem:

$$v_{\mathrm{D}} = \inf_{\lambda\geq 0}\left\{\lambda\bar{\rho} + \lambda\epsilon\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}}\Big[h_1\Big(\mathbb{E}_{\boldsymbol{\xi}_1}\Big[h_2\Big(\mathbb{E}_{\widehat{\boldsymbol{y}}|\widehat{\boldsymbol{x}}}\Big[h_1\Big(\mathbb{E}_{\boldsymbol{\xi}_2}\Big[h_3\left(\lambda;\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right)\Big]\Big)\Big]\Big)\Big]\Big)\Big]\right\},$$

where the functions h_1, h_2, and h_3 are defined as

$$\begin{aligned}
&h_1:\mathbb{R}_+\to\mathbb{R}, && h_1(z)=\log(z),\\
&h_2:\mathbb{R}\to\mathbb{R}_+, && h_2(z)=e^z,\\
&h_3:\mathbb{R}_+\times\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\times\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\to\mathbb{R}, && h_3(\lambda;\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)=\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right).
\end{aligned}$$

Existing literature has provided different variants of optimization algorithms for solving this kind of formulation, such as M. Wang et al. (2017) and T. Chen et al. (2021). Both methods are gradient-based algorithms, introducing auxiliary variables to track the iterative update of inner expectations in the gradient computation. ♣

Remark 3. (Soft-constrained Causal-SDRO). In problem (Causal-SDRO) with hard CSD constraint, the radius $\bar{\rho}$ is a hyperparameter to be tuned while λ is the dual variable. However, if we regard the hard constraint as a soft one, it suffices to tune λ as a hyperparameter. The soft-constrained Causal-SDRO problem is given by

$$\inf_{f\in\mathcal{F}}\max_{\mathbb{P}\in\mathfrak{P}(\widehat{\mathbb{P}})}\quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\Big] - \lambda\cdot R_p(\widehat{\mathbb{P}},\mathbb{P})^p. \tag{Soft-Causal-SDRO}$$

For the inner problem in (Soft-Causal-SDRO), we derive its dual formulation by the Fenchel duality, then the dual problem is given by

$$\inf_{f\in\mathcal{F}}\quad \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon\log\int_{\mathcal{X}}\exp\left(\frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right], \tag{7}$$

where the definition of function g is the same as Equation (2b). Under Assumption 1, it can be shown that the strong duality of the inner problem in (Soft-Causal-SDRO) holds (by the similar proof process as in Theorem 1). The dual problem (7) is also a stochastic multi-level compositional optimization problem. The soft-constrained problem (Soft-Causal-SDRO) is easier to solve compared with (Causal-SDRO).

Remark 4. (Connection with KL-Divergence DRO). Let $\mathbb{D}_{\mathrm{KL}}\Big(\mathbb{P}\|\mathbb{Q}\Big)$ be the Kullback–Leibler (KL) divergence from distribution $\mathbb{P}$ to $\mathbb{Q}$. Then, the constraint $R_p(\widehat{\mathbb{P}},\mathbb{P})^p \le \rho^p$ in (Causal-SDRO) can be rewritten as (see E-companion EC.1.2 for details)

$$\mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\Big[\mathbb{D}_{\mathrm{KL}}\Big(\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}\|\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}\Big)\Big] \le \frac{\bar{\rho}}{\epsilon}, \tag{8}$$

where $\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}$ is the conditional distribution of γ given the value of the marginal $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})$, and $\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}$ is a kernel probability distribution defined by kernel distributions Q_ϵ and W_ϵ in Equation (3):

$$\mathrm{d}\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}(\boldsymbol{x},\boldsymbol{y}) := \mathrm{d}Q_\epsilon(\boldsymbol{x})\cdot\mathrm{d}W_\epsilon(\boldsymbol{y}).$$

Compare with the KL-divergence-based DRO problem (Ben-Tal et al., 2013; Blanchet et al., 2023) with constraint $\mathbb{D}_{\mathrm{KL}}\Big(\mathbb{P}\|\widehat{\mathbb{P}}\Big) \le \rho$, in constraint (8), the conditional distribution $\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}$ is remained due to the causal consideration in (Causal-SDRO), and the distribution $\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}$ can be viewed as a non-parametric kernel estimation constructed from $\widehat{\mathbb{P}}$. ♣

3.2 Worst-Case Distribution

As demonstrated in the proof of Theorem 1 (see E-companion EC.1.2 for details), Theorem 2 characterizes the worst-case distribution of the primal problem v_{P}.

Theorem 2. (Worst-case Distribution of Problem (Causal-SDRO)). *Under Assumption 1, suppose the dual problem v_{D} has an optimal solution $\lambda^* > 0$. Then the dual optimal solution λ^* is unique, and the density of worst-case distribution $\mathbb{P}^*$ of the primal problem v_{P} is given by*

$$\frac{\mathrm{d}\mathbb{P}^*(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\Big[\alpha_{\widehat{\boldsymbol{x}}}\cdot\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}}\cdot e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})+s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\Big], \tag{9}$$

where $\alpha_{\widehat{\boldsymbol{x}}} = \Big(\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\Big)^{-1}$, $\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}} = \Big(\int_{\mathcal{Y}} e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\Big)^{-1}$,

$$s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y}) = \frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda^* c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda^*\epsilon},$$

and

$$r(\widehat{\boldsymbol{x}},\boldsymbol{x}) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\Big[\log\int_{\mathcal{Y}} e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\Big].$$

We provide the proof of Theorem 2 in E-companion EC.1.4. Theorem 2 reveals that the worst-case distribution $\mathbb{P}^*$ is a mixture of Gibbs distributions, and Causal-SDRO spreads probability mass continuously over the support of the reference measure, governed by the regularization parameter ϵ. This result leads to the following Corollary 1.

Corollary 1. (Worst-case Distribution of Problem (7)). *Under Assumption 1, the density of worst-case distribution $\mathbb{P}^*_\lambda$ of the inner problem of (Soft-Causal-SDRO) for any $\lambda > 0$ is given by*

$$\frac{\mathrm{d}\mathbb{P}^*_\lambda(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\Big[\alpha_{\widehat{\boldsymbol{x}}}(\lambda)\cdot\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}}(\lambda)\cdot e^{r'(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})+s'(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\Big], \tag{10}$$

where $\alpha_{\widehat{\boldsymbol{x}}}(\lambda) = \left(\int_{\mathcal{X}} e^{r'(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \right)^{-1}$, $\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}}(\lambda) = \left(\int_{\mathcal{Y}} e^{s'(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})} \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y}) \right)^{-1}$,

$$s'(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y}) = \frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon},$$

and

$$r'(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x}) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\log \int_{\mathcal{Y}} e^{s'(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})} \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y}) \right].$$

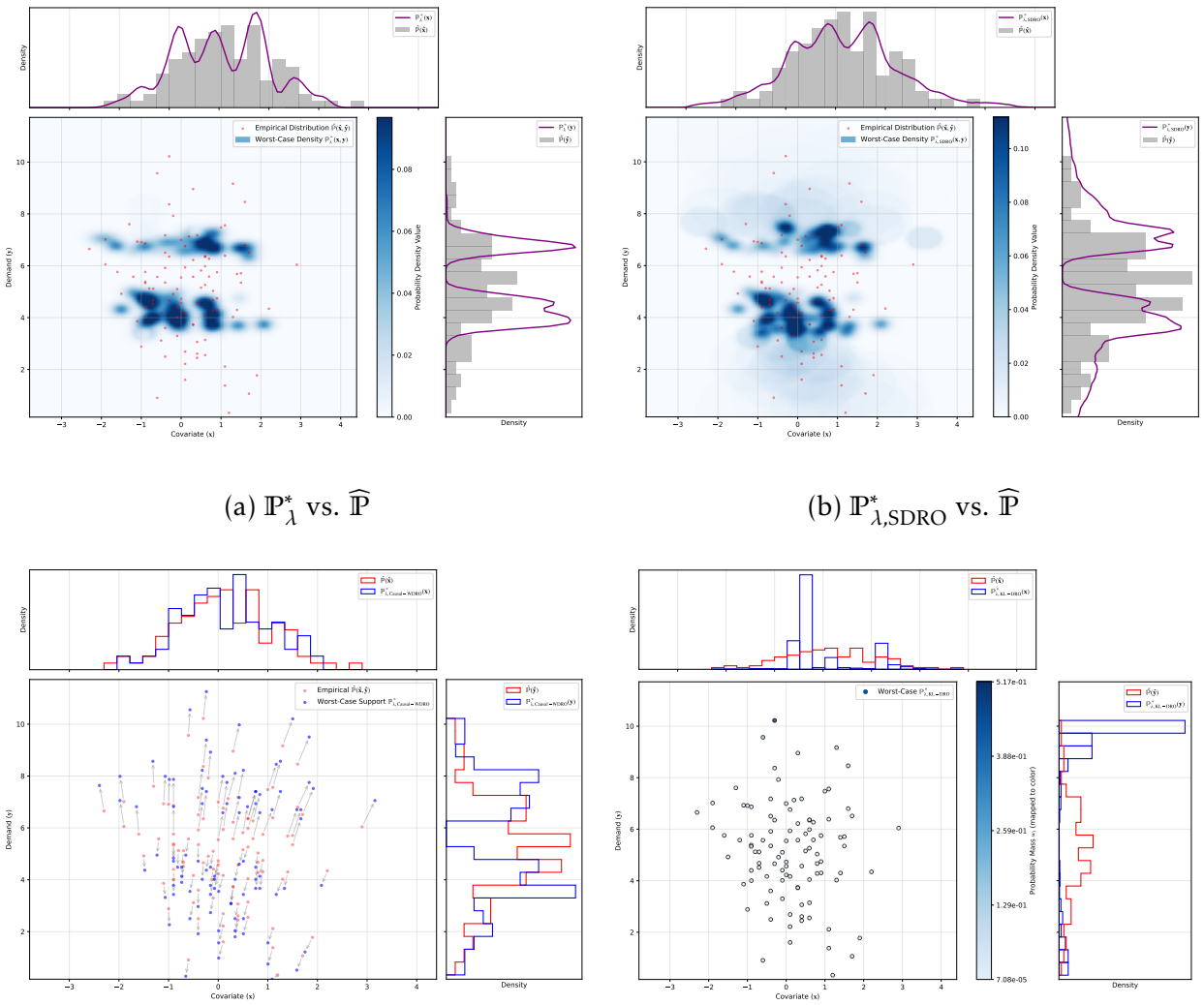
We next compare $\mathbb{P}^*_\lambda$ with the worst-case distribution formulation of soft-constrained Sinkhorn DRO (SDRO), Causal Wasserstein DRO (Causal-WDRO), and KL-divergence-based DRO (KL-DRO) models in contextual settings, denoted by $\mathbb{P}^*_{\lambda,\text{SDRO}}$, $\mathbb{P}^*_{\lambda,\text{Causal-WDRO}}$, and $\mathbb{P}^*_{\lambda,\text{KL-DRO}}$. We summarize the formulations of these models and worst-case distributions in E-companion EC.2.

Example 1. (Revisited). Consider a single-product feature-based newsvendor problem with one covariate, i.e., $d_x = d_y = d_z = 1$. Consider a true decision rule, i.e., $f = f_{\text{true}}$, in Figure 2, we show the structure of distributions $\widehat{\mathbb{P}}$, $\mathbb{P}^*_\lambda$, $\mathbb{P}^*_{\lambda,\text{SDRO}}$, $\mathbb{P}^*_{\lambda,\text{Causal-WDRO}}$, and $\mathbb{P}^*_{\lambda,\text{KL-DRO}}$, as well as their marginal probability density or mass. In Figures 2(a)-2(c), historical data points (i.e., empirical distribution $\widehat{\mathbb{P}}$) are marked in red. Figure 2(a) shows the structure of $\mathbb{P}^*_\lambda$, while Figure 2(b) shows the corresponding structure of $\mathbb{P}^*_{\lambda,\text{SDRO}}$. Comparing Figure 2(a) with Figure 2(b), the probability density of $\mathbb{P}^*_\lambda$ is more concentrated than that of $\mathbb{P}^*_{\lambda,\text{SDRO}}$. This is because the causal transport constraint prevents transport plans that violate the conditional independence between $\boldsymbol{x}$ and $\widehat{\boldsymbol{y}}$ given $\widehat{\boldsymbol{x}}$, which allows Causal-SDRO to avoid overly conservative results. In Figure 2(c), for each point in $\mathbb{P}^*_{\lambda,\text{Causal-WDRO}}$, we mark how they are transported from the empirical distribution with arrows. In Figure 2(d), as $\mathbb{P}^*_{\lambda,\text{KL-DRO}}$ has the same support as $\widehat{\mathbb{P}}$, we show the structure of $\mathbb{P}^*_{\lambda,\text{KL-DRO}}$ by color depth, where a darker color of a point means a greater probability mass. ♣

The visualization in Example 1 corroborates our theoretical findings regarding the structure of worst-case distributions. As illustrated, the worst-case distributions for Causal-WDRO and KL-DRO are inherently discrete (supported on finite points), while for Causal-SDRO and SDRO, the entropic regularization leads to continuous worst-case distributions. Crucially, distinguishing Causal-SDRO from standard SDRO, our worst-case distribution strictly remains causally consistent, thereby avoiding causally implausible robustness scenarios.

4 Soft Regression Forest Decision Rule

Optimizing policies in a general measurable function space is computationally challenging due to the infinite-dimensional functional optimization involved. Instead, we consider a parametric decision rule approach $f : \mathcal{X} \to \mathcal{Z}$ that approximates the optimal mapping between covariates and decisions. In this section, we propose a parametric and interpretable Soft Regression Forest (SRF) decision rule. Section 4.1 introduces the structure of this decision rule, and Section 4.2 discusses its intrinsic interpretability.



(a) $\mathbb{P}^*_\lambda$ vs. $\widehat{\mathbb{P}}$

(b) $\mathbb{P}^*_{\lambda,\text{SDRO}}$ vs. $\widehat{\mathbb{P}}$

(c) $\mathbb{P}^*_{\lambda,\text{Causal-WDRO}}$ vs. $\widehat{\mathbb{P}}$

(d) $\mathbb{P}^*_{\lambda,\text{KL-DRO}}$

Figure 2. Structure of distributions $\widehat{\mathbb{P}}$ (red points in 2(a)-2(c)), $\mathbb{P}^*_\lambda$, $\mathbb{P}^*_{\lambda,\text{SDRO}}$, $\mathbb{P}^*_{\lambda,\text{Causal-WDRO}}$, and $\mathbb{P}^*_{\lambda,\text{KL-DRO}}$ ($p = 2$, $\lambda = 0.5$, $\epsilon = 0.05$, and sample size $N = 100$)

4.1 Structure of the Soft Regression Forest

In practice, the decision-making process may follow a hierarchical and interpretable structure, such as an 'if-then' structure, rather than a fixed and continuous function. To capture this structure, unlike the traditional deep-learning-based methods, the proposed SRF decision rule is based on the principles of soft decision trees (Frosst and Hinton, 2017) and ensemble learning. Compared with the traditional *hard* decision-tree-based methods, SRF is parametric, differentiable, and can be end-to-end trained by gradient-based algorithms, while maintaining the intrinsic interpretability.

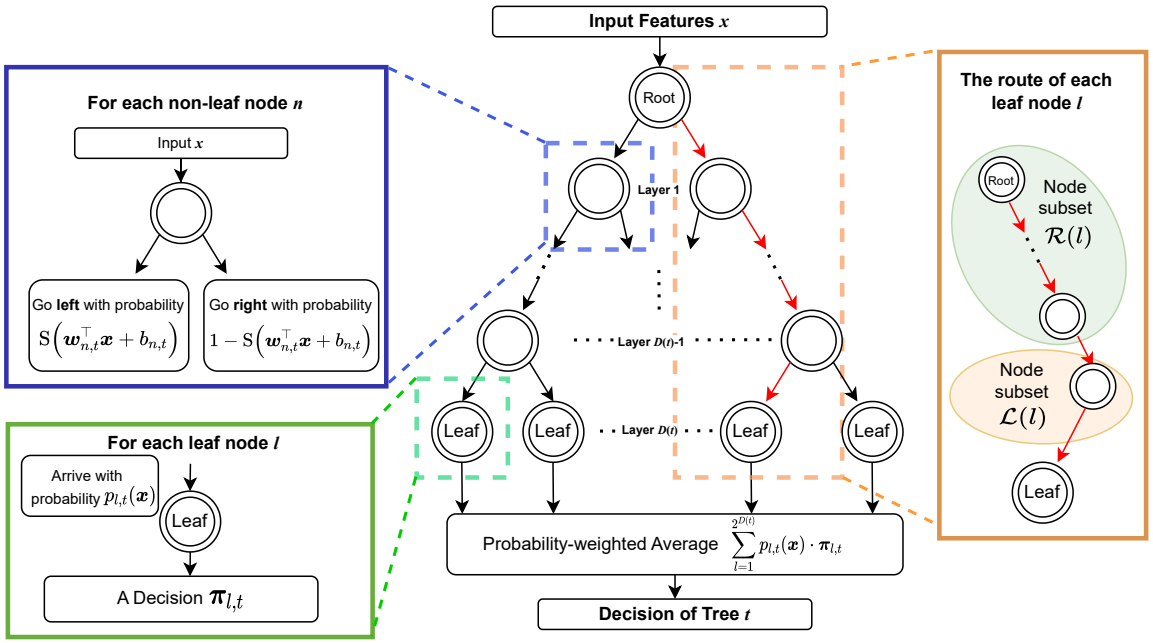


Figure 3. Structure of a soft regression tree t with $D(t)$ depth

The SRF consists of an ensemble of T full binary Soft Regression Trees (SRTs). For the t-th ($t \in [T]$) tree with depth $D(t)$, each leaf node $l \in \left[2^{D(t)}\right]$ (i.e., a node has no child node) corresponds to a decision $\pi_{l,t} \in \mathbb{R}_+^{d_z}$ and a unique route from the root node in the tree. For each route of leaf node $l \in \left[2^{D(t)}\right]$ for any $t \in [T]$, denote the left-hand-side and right-hand-side node sets on the route as $\mathcal{L}(l)$ and $\mathcal{R}(l)$, respectively, and $\Lambda(l) := \mathcal{L}(l) \cup \mathcal{R}(l)$.

Distinct from hard regression tree method that select a determined child-node at each branch, in an SRT t, at each internal node $j \in [2^{D(t)} - 1]$ in SRT t, a gating function $S(w_{j,t}^\top x + b_{j,t})$ determines the probability of directing the input x to the left child, where $S(\cdot)$ represents the Sigmoid function and $w_{j,t} \in \mathbb{R}^{d_x}$, $b_{j,t} \in \mathbb{R}$. Consequently, the probability that a given input covariate x reaches leaf node l (i.e., the decision $\pi_{l,t}$) in tree t is given by

$$p_{l,t}(x) = \prod_{i \in \mathcal{L}(l)} S(w_{i,t}^\top x + b_{i,t}) \cdot \prod_{j \in \mathcal{R}(l)} \left(1 - S(w_{j,t}^\top x + b_{j,t})\right).$$

The final output of the SRF is the ensemble average of the expected decisions from all trees, and thus the SRF decision rule $f_\theta^{\mathrm{SRF}} : \mathcal{X} \to \mathcal{Z}$ is explicitly defined as

$$\left[f_\theta^{\mathrm{SRF}}(x)\right]_k := \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} p_{l,t}(x) \cdot \left[\pi_{l,t}\right]_k, \quad \forall k \in [d_z], \tag{SRF}$$

where $[\pi_{l,t}]_k$ represents the k-th decision for any $k \in [d_z]$, the vector $\theta \in \Theta$ is the collection of all individual parameters $\{w_{i,t}, b_{i,t}, \pi_{l,t}\}$ for each non-leaf node $i \in \Lambda(l)$, leaf node $l \in \left[2^{D(t)}\right]$ and tree $t \in [T]$. This decision rule needs to train $(d_x + 1) \cdot \sum_{t=1}^{T} (2^{D(t)} - 1) + d_z \cdot \sum_{t=1}^{T} 2^{D(t)}$ parameters in total.

Figure 3 shows the structure of a single SRT $t \in [T]$ with $D(t)$ layers (let the root node be in layer 0). As illustrated, on the left-hand side, we show the input and output structure of each node, while on the right-hand side, we show the route and node subsets $\mathcal{L}(l)$ and $\mathcal{R}(l)$ for each leaf node. The SRF employs an ensemble of SRTs to mitigate the high variance and potential overfitting risks associated with individual trees (Breiman, 2001).

Compared with existing deep-learning-based decision rules, this proposed decision rule with a hierarchical structure and probabilistic decisions possesses intrinsic interpretability. In SRF, each SRT is similar to a distilled non-fully connected multi-layer neural network where only the important nodes are connected, which enhances the efficiency of feature representation.

Recall that Bertsimas and Kallus (2020) and Kallus and Mao (2023) introduce tree-based model for solving CSO. In their framework, they use hard split decision trees to estimate the conditional local weights in the weighted sample average approximation method. In comparison, the proposed parametric SRF decision rule makes end-to-end decisions based on covariates and is applicable for both CSO and contextual DRO.

4.2 Interpretability of the Soft Regression Forest

In this subsection, we demonstrate the intrinsic interpretability of SRF by its transparent structure and stability for decision-making.

As a tree-based model, SRF inherits the transparent structure from the traditional methods.

Remark 5. (Asymptotic Consistency to Hard Regression Forest). Adding a scaling parameter τ to the linear transformation part at each internal node in all SRTs, i.e., $\mathrm{S}\Big((\boldsymbol{w}^\top \boldsymbol{x} + b)/\tau\Big)$ for all $j(t) \in [2^{D(t)} - 1]$, we obtain a variant of SRF decision rule termed $f_{\theta,\tau}^{SRF}(\boldsymbol{x})$ where each leaf node $l \in [2^{D(t)}]$ in tree t can be reached with probability $p_{l,t,\tau}$. The proposed SRF $f_{\theta}^{SRF}(\boldsymbol{x})$ shown in Section 4.1 is a special case with $\tau = 1$. For any input covariate $\boldsymbol{x}$ not lying on any decision boundary (i.e., $\{\boldsymbol{x} \in \mathcal{X} \mid \boldsymbol{w}_{j,t}^\top \boldsymbol{x} + \boldsymbol{b}_{j,t} \neq 0, \forall j, t\}$), when $\tau \to 0$, the structure of SRF converges to a hard regression forest, which implies that the sigmoid function takes value only in $\{0, 1\}$ and thus only one deterministic leaf node can be selected as the final decision, i.e., $\sum_{i=1}^{2^{D(t)}} p_{i,t,\tau} = 1$ and $\lim_{\tau \to 0} p_{i,t,\tau} \in \{0, 1\}$. Unlike traditional univariate decision trees, all trees in the resulting hard regression forest provide multivariate splits at all nodes, which improve the accuracy and interpretability by reducing tree depth (Bertsimas and Dunn, 2017; Bertsimas and Stellato, 2021).

Although the SRT theoretically aggregates outputs across all routes, probabilities for weakly correlated routes effectively vanish as they are calculated as products of several Sigmoid functions. Consequently, the final decision is typically dominated by a few high-probability routes. This inherent sparsity enhances interpretability, enabling decision-makers to easily trace the primary routes driving the final prescription. We provide empirical evidence for this in Section 6.3.

As traditional decision trees allow for tracing the decision process via routes and identifying the impact of each feature, the SRT can also explicitly trace the influence of features and their interaction effects along each route.

Proposition 1. (Traceability of Decisions in SRF). *In SRF, given an input covariate $\boldsymbol{x} \in \mathbb{R}^{d_x}$, for each route selected with probability $p_{l,t}(\boldsymbol{x})$, the marginal contribution of each feature and the interaction effects among features along that route are explicitly characterized by*

$$\begin{aligned}
\frac{\partial p_{l,t}(\boldsymbol{x})}{\partial x_j} &= p_{l,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l)} \psi_{i,t} \Big[\boldsymbol{w}_{i,t}\Big]_j, \\
\frac{\partial^2 p_{l,t}(\boldsymbol{x})}{\partial x_j \partial x_k} &= p_{l,t}(\boldsymbol{x}) \cdot \Bigg[\Big(\sum_{i \in \Lambda(l)} \psi_{i,t} \Big[\boldsymbol{w}_{i,t}\Big]_j \Big) \Big(\sum_{i \in \Lambda(l)} \psi_{i,t} \Big[\boldsymbol{w}_{i,t}\Big]_k \Big) - \\
&\qquad \sum_{i \in \Lambda(l)} S\Big(\boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}\Big)\Big(1 - S\Big(\boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}\Big)\Big) \Big[\boldsymbol{w}_{i,t}\Big]_j \Big[\boldsymbol{w}_{i,t}\Big]_k \Bigg], \\
&\forall j,k \in [d_x],\, l \in [2^{D(t)}],\, t \in T,
\end{aligned}$$

where for each $i \in \Lambda(l), l \in [2^{D(t)}], t \in T$,

$$\psi_{i,t} := \begin{cases} 1 - S\Big(\boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}\Big), & \text{if route } l \text{ goes left at node } i; \\ -S\Big(\boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}\Big), & \text{if route } l \text{ goes right at node } i. \end{cases}$$

We provide the proof of Proposition 1 in E-companion EC.1.5. Although many uninterpretable deep learning models are also differentiable, their gradients are typically aggregated through opaque dense layers, obscuring the internal decision mechanism. In contrast, the SRF derivatives explicitly decompose the feature influence into specific decision nodes along the route, allowing us to exactly trace *where* (at which node) and *how* (direction and magnitude) a feature contributes to the decision process.

Beyond the transparency and traceability, we next show that the mathematical smoothness of the SRF structure also contributes to interpretability. We define $W_{\max} := \max_{i,t} \|\boldsymbol{w}_{i,t}\|_2$ as the maximum norm of the internal node weights, $\Pi_{\max} := \max_{l,t} \|\boldsymbol{\pi}_{l,t}\|_2$ as the maximum norm of leaf vectors, and $D_{\max}$ as the maximum tree depth. Then, the following proposition holds.

Proposition 2. (Lipschitz Continuity and Smoothness of SRF). *The SRF decision rule $f_{\boldsymbol{\theta}}^{\mathrm{SRF}} : \mathcal{X} \to \mathcal{Z}$ is L^{SRF}-Lipschitz continuous and S^{SRF}-Lipschitz on the compact set $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, where*

$$\begin{aligned}
L^{\mathrm{SRF}} &= W_{\max} \cdot \Pi_{\max} \cdot (D_{\max} - 1), \\
S^{\mathrm{SRF}} &= W_{\max}^2 \cdot \Pi_{\max} \cdot (D_{\max} - 1) \cdot (D_{\max} - \frac{3}{4}).
\end{aligned}$$

We provide the proof of Proposition 2 in E-companion EC.1.6. These Lipschitz properties confirm the decision stability and robustness (interpretation stability) of SRF, distinguishing it from existing uninterpretable deep learning models and post-hoc explanation methods, which are typically not Lipschitz as small input perturbations may lead to abrupt changes in decisions and explanations (Alvarez-Melis and Jaakkola, 2018).

All analyses above demonstrate the structural transparency, stability, and robustness of the SRF decision rule. In E-companion EC.3, we further introduce both global and local intrinsic interpretation measures for SRF, which depend only on the structure of SRF and avoid post-hoc

explanation analyses. In the following Section 6.3, we confirm the practical interpretability of SRF based on its structure and the proposed intrinsic interpretation measures on the portfolio problem shown in Example 3 with real data.

5 Solving Causal-SDRO

In this section, we discuss the algorithms to solve the Causal-SDRO problem. In Section 5.1, we reformulate (7) as a three-level stochastic compositional optimization. For this tractable formulation, we analyze the sample and computational complexity of the sample average approximation method in Section 5.2, and develop a gradient-based algorithm to solve it in Section 5.3.

5.1 Tractable Reformulation for Causal-SDRO

Let each decision rule $f_{\boldsymbol{\theta}}$ be parameterized by a parameter vector $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{d_\theta}$. In this subsection, we reformulate the dual problem of (Soft-Causal-SDRO), which is given by:

$$\min_{\boldsymbol{\theta}\in\Theta} \quad F(\boldsymbol{\theta}) = \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\exp\left(\frac{h(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon}\right)\right]\right] \tag{11a}$$

where

$$h(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right]\right]. \tag{11b}$$

Since the nominal distribution $\widehat{\mathbb{P}}$ is the discrete empirical distribution from the training data, the conditional distribution $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{y}}|\widehat{\boldsymbol{x}}}$ also has a finite support for any given historical covariate $\widehat{\boldsymbol{x}}$. As all historical data are available, the conditional probability $\widehat{p}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{x}})$ given $\widehat{\boldsymbol{x}}$ can be estimated by the empirical frequency, i.e.,

$$\widehat{p}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{x}}) = \frac{1}{n_{\widehat{\boldsymbol{x}}}} \sum_{j=1}^{n_{\widehat{\boldsymbol{x}}}} \mathbb{I}\left(\widehat{\boldsymbol{y}}_j = \widehat{\boldsymbol{y}}_i\right),$$

where $n_{\widehat{\boldsymbol{x}}}$ is the number of observed outcomes for $\widehat{\boldsymbol{y}}$ associated with covariate $\widehat{\boldsymbol{x}}$, and function $\mathbb{I}(\cdot)$ is an indicator function. Therefore, the conditional expectation in Equation (11b) can be computed by

$$h(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda) = \lambda\epsilon \cdot \sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}} \widehat{p}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{x}}) \cdot \log \mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_i+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right].$$

Then, the problem (11) is equivalent to a three-level stochastic compositional optimization (SCO) problem, driven by the three independent random vectors $\widehat{\boldsymbol{x}}$, $\boldsymbol{\xi}_1$, and $\boldsymbol{\xi}_2$:

$$\min_{\boldsymbol{\theta}\in\Theta} \quad F(\boldsymbol{\theta}) = \lambda\epsilon \cdot \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[t_1\left(\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[t_3\left(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2\right)\right];\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1\right)\right];\widehat{\boldsymbol{x}}\right)\right] \tag{SCO}$$

where

$$\begin{aligned}
&t_1 : \mathbb{R}_+ \to \mathbb{R}, && t_1(z;\widehat{\boldsymbol{x}}) = \log(z),\\
&t_2 : \mathbb{R}^{n_{\widehat{\boldsymbol{x}}}} \to \mathbb{R}_+, && t_2(\boldsymbol{v};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) = \exp\left(\sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}} \widehat{p}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{x}}) \cdot \log(v_i)\right),\\
&t_3 : \mathbb{R}^{d_\theta} \to \mathbb{R}^{n_{\widehat{\boldsymbol{x}}}}, && \left[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\right]_i = \exp\left(\frac{\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_i+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right), \forall i \in [n_{\widehat{\boldsymbol{x}}}],
\end{aligned} \tag{12}$$

where the set of vectors $\{\widehat{\boldsymbol{y}}_i\}_{i=1}^{n_{\widehat{\boldsymbol{x}}}}$ is implicitly defined by a covariate $\widehat{\boldsymbol{x}}$. For brevity, we denote functions $t_1(z;\widehat{\boldsymbol{x}})$, $t_2(\boldsymbol{v};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)$, and $t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)$ as $t_1(z)$, $t_2(\boldsymbol{v})$, and $t_3(\boldsymbol{\theta})$ respectively. We also define

$$\phi^{(0)}_{\widehat{\boldsymbol{x}}}(\boldsymbol{\theta}) := \mathbb{E}_{\widehat{\boldsymbol{x}}}\Big[t_1(\boldsymbol{\theta};\widehat{\boldsymbol{x}})\Big], \quad \phi^{(1)}_{\boldsymbol{\xi}_1}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_1}\Big[t_2(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)\Big], \text{ and } \phi^{(2)}_{\boldsymbol{\xi}_2}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\Big].$$

In the following subsections, we introduce several assumptions for this problem in Assumption 2, which are commonly used by related literature, e.g., Hu et al. (2020) and Shapiro et al. (2021).

Assumption 2. *We assume that*

(I) *(Bounded Diameter). The decision set $\Theta \subseteq \mathbb{R}^{d_\theta}$ has a positive finite diameter $D_\Theta > 0$, that is, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \le D_\Theta$.*

(II) *(Lipschitz Continuity). For any fixed $\boldsymbol{x}$ and $\boldsymbol{y}$ and given parameteric decision rule, the loss function $L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) := \Psi(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{y})$ is $L_{\boldsymbol{\theta}}$-Lipschitz continuous with respect to $\boldsymbol{\theta}$.*

(III) *(Bounded Cost). The loss function $\Psi(\boldsymbol{z},\boldsymbol{y})$ satisfies $0 \le \Psi(\boldsymbol{z},\boldsymbol{y}) \le B$ for any $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{y} \in \mathcal{Y}$.*

Assumption 2(I) on the diameter of the decision space is used for sample complexity analysis. Assumption 2(II) is crucial for deriving the convergence rate of the gradient-based algorithms. From Assumption 2, we have the following Proposition 3.

Proposition 3. (Properties of Problem (SCO)). *Under Assumption 2, functions t_1, t_2, and t_3 in Equation (12):*

(I) *are L_1-, L_2-, and L_3-Lipschitz continuous;*

(II) *are S_1-, S_2-, and S_3-Lipschitz smooth;*

(III) *have bounded stochastic gradients in expectation, i.e., $\mathbb{E}\Big[|\nabla t_1(z)|^2\Big] \le C_1^2$, $\mathbb{E}\Big[\|\nabla t_2(\boldsymbol{v})\|_2^2\Big] \le C_2^2$, and $\mathbb{E}\Big[\|\nabla t_3(\boldsymbol{\theta})\|_2^2\Big] \le C_3^2$;*

(IV) *have finite variances, i.e., $\sigma_1^2 = \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{V}_{\widehat{\boldsymbol{x}}}\Big(t_1\Big(\phi^{(1)}_{\boldsymbol{\xi}_1}(\boldsymbol{\theta})\Big)\Big) < \infty$, $\sigma_2^2 = \sup_{\boldsymbol{\theta}\in\Theta,\widehat{\boldsymbol{x}}} \mathbb{V}_{\boldsymbol{\xi}_1}\Big(t_2\Big(\phi^{(2)}_{\boldsymbol{\xi}_2}(\boldsymbol{\theta})\Big)\Big) < \infty$, and $\sigma_3^2 = \sup_{\boldsymbol{\theta}\in\Theta,\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}}} \mathbb{V}_{\boldsymbol{\xi}_2}\Big(t_3\Big(\boldsymbol{\theta}\Big)\Big) < \infty$;*

where

$$L_1 = S_1 = C_1 = 1; \qquad L_2 = C_2 = \exp\Big(B/\lambda\epsilon\Big), \quad S_2 = \sqrt{2}L_2;$$
$$L_3 = \frac{1}{\lambda\epsilon}\exp\Big(B/\lambda\epsilon\Big), \quad S_3 = \frac{1}{\lambda\epsilon}L_3, \quad C_3 = L_3 L_{\boldsymbol{\theta}}\sqrt{\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[n_{\widehat{\boldsymbol{x}}}\Big]};$$

where $\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[n_{\widehat{\boldsymbol{x}}}\Big]$ is a finite positive constant since the expectation is over the finite support of the empirical distribution $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$.

We provide the proof of Proposition 3 in E-companion EC.1.7.

5.2 Complexity of Sample Average Approximation

A standard method for solving the stochastic compositional optimization problem (SCO) is the Sample Average Approximation (SAA), which replaces each nested expectation with its corresponding empirical average, constructed from finite samples generated by the Monte Carlo sampling technique. In this subsection, we analyze the sample and computational complexity of SAA on our problem.

Specifically, we draw N_1 independent and identically distributed (i.i.d.) samples $\{(\widehat{\boldsymbol{x}}^i, \widehat{\boldsymbol{y}}^i)\}_{i=1}^{N_1}$ from the nominal distribution $\widehat{\mathbb{P}}$, N_2 i.i.d. samples $\{\boldsymbol{\xi}_1^j\}_{j=1}^{N_2}$ from the kernel distribution Q_ϵ, and N_3 i.i.d. samples $\{\boldsymbol{\xi}_2^k\}_{k=1}^{N_3}$ from the kernel distribution W_ϵ. This leads to the following SAA formulation for (SCO):

$$\min_{\boldsymbol{\theta}\in\Theta} \quad \widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}) = \frac{\lambda\epsilon}{N_1}\sum_{i=1}^{N_1} t_1\Big(\frac{1}{N_2}\sum_{j=1}^{N_2} t_2\Big(\frac{1}{N_3}\sum_{k=1}^{N_3} t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\widehat{\boldsymbol{y}}^i,\boldsymbol{\xi}_2^k\Big);\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j\Big);\widehat{\boldsymbol{x}}\Big). \tag{SAA}$$

Let $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}$ be the optimal solutions of the problems (SCO) and (SAA), respectively. We next analyze the number of samples required for the solution to the problem (SAA) to be δ-optimal of the problem (SCO) with high probability, i.e. $\Pr\Big(F\Big(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\Big) - F(\boldsymbol{\theta}^*) \le \delta\Big) \ge 1-\alpha$ for any $\delta > 0$ and $\alpha \in (0,1)$.

Using Assumption 2, we derive the sample complexity of the SAA method on this problem in the following Theorem 3.

Theorem 3. (Sample Complexity for Problem (SAA)). *Under Assumption 2, the following results hold.*

(I) *For any $\kappa > 0$, there exists an $\delta_1 > 0$ such that for any $\delta \in (0,\delta_1)$, it holds that*

$$\begin{aligned}
&\Pr\Big(F\Big(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\Big) - F(\boldsymbol{\theta}^*) > \delta\Big)\\
\le&\mathcal{O}(1)\Big(\frac{8L_1L_2L_3D_\Theta}{\delta}\Big)^{d_\theta}\Big(N_1N_2n_{\widehat{\boldsymbol{x}}}\exp\Big(-\frac{N_3\delta^2}{144(2+\kappa)\lambda^2\epsilon^2L_1^2L_2^2\sigma_3^2}\Big)\\
&\qquad + N_1\exp\Big(-\frac{N_2\delta^2}{144(2+\kappa)\lambda^2\epsilon^2L_1^2\sigma_2^2}\Big) + \exp\Big(-\frac{N_1\delta^2}{144(2+\kappa)\lambda^2\epsilon^2\sigma_1^2}\Big)\Big).
\end{aligned}$$

(II) *With probability at least $1-\alpha$, the solution to Problem (SAA) is δ-optimal to the original problem (SCO) if the sample sizes N_1, N_2, and N_3 satisfy that*

$$\begin{aligned}
N_1 &> \frac{\mathcal{O}(1)\sigma_1^2}{\delta^2}\Big[d_\theta\log\Big(\frac{8L_1L_2L_3D_\Theta}{\delta}\Big) + \log\Big(\frac{1}{\alpha}\Big)\Big],\\
N_2 &> \frac{\mathcal{O}(1)L_1^2\sigma_2^2}{\delta^2}\Big[d_\theta\log\Big(\frac{8L_1L_2L_3D_\Theta}{\delta}\Big) + \log\Big(\frac{1}{\alpha}\Big) + \log(N_1)\Big],
\end{aligned}$$

and

$$N_3 > \frac{\mathcal{O}(1)L_1^2L_2^2\sigma_3^2}{\delta^2}\Big[d_\theta\log\Big(\frac{8L_1L_2L_3D_\Theta}{\delta}\Big) + \log\Big(\frac{1}{\alpha}\Big) + \log(N_1N_2n_{\widehat{\boldsymbol{x}}})\Big].$$

Ignoring the log factors, the total sample complexity of problem (SAA) for achieving a δ-optimal solution is $T = N_1 + N_2 + N_3 = \mathcal{O}\left(d_\theta/\delta^2\right)$.

The proof of Theorem 3 is provided in E-companion EC.1.8. For the problem (SAA), the computational complexity of using the gradient descent (GD) algorithm is at least $\mathcal{O}\left(d_\theta^3/\delta^6\right)$, since a single iteration requires $N_1 \times N_2 \times N_3$ gradient updates. Similarly, the computational complexity of using the unbiased stochastic gradient descent (SGD) algorithm is at least $\mathcal{O}\left(d_\theta^2/\delta^4\right)$, which is still computationally challenging. This motivates us to develop an efficient algorithm for solving this SCO problem.

5.3 Stochastic Compositional Algorithm

In this subsection, we introduce a gradient-based algorithm for the problem (SCO). Before introducing the gradient algorithm for the stochastic compositional optimization problem, we first show the inherent challenge of applying the standard stochastic gradient descent (SGD) method to Problem (SCO). The true gradient of function $F(\boldsymbol{\theta})$ at point $\boldsymbol{\theta}^k$ (i.e., the vector at iteration k) is given by

$$\nabla F(\boldsymbol{\theta}^k) = \mathbb{E}_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2}\left[\nabla t_1\left(\phi^{(1)}_{\boldsymbol{\xi}_1}(\boldsymbol{\theta}^k)\right)\cdot\nabla t_2\left(\phi^{(2)}_{\boldsymbol{\xi}_2}(\boldsymbol{\theta}^k)\right)\cdot\nabla t_3(\boldsymbol{\theta}^k)\right].$$

Given the samples $(\widehat{\boldsymbol{x}}^k,\widehat{\boldsymbol{y}}^k)$, $\boldsymbol{\xi}_1^k$, and $\boldsymbol{\xi}_2^k$ drawn at iteration k, for brevity, we define

$$t_1^k(z) := t_1\left(z;\widehat{\boldsymbol{x}}^k\right),\quad t_2^k(\boldsymbol{v}) := t_2\left(\boldsymbol{v};\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k\right),\text{ and } t_3^k(\boldsymbol{\theta}) := t_3\left(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k,\widehat{\boldsymbol{y}}^k,\boldsymbol{\xi}_2^k\right).$$

The SGD method replaces $t_2\left(\phi^{(2)}_{\boldsymbol{\xi}_2}(\boldsymbol{\theta}^k)\right)$ by $t_2^k\left(t_3^k\left(\boldsymbol{\theta}^k\right)\right)$, and $t_1\left(\phi^{(1)}_{\boldsymbol{\xi}_1}(\boldsymbol{\theta}^k)\right)$ by $t_1^k\left(t_2^k\left(t_3^k\left(\boldsymbol{\theta}^k\right)\right)\right)$ in each iteration. That is, it simplifies the computation of the true gradient $\nabla F(\boldsymbol{\theta})$ by replacing the expected values with stochastic estimates computed from single random samples. However, as functions t_1, t_2, and t_3 are all non-linear, the stochastic gradient of the SGD method, denoted as $\left(\nabla F(\boldsymbol{\theta}^k)\right)_{\text{SGD}}$, is biased, i.e.,

$$\left(\nabla F(\boldsymbol{\theta}^k)\right)_{\text{SGD}} = \mathbb{E}_{\widehat{\boldsymbol{x}}^k,\widehat{\boldsymbol{y}}^k,\boldsymbol{\xi}_1^k,\boldsymbol{\xi}_2^k}\left[\nabla t_1^k\left(t_2^k\left(t_3^k\left(\boldsymbol{\theta}^k\right)\right)\right)\cdot\nabla t_2^k\left(t_3^k\left(\boldsymbol{\theta}^k\right)\right)\cdot\nabla t_3^k(\boldsymbol{\theta}^k)\right] \neq \nabla F(\boldsymbol{\theta}^k).$$

Since the bias of the standard SGD method is uncontrollable, it cannot be used to solve the problem (SCO) directly.

Therefore, we provide a Stochastically Corrected Stochastic Compositional gradient method (SCSC, T. Chen et al., 2021) to solve the problem (SCO), which controls the bias using momentum gradient updates. This method provides estimators for the expectations in $\nabla F(\boldsymbol{\theta})$ at each iteration. Specifically, in each iteration k with samples $(\widehat{\boldsymbol{x}}^k,\widehat{\boldsymbol{y}}^k)$, $\boldsymbol{\xi}_1^k$, and $\boldsymbol{\xi}_2^k$, functions $\phi^{(1)}_{\boldsymbol{\xi}_1}$ and $\phi^{(2)}_{\boldsymbol{\xi}_2}$ are estimated by y_1^k and $\boldsymbol{y}_2^k$, respectively, and thereby the parameters of decision rule are updated by

$$\boldsymbol{\theta}^{k+1} := \boldsymbol{\theta}^k - \alpha_k\cdot\nabla t_1^k(y_1^k)\cdot\nabla t_2^k(\boldsymbol{y}_2^k)\cdot\nabla t_3^k(\boldsymbol{\theta}^k), \tag{13}$$

where

$$y_1^{k+1} = (1-\beta_k)\cdot\left(y_1^k + t_2^k(\boldsymbol{y}_2^{k+1}) - t_2^k(\boldsymbol{y}_2^k)\right) + \beta_k\cdot t_2^k(\boldsymbol{y}_2^{k+1}), \tag{14}$$

and

$$\boldsymbol{y}_2^{k+1} = (1-\beta_k)\cdot\left(\boldsymbol{y}_2^k + t_3^k(\boldsymbol{\theta}^k) - t_3^k(\boldsymbol{\theta}^{k-1})\right) + \beta_k \cdot t_3^k(\boldsymbol{\theta}^k). \tag{15}$$

Compared to the stochastic compositional gradient descent method proposed by M. Wang et al. (2017) which update the y_1 and $\boldsymbol{y}_2$ by

$$y_1^{k+1} = (1-\beta_k)\cdot y_1^k + \beta_k \cdot t_2^k(\boldsymbol{\theta}^k),$$

and

$$\boldsymbol{y}_2^{k+1} = (1-\beta_k)\cdot \boldsymbol{y}_2^k + \beta_k \cdot t_3^k(\boldsymbol{\theta}^k),$$

the SCSC method adds a correction on $\boldsymbol{y}^k$ to avoid the information lag as $\boldsymbol{y}^k$ is updated by the outdated $\boldsymbol{\theta}^{k-1}$.

The pseudo-code of the SCSC is shown as the following Algorithm 1.

Algorithm 1 : SCSC for problem (SCO)

1: Initialize $\boldsymbol{\theta}^0$, y_1^0, $\boldsymbol{y}_2^0$, stepsizes α_0, β_0
2: **for** $k = 1, \cdots, K$, **do**
3: select $(\widehat{\boldsymbol{x}}^k, \widehat{\boldsymbol{y}}^k), \boldsymbol{\xi}_1^k, \boldsymbol{\xi}_2^k$ randomly;
4: compute $t_2^k(\boldsymbol{y}_2^k), \nabla t_2^k(\boldsymbol{y}_2^k)$ and $t_3^k(\boldsymbol{\theta}^k), \nabla t_3^k(\boldsymbol{\theta}^k)$;
5: update y_1^{k+1} and $\boldsymbol{y}_2^{k+1}$ by Equations (14) and (15);
6: compute $\nabla t_1^k(y_1^k)$ and $\nabla t_2^k(\boldsymbol{y}_2^k)$;
7: update $\boldsymbol{\theta}^{k+1}$ by Equation (13);
8: **end for**

We analyze the convergence of SCSC based on the following assumption.

Assumption 3. (Unbiased Oracle). *We assume that the sampling oracle satisfies that for each $k \in [K]$,*

(I) $\phi_{\widehat{\boldsymbol{x}}^k}^0(\boldsymbol{\theta}) = \phi_{\widehat{\boldsymbol{x}}}^{(0)}(\boldsymbol{\theta})$, $\phi_{\boldsymbol{\xi}_1^k}^1(\boldsymbol{\theta}) = \phi_{\boldsymbol{\xi}_1}^{(1)}(\boldsymbol{\theta})$, *and* $\phi_{\boldsymbol{\xi}_2^k}^2(\boldsymbol{\theta}) = \phi_{\boldsymbol{\xi}_2}^{(2)}(\boldsymbol{\theta})$;

(II) $\mathbb{E}_{\widehat{\boldsymbol{x}}^k, \widehat{\boldsymbol{y}}^k, \boldsymbol{\xi}_1^k, \boldsymbol{\xi}_2^k}\left[\nabla t_1^k(y_1)\nabla t_2^k(\boldsymbol{y}_2)\nabla t_3^k(\boldsymbol{\theta})\right] = \mathbb{E}_{\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2}\left[\nabla t_1(y_1)\nabla t_2(\boldsymbol{y}_2)\nabla t_3(\boldsymbol{\theta})\right]$.

Assumption 3 is standard in stochastic compositional optimization (T. Chen et al., 2021), and is analogous to the unbiasedness assumption for stochastic non-compositional problems. According to T. Chen et al. (2021), we have the following convergence results.

Theorem 4. (Convergence of SCSC for Problem (SCO)). *Under Assumptions 2 and 3, if we choose the step-sizes as $\alpha_k = \frac{2\beta_k}{A_1^2 + A_2^2} = \frac{1}{\sqrt{K}}$, then*

(I) *the iterates $\{\boldsymbol{\theta}^k\}$ of the Algorithm 1 satisfy:*

$$\frac{\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla F(\boldsymbol{\theta}^k)\|^2\right]}{K} \leq \frac{C_{\text{const}}}{\sqrt{K}},$$

where $A_1, A_2, C_{\text{const}}$ are constants that depend on the initial setting of the algorithm and constants $C_1, C_2, C_3, S_1, S_2, S_3$;

(II) *to obtain an ε-stationary point of function F, i.e., a point $\widehat{\boldsymbol{\theta}}$ satisfying $\mathbb{E}\left[\|\nabla F(\widehat{\boldsymbol{\theta}})\|^2\right] \leq \varepsilon^2$, the number of iterations required, the sample complexity, and the gradient complexity of functions t_1, t_2, and t_3 are all at the order of $\mathcal{O}(\varepsilon^{-4})$, and this result is nearly optimal.*

The proof of Theorem 4 is provided in E-companion EC.1.9. Theorem 4 also shows that the convergence rate of SCSC is $\mathcal{O}(k^{-1/2})$, which is on the same order as SGD's rate for the stochastic non-compositional nonconvex problems.

6 Applications and Numerical Results

In this section, we validate the efficiency of the proposed approach across three applications: the newsvendor problem (Example 1) in Section 6.1, the inventory substitution problem (Example 2) in Section 6.2, and the portfolio selection problem with real data (Example 3) in Section 6.3. Additionally, we demonstrate the interpretability of the SRF decision rule for the portfolio problem in Section 6.3.2.

In all experiments, we train the decision rule by solving the soft-constrained Causal-SDRO model. Following Remark 3, we treat the penalty coefficient λ as a tunable hyperparameter instead of the radius $\bar{\rho}$. We take ℓ_p-norm as transportation cost for the causal Sinkhorn discrepancy, where $p \in \{1, 2\}$. For brevity, we denote the resulting model as *p-Causal-SDRO*, and solve it using the SCSC algorithm described in Section 5.3. We take $T = 20, D(t) = \lceil \log_2 d_x \rceil + 1$ for each $t \in [T]$ for the SRF decision rule. The experiments are coded in Python 3.8 and conducted on a personal computer equipped with an Intel Core i9-13900HX CPU, 32 GB of RAM, and an Nvidia GeForce RTX 4060 GPU. All GPU computations are performed using PyTorch 2.0.1 (utilizing CUDA 11.8).

For benchmark comparison, we also examine the performance of a two-layer neural network (2NN) decision rule, which is a learning-based parametric decision rule but lacks interpretability. Let m be the dimension of the hidden layer in the 2NN. Then, the 2NN decision rule $f_{\boldsymbol{\theta}}^{\text{2NN}} : \mathcal{X} \to \mathcal{Z}$, parametrized by $\boldsymbol{\theta} \in \Theta$ that collects all parameters $\left\{\boldsymbol{a}^k \in \mathbb{R}^m, \boldsymbol{b}^k \in \mathbb{R}^m, \boldsymbol{w}_i^k \in \mathbb{R}^{d_x}\right\}$ for all $i \in [m]$ and $k \in [d_z]$, is given by

$$\left[f_{\boldsymbol{\theta}}^{\text{2NN}}(\boldsymbol{x})\right]_k := \frac{1}{m}\sum_{i=1}^{m} a_i^k \cdot \text{ReLu}\left((\boldsymbol{w}_i^k)^\top \boldsymbol{x} + b_i^k\right), \quad \forall k \in [d_z], \tag{2NN}$$

where $\text{ReLu} : \mathbb{R} \to \mathbb{R}_+$ represents the ReLu activation function, while a_i^k and b_i^k represent the i-th element in vectors $\boldsymbol{a}^k$ and $\boldsymbol{b}^k$, respectively, for any $i \in [m]$. This decision rule requires training a total of $m \cdot d_z(d_x + 2)$ parameters. According to Ma et al. (2018), over-parameterized 2NNs can effectively approximate optimal policies within the Barron space with dimension-independent convergence rates. We take $m = 64 \times d_x$ for the 2NN decision rule to ensure its number of trainable parameters approximates that of SRF.

We evaluate the out-of-sample performance using the coefficient of prescriptiveness (Bertsimas and Kallus, 2020). Let $\mathcal{S}$ denote an independent test dataset sampled from the true joint

distribution, disjoint from the training data. We define the average out-of-sample loss as

$$\mathcal{H}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}} \Psi\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y}\Big).$$

where $|\mathcal{S}|$ represents the cardinality of set $\mathcal{S}$. Let $\mathcal{H}^*$ represent the oracle loss under perfect information, and then the coefficient of prescriptiveness is given by

$$\text{Prescriptiveness}(\boldsymbol{\theta}) = \left(1 - \frac{\mathcal{H}(\boldsymbol{\theta}) - \mathcal{H}^*}{\mathcal{H}(\boldsymbol{\theta}^{\text{ERM}}) - \mathcal{H}^*}\right) \times 100\%, \tag{16}$$

where $\boldsymbol{\theta}^{\text{ERM}}$ denotes the parameter trained by the empirical risk minimization (ERM) model, i.e.,

$$\boldsymbol{\theta}^{\text{ERM}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{N} \sum_{i=1}^{N} \Psi\Big(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}_i), \widehat{\boldsymbol{y}}_i\Big).$$

A higher value indicates a better out-of-sample performance of decision rules. In the experiments in Section 6.1 and 6.2, we set a testing dataset size of $|\mathcal{S}| = 10^5$.

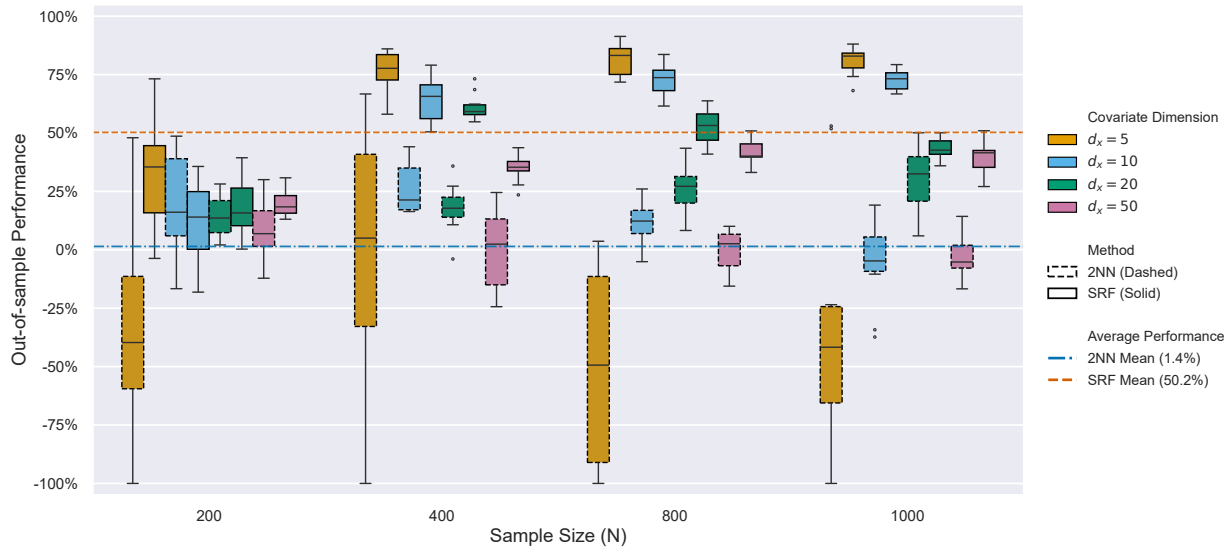
6.1 Feature-based Newsvendor Problem

In this subsection, we consider a feature-based newsvendor problem, adopting an experimental setup similar to that of J. Yang et al. (2022), where the demand $y \in \mathbb{R}_+$ depends on the covariate $\boldsymbol{x}$ in a nonlinear way:

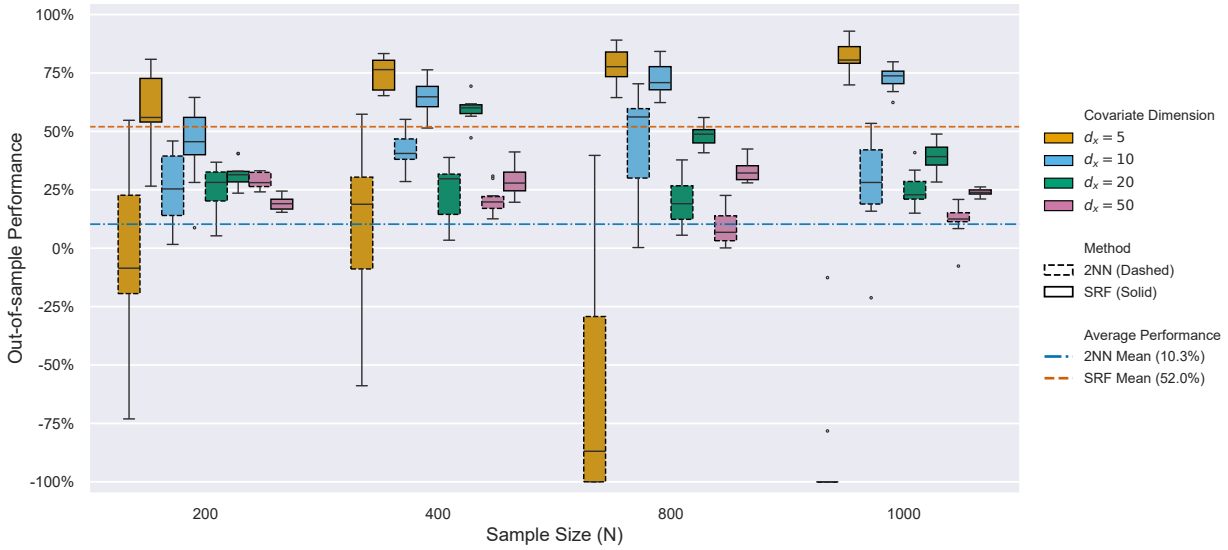
$$y = f_{\text{true}}(\boldsymbol{\beta}^\top \boldsymbol{x}) + \varsigma, \quad \text{where } f_{\text{true}}(\lambda) = c\Big[\sin(2\lambda) + 2\exp(-16\lambda^2) + 1\Big].$$

Here, $\varsigma \sim \mathcal{N}(0,1)$ represents an independent Gaussian noise, the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{d_x}$ is generated by taking each component sampled from the uniform distribution $\mathcal{U}([-0.1, 0.1])$, the covariate $\boldsymbol{x} \in \mathbb{R}^{d_x}$ is generated from a multivariate normal distribution with zero mean and covariance matrix Σ with $\Sigma_{ij} = 0.5^{|i-j|}$ for each $i, j \in [d_x]$, the constant $c = 1.7$. To ensure non-negativity, the demand y is simulated via the acceptance and rejection method. We conduct experiments across observed historical sample size $N \in \{200, 400, 800, 1000\}$, feature dimension $d_x \in \{5, 10, 20, 50\}$. The unit holding and stock-out costs are set to $h = 0.6$ and $b = 1.0$, respectively.

Figure 4 compares the out-of-sample performance across different sample sizes and feature dimensions. In these plots, the results of SRF are distinguished by boxes with solid borders. As illustrated, the proposed SRF decision rule outperforms the 2NN benchmark and the ERM baseline across nearly all tested instances. Under 2-Causal-SDRO, the SRF provides positive out-of-sample performance on all instances. Quantitatively, the SRF achieves average prescriptiveness scores of 50.2% (1-Causal-SDRO) and 52.0% (2-Causal-SDRO), marking a significant advantage over the 2NN, which yields only 1.4% and 10.3%, respectively. Beyond average performance, the box plots reveal that the SRF exhibits significantly lower variance (indicated by shorter interquartile ranges) compared to the 2NN, highlighting the stability of our approach. Notably, the SRF achieves its peak performance at $d_x = 5$ across all sample sizes, whereas the 2NN performs worst in this setting. These results show that the proposed SRF rule is highly effective for decision-making tasks, even when historical data is limited.



(a) 1-Causal-SDRO

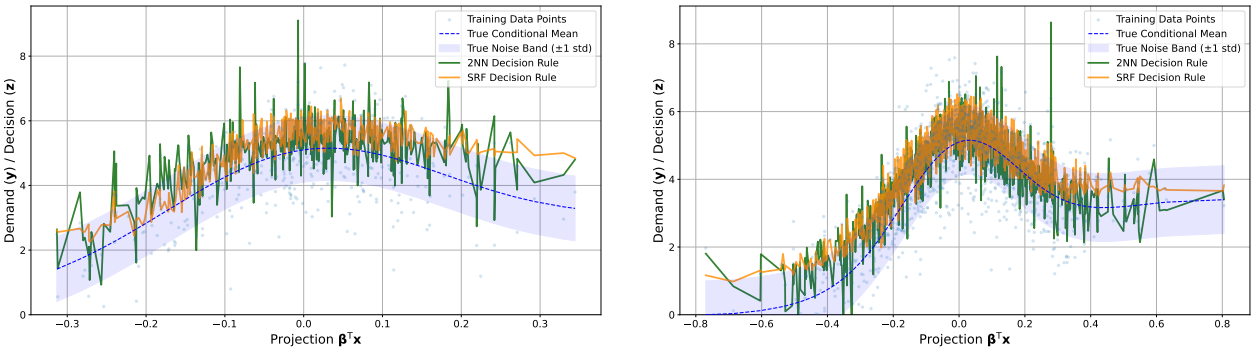


(b) 2-Causal-SDRO

Figure 4. Out-of-sample performance of the decision rules on the newsvendor problem

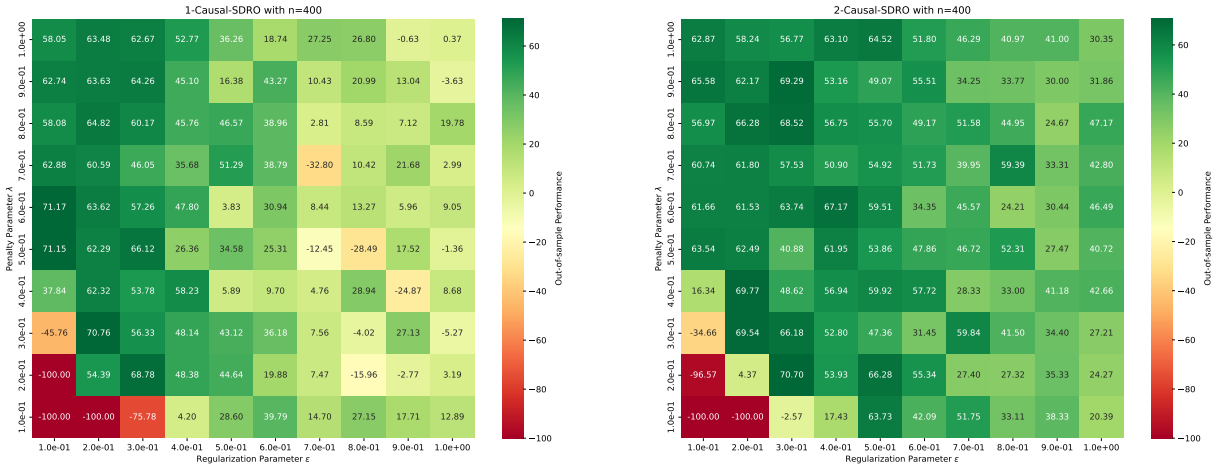
Figure 5 visualizes the fitted 2NN and SRF decision rules against the true conditional mean (blue dashed line) for sample size $N \in \{400, 1000\}$. The 2NN decision rule (green line) exhibits high variance, resulting in overfitting to the observed noise. In contrast, the SRF decision rule (orange line) yields a stable fit that captures the underlying shape of the true function well. Note that the SRF curve lies consistently above the conditional mean. This alignment correctly reflects that the unit holding cost is lower than the unit stock-out cost ($h < b$), thereby decision-makers prefer maintaining higher inventory levels to mitigate stock-out risks.

Figure 6 reports the out-of-sample performance of the proposed method for the Causal-SDRO



(a) $N = 400$ (b) $N = 1000$

Figure 5. True distribution vs. Trained decision rules for 2-Causal-SDRO ($d_x = 10$)



(a) 1-Causal-SDRO (b) 2-Causal-SDRO

Figure 6. Out-of-sample performance of the newsvendor problem with different parameters ($N = 400, d_x = 10$)

model across different parameter combinations, including penalty parameter λ, regularization parameter ϵ, and norm p, taking instances where $N = 400$ and $d_x = 10$ as examples. As shown in these plots, though both models achieve positive out-of-sample performance on almost all instances, the 2-Causal-SDRO illustrates a higher out-of-sample performance on most parameter combinations. These results indicate that performance improves by moderately increasing λ and decreasing ϵ. This is because a small λ leads to excessive conservatism, while a large λ reduces the model to ERM. Similarly, an insufficient ϵ fails to adequately characterize the continuity of the underlying distribution, while an excessive ϵ dilutes the correlation between covariates and uncertain parameters.

6.2 Feature-based Inventory Substitution Problem

In this subsection, we conduct a numerical study on the feature-based inventory substitution problem introduced in Example 2. This application serves as a representative two-stage contextual DRO problem.

The inventory substitution problem with a soft CSD constraint is equivalent to the following stochastic compositional optimization problem

$$\min_{\boldsymbol{\theta}\in\Theta} \quad F(\boldsymbol{\theta}) = \lambda\epsilon \cdot \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[t_1\left(\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[t_3'\left(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2\right)\right];\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1\right)\right];\widehat{\boldsymbol{x}}\right)\right]$$

where functions t_1 and t_2 are defined in (12), and the inner function t_3' involves the dual of the second-stage recourse problem, denoted by Ψ_{I}^*:

$$\left[t_3'(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\right]_i = \exp\left(\frac{\Psi_{\mathrm{I}}^*(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_i+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right), \quad \forall i\in[n_{\widehat{\boldsymbol{x}}}],$$

and

$$\Psi_{\mathrm{I}}^*(f(\boldsymbol{x}),\boldsymbol{y}) := \max_{\boldsymbol{\eta}\in\mathbb{R}^{d_z},\boldsymbol{\upsilon}\in\mathbb{R}^{d_y}} \left\{ \sum_{i=1}^{d_z}[f(\boldsymbol{x})]_i(\eta_i+c_i) + \sum_{j=1}^{d_y} y_j\upsilon_j \quad \middle| \quad \begin{array}{ll} \eta_i \le h_i, & \forall i\in[d_z], \\ \upsilon_j \le b_j, & \forall j\in[d_y], \\ \eta_i+\upsilon_j \le s_{i,j}, & \forall j\in\left\{i,i+1,\cdots,d_y\right\}, i\in[d_z] \end{array} \right\}.$$

Detailed derivations are provided in E-companion EC.4. We solve $\Psi_{\mathrm{I}}^*(f(\boldsymbol{x}),\boldsymbol{y})$ using the commercial solver Gurobi (version 12.0.1).

We examine a scenario with $d_z = d_y = 3$ products, varying feature dimensions $d_x \in \{3,5,8,10\}$ and sample sizes $N \in \{100,200,400,800\}$. Let the conditional demand distributions of the products be exponential and Gamma distributions parametrized by covariates:

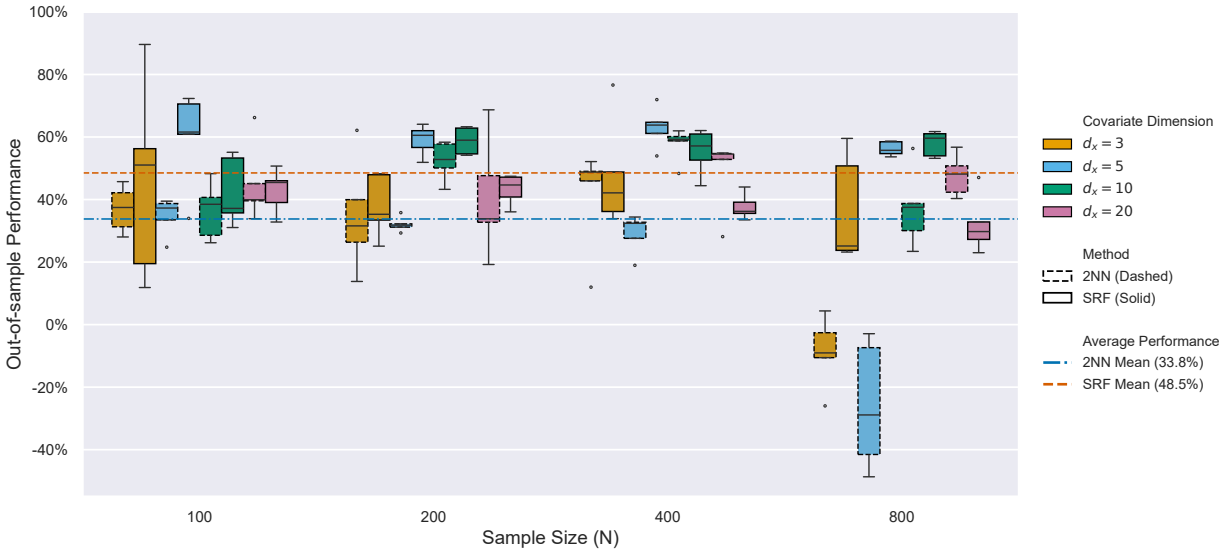
$$y_1 \mid \boldsymbol{x} \sim \mathrm{Exp}(e^{\boldsymbol{\beta}^\top\boldsymbol{x}}), \quad y_2 \mid \boldsymbol{x} \sim \mathrm{Gamma}(2,e^{\boldsymbol{\beta}^\top\boldsymbol{x}}), \quad y_3 \mid \boldsymbol{x} \sim \mathrm{Gamma}(4,e^{\boldsymbol{\beta}^\top\boldsymbol{x}}),$$

where $\boldsymbol{\beta}\in\mathbb{R}^{d_x}$ is sampled from $\mathcal{U}([-0.1,0.1])$, and $\boldsymbol{x}\in\mathbb{R}^{d_x}$ is constructed by the procedure in Section 6.1. We specify the cost parameters as

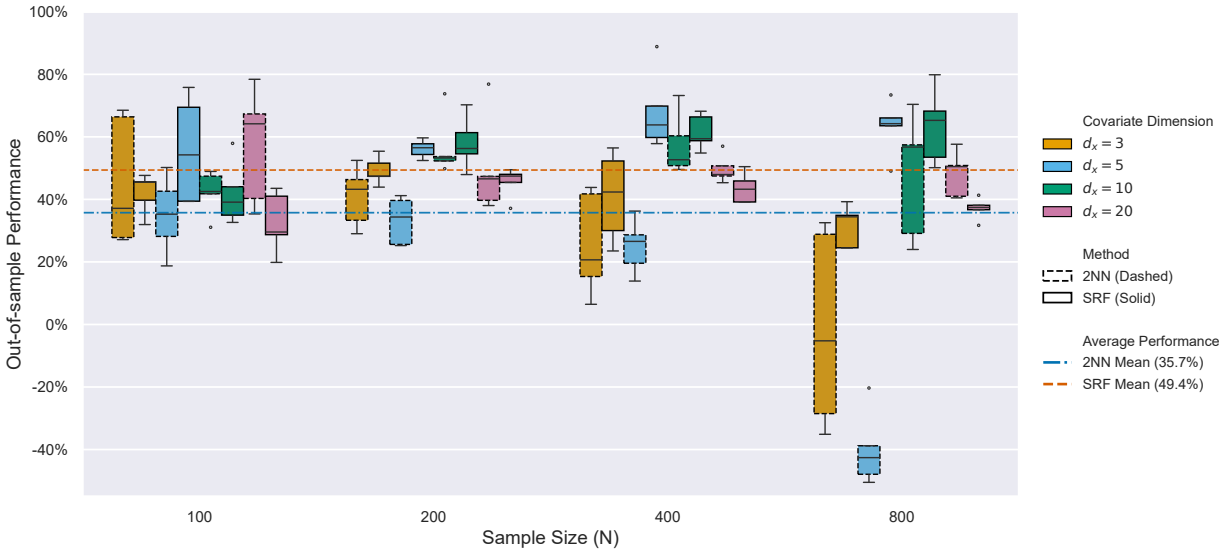
$$\boldsymbol{h} = \begin{bmatrix} 1 \\ 0.7 \\ 0.6 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1.8 \\ 1.6 \\ 1.2 \end{bmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} 0 & 1.7 & 2 \\ \infty & 0 & 1.5 \\ \infty & \infty & 0 \end{pmatrix},$$

and $\boldsymbol{c} = \boldsymbol{0}$. Note that in the substitution cost matrix $\boldsymbol{S}$, $S_{i,j} = \infty$ when $i > j$ as lower-quality products cannot substitute for higher-quality ones.

Figure 7 reports the out-of-sample performance of the proposed approach across varying p, N, and d_x. The SRF decision rule demonstrates superior efficacy, achieving average prescriptiveness scores of 48.5% (1-Causal-SDRO) and 49.4% (2-Causal-SDRO), while consistently maintaining positive out-of-sample performance across all instances. In comparison, the 2NN benchmark



(a) 1-Causal-SDRO



(b) 2-Causal-SDRO

Figure 7. Out-of-sample performance of the decision rules on the inventory substitution problem

yields only 33.8% (1-Causal-SDRO) and 35.7% (2-Causal-SDRO). These results validate that our proposed approach generalizes effectively to two-stage contextual DRO problems.

Regarding parameter sensitivity, the Causal-SDRO model exhibits trends consistent with the newsvendor problem. Please see E-companion EC.5 for details.

6.3 Data-driven Portfolio Selection Problem

In this subsection, we consider a data-driven portfolio selection problem using real-world market data adapted from Nguyen et al. (2025). This dataset contains the historical asset returns of S&P500 constituents from January 1, 2017, to March 31, 2023. All of the selected 399 assets have been in the S&P500 since 2010. The features include five publicly available market indices: (i) Volatility Index (VIX), (ii) 10-year Treasury Yield Index (TNX), (iii) Crude Oil Index (CL=F), (iv) S&P 500 (GSPC), and (v) Dow Jones Index (DJI) to construct the covariate $\boldsymbol{x} \in \mathbb{R}^5$. These features capture the macro market environment and economic conditions, and it is reasonable to assume they are exogenous to the historical returns of individual assets given historical values, thereby satisfying the causal structure in our model.

In the following, Section 6.3.1 shows the performance of our approach on the portfolio problem, and Section 6.3.2 shows its intrinsic interpretability from an empirical perspective.

6.3.1 Performance of the Proposed Approach

We implement a rolling-horizon experiment to validate the performance of the proposed approach. For the first trade day of each month between January 2021 and December 2022, we randomly sample $d_y = 50$ assets from the universe to form the stock pool. We make portfolio decisions based on an empirical distribution formed by the prior two-year window data on the covariates and asset returns, targeting the best return over the subsequent 60-day holding period.

Let $r_{i,j}$ denote the return of asset i on day j within the testing horizon ($j \in [60]$). We compare the following methods:

(I) The post-hoc testing (PT) model. This benchmark provides a theoretically optimal objective value (i.e., $\mathcal{H}^*$) when the information of the future 60 days is completely known:

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \quad \frac{1}{60}\sum_{j=1}^{60}\Bigg[-\omega\cdot\sum_{i=1}^{d_y} r_{i,j}z_i + \Bigg(\sum_{i=1}^{d_y} r_{i,j}z_i - z_0\Bigg)^2\Bigg].$$

(II) The equal-weighted (EW) model. This model provides equal weights to each selected asset, i.e., the investment amount for all assets is $1/d_y$.

(III) The unconditional mean-variance (MV) model. This model is a traditional portfolio model that makes a decision based on the empirical distribution $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}}$:

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \quad \mathbb{E}_{\boldsymbol{y}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}}}\Big[\Psi_{\mathrm{P}}(\boldsymbol{z},\boldsymbol{y})\Big],$$

where the portfolio loss function Ψ_{P} is defined in Example 3.

(IV) The conditional mean-variance (CMV) model. This is a contextual stochastic optimization (CSO) approach that trains a decision rule $f \in \mathcal{F}$ to minimize the empirical conditional risk:

$$\inf_{f\in\mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\widehat{\mathbb{P}}}\Big[\Psi_{\mathrm{P}}(f(\boldsymbol{x}),\boldsymbol{y})\Big].$$

(V) Our conditional p-Causal-SDRO mean-variance (p-DRO-CMV) model shown in Example 3.

Let $\widetilde{r}_t$ denote the realized daily return of the portfolio on day t within the 60-day holding period ($t \in [60]$). We compute and report the following performance metrics: (1) The mean return $\text{Mean}(\{\widetilde{r}_t\}_{t\in[60]})$, and the standard deviation of return $\text{stdDev}(\{\widetilde{r}_t\}_{t\in[60]})$. (2) The portfolio loss value shown in Example 3. (3) The annualized Sharpe ratio $\sqrt{252}\times\text{Mean}(\{\widetilde{r}_t\}_{t\in[60]})/\text{stdDev}(\{\widetilde{r}_t\}_{t\in[60]})$. (4) The Conditional Value-at-Risk (CVaR) at the 5% level, which quantifies the expected loss in the worst-case scenarios. (5) The out-of-sample performance for all decision rules following Equation (16).

We employ the proposed SRF as the parametric decision rule for both the CMV and Causal-SDRO models. Table 1 shows the experimental results across varying risk aversion levels $\omega \in \{1,3,5,7,9\}$, representing different tradeoffs between portfolio mean return and variance. As illustrated, the Causal-SDRO model achieves the lowest average loss among all implementable baselines (excluding the PT oracle) for the four cases where $\omega \in \{1,5,7,9\}$. The only exception occurs at $\omega = 3$, where the CMV yields a slightly smaller average loss and yet Causal-SDRO delivers the highest mean portfolio return. Across almost all values of ω, both CMV and Causal-SDRO outperform the unconditional MV model, underscoring the value of incorporating covariates for decision-making. In practice, although decision-makers cannot obtain all covariates, our results demonstrate that our approach maintains robust performance even when the observed features may be imperfect.

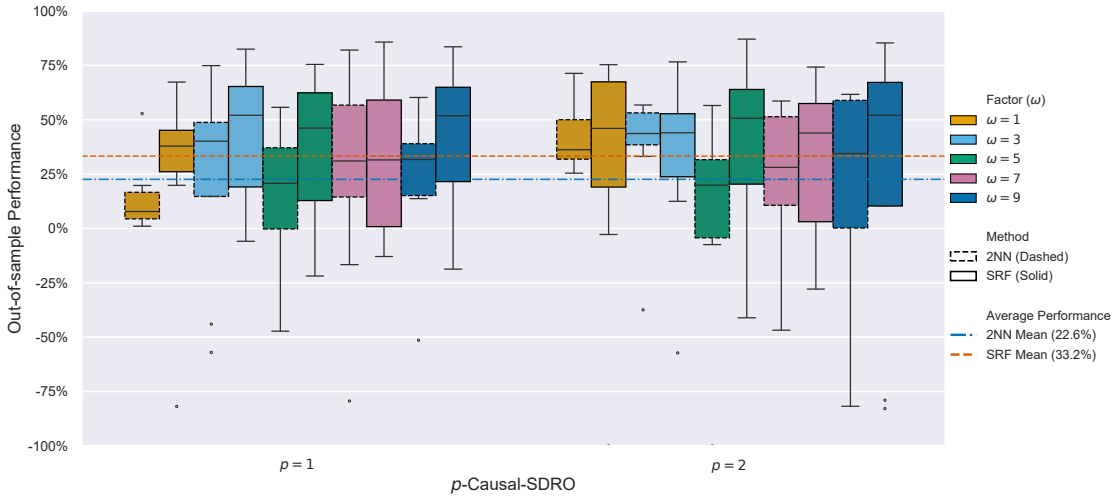


Figure 8. Out-of-sample performance of the decision rules on the portfolio problem

Figure 8 compares the out-of-sample performance of decision rules on the portfolio selection problem. As illustrated, the SRF decision rule outperforms the 2NN benchmark, achieving an average score of 33.2% compared to 22.6%. Furthermore, the SRF consistently maintains a higher median performance across all tested values of p and ω.

Table 1: Average performance for mean-variance methods

ω	Methods	Average Loss ↓	Sharpe ↑	Mean ↑	stdDev ↓	CVaR ↓
1	PT	0.468	4.125	0.176	0.784	1.971
	EW	1.425	1.163	0.068	1.195	2.450
	MV	1.009	**1.578**	0.072	1.008	1.971
	CMV	0.984	1.156	0.061	1.003	2.108
	1-DRO-CMV	0.950	1.352	0.067	0.985	2.004
	2-DRO-CMV	**0.925**	1.559	**0.081**	**0.978**	**1.965**
3	PT	0.016	5.189	0.263	0.883	1.543
	EW	1.288	1.163	0.068	1.195	2.450
	MV	0.936	**1.512**	0.069	1.037	**2.033**
	CMV	**0.865**	1.367	0.069	**1.016**	2.089
	1-DRO-CMV	0.929	1.387	**0.072**	1.044	2.109
	2-DRO-CMV	0.895	1.260	0.065	1.022	2.058
5	PT	-0.569	5.383	0.315	0.991	1.705
	EW	1.152	1.163	0.068	1.195	2.450
	MV	0.890	**1.445**	0.066	1.073	**2.103**
	CMV	0.836	1.293	0.063	**1.045**	2.132
	1-DRO-CMV	**0.800**	1.353	**0.074**	1.060	2.127
	2-DRO-CMV	0.851	1.107	0.059	1.053	2.147
7	PT	-1.234	5.360	0.345	1.075	1.818
	EW	1.016	1.163	**0.068**	1.195	2.450
	MV	0.874	**1.367**	0.061	1.111	**2.176**
	CMV	0.895	0.917	0.046	1.087	2.258
	1-DRO-CMV	**0.775**	1.192	0.062	1.085	2.222
	2-DRO-CMV	0.823	1.203	0.058	**1.084**	2.310
9	PT	-1.943	5.261	0.361	1.135	1.897
	EW	0.879	1.163	0.068	1.195	2.450
	MV	0.861	**1.317**	0.058	1.147	**2.236**
	CMV	0.752	1.141	0.052	**1.081**	2.265
	1-DRO-CMV	0.698	1.115	0.062	1.102	2.337
	2-DRO-CMV	**0.618**	1.253	**0.074**	1.113	2.330

Note. Bold values represent the best performance except for the PT model for each value of ω.

6.3.2 Interpretability of the Proposed Approach

To further understand the intrinsic interpretability, we examine a simple Soft Regression Tree (SRT) with three layers, trained on the mean-variance portfolio problem with $\omega = 5$. This SRT is trained using covariates and asset returns from the preceding two-year rolling window.

Figure 9 illustrates the structure of the trained SRT, where normalized feature weights are displayed at each internal node. Consider two input covariates $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^5$. The decision of $\boldsymbol{x}_1$ becomes π_8 with weight probability (approximately) 1. The model maps $\boldsymbol{x}_2$ to decisions π_7 and π_4 with weight probabilities 0.996 and 0.004, respectively. These results confirm that the final decision is dominated by a few high-probability routes.

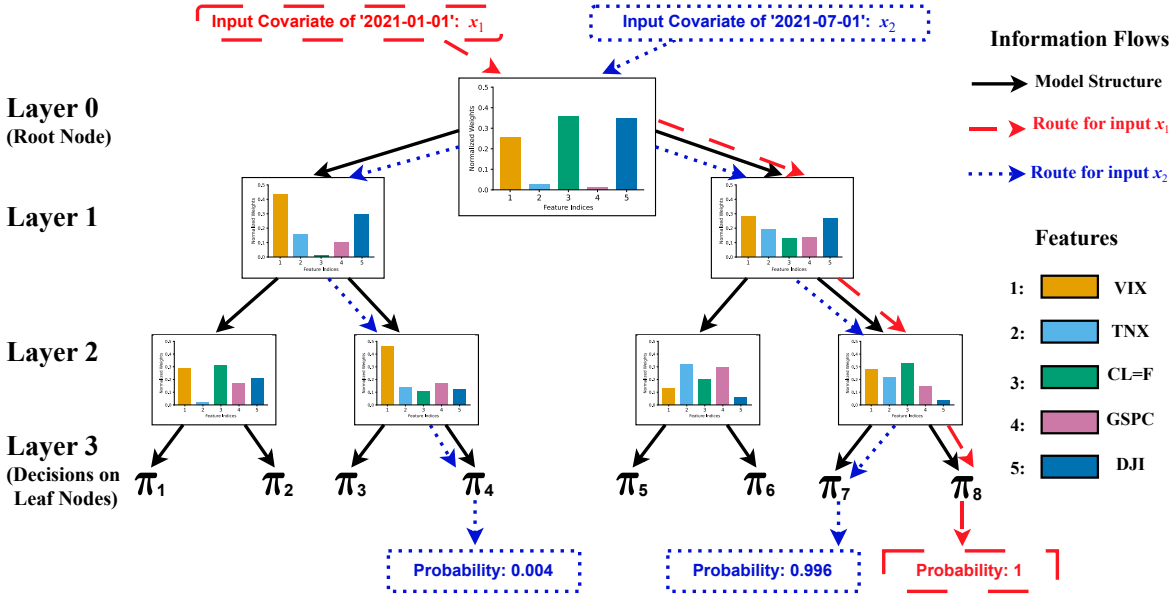


Figure 9. The structure of the trained SRT with three layers. Here, the solid black lines denote the model structure, while the red and blue dashed lines illustrate the decision routes and their corresponding selection probabilities for two input covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

We next demonstrate the intrinsic interpretability of this SRT from both global and local perspectives. The global interpretation provides a holistic view of the model by quantifying the contribution of each feature to the overall decision-making process, while the local shows how an individual decision is derived given a specific covariate (Dwivedi et al., 2023). In E-companion EC.3, we introduce a global feature importance measure and a local feature attribution measure for SRF based on differentiability. In the following, we analyze the SRT using these proposed interpretation measures and compare them with traditional posh-hoc explanation methods.

Globally, Figure 10 visualizes the relative feature importance of this SRT using the proposed interpretation measure in E-companion EC.3 and a perturbation-based measure (Hastie et al., 2009). The difference between the two measures is that our measure shows the feature importance in an intrinsic way that depends only on the derivatives of SRF and avoids post-hoc perturbation analyses. As illustrated, these measures are highly correlated with a Pearson correlation coefficient of 0.820. This result confirms that the intrinsic interpretation of SRF is consistent with established global post-hoc explanations.

Locally, Figure 11 visualizes the feature attributions derived from intrinsic interpretation (the proposed EIG measure in E-companion EC.3) and a traditional post-hoc explanation (SHAP mea-

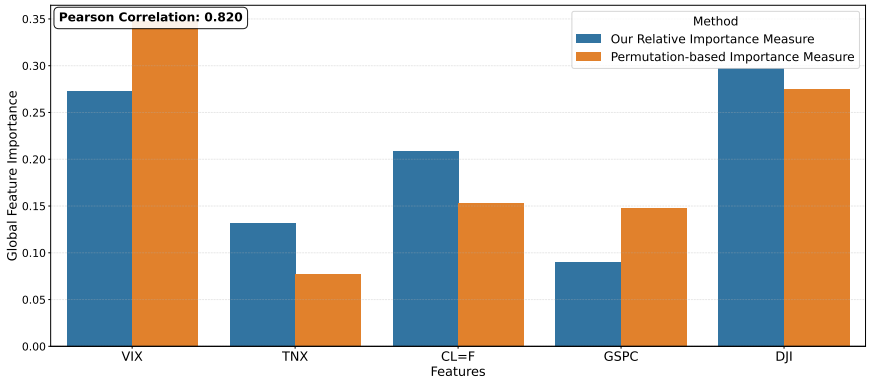
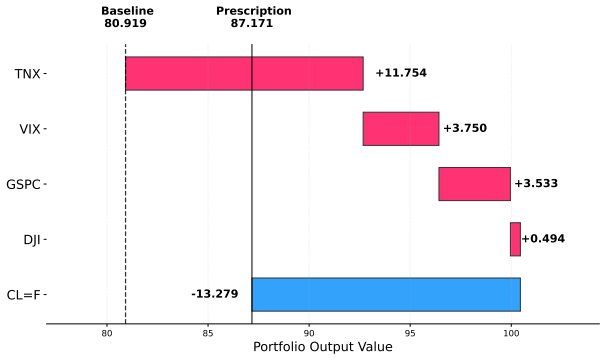
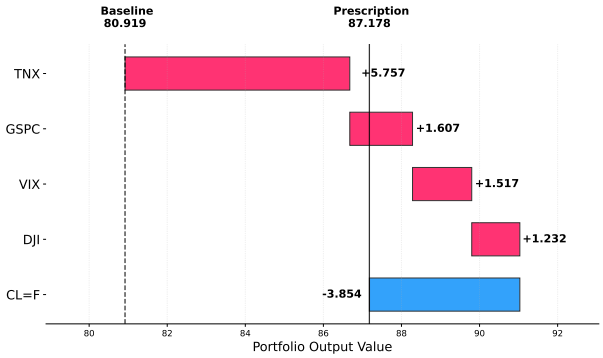


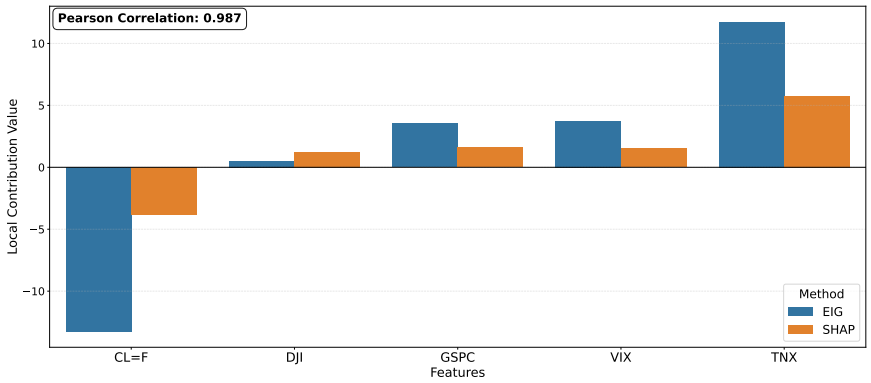
Figure 10. Global feature importance comparison for the trained SRT



(a) Intrinsic interpretation by EIG measure



(b) Post-hoc explanation by SHAP measure



(c) Comparison of the two measures

Figure 11. Local feature attribution comparison for the trained SRT

sure, Lundberg et al., 2020) for this SRT given a specific covariate. Figures 11(a) and 11(b) display the waterfall plots for EIG and SHAP measures, respectively, where red and blue bars indicate positive and negative contributions, respectively. The prescription and baseline value are calculated by $\sum_{k=1}^{d_z}[f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x})]_k$ and $\sum_{i=1}^{N}\sum_{k=1}^{d_z}[f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x}^i)]_k/N$, respectively. Figure 11(c) directly compares the EIG and SHAP values across all features. As illustrated, the high correlation coefficient of

0.987 between these measures reflects the consistency between our local intrinsic interpretation and the local post-hoc explanation.

7 Conclusion

In this paper, we consider the causal and continuous structure of the underlying distribution for contextual DRO. We develop a new framework termed Causal-SDRO, which builds the ambiguity set using the entropy-regularized causal Wasserstein distance, excluding discrete and causally implausible distributions. To maintain interpretability and computational tractability, we develop a Soft Regression Forest (SRF) decision rule. The SRF possesses universal approximation capabilities to approach optimal policies within arbitrary measurable function spaces and maintains the interpretability of tree-based models, enabling intrinsic interpretation from both global and local perspectives. To solve the resulting model, we present a gradient-based algorithm with a convergence rate at the order of $\mathcal{O}(\varepsilon^{-4})$, which is nearly optimal. The proposed approach empirically outperforms baselines in both decision out-of-sample performance and interpretability.

There are several promising directions to explore. Theoretically, it is interesting to incorporate prior information to design ambiguity sets that contain more plausible distributions for contextual DRO. Moreover, it is important to develop accelerated and posterior update algorithms for SRF decision rule-based optimization. Finally, it is promising to apply our proposed framework for practical applications that require safety, robustness, and interpretability.

References

Aghaei, S., Gómez, A., & Vayanos, P. (2025). Strong optimal classification trees. *Operations Research*, *73*(4), 2223–2241.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Azizian, W., Iutzeler, F., & Malick, J. (2023a). Exact generalization guarantees for (regularized) Wasserstein distributionally robust models. *Advances in Neural Information Processing Systems*, *36*, 14584–14596.

Azizian, W., Iutzeler, F., & Malick, J. (2023b). Regularization for Wasserstein distributionally robust optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, *29*, 33.

Ban, G.-Y., & Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, *67*(1), 90–108.

Bennouna, A., & Van Parys, B. (2022). Holistic robust data-driven decisions. *arXiv preprint arXiv: 2207.09560*.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., & Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, *59*(2), 341–357.

Bertsimas, D., Delarue, A., Jaillet, P., & Martin, S. (2019). The price of interpretability. *arXiv preprint arXiv:1907.03419*.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039–1082.

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044.

Bertsimas, D., & Koduri, N. (2022). Data-driven optimization: A reproducing kernel Hilbert space approach. *Operations Research*, *70*(1), 454–471.

Bertsimas, D., & Stellato, B. (2021). The voice of optimization. *Machine Learning*, *110*(2), 249–277.

Birrell, J., & Ebrahimi, R. (2025). Optimal transport regularized divergences: Application to adversarial robustness. *SIAM Journal on Mathematics of Data Science*, *7*(4), 1801–1827.

Blanchet, J., Kuhn, D., Li, J., & Taskesen, B. (2023). Unifying distributionally robust optimization via optimal transport theory. *arXiv preprint arXiv:2308.05414*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Cescon, R., Martin, A., & Ferrari-Trecate, G. (2025). Data-driven distributionally robust control based on Sinkhorn ambiguity sets. *arXiv preprint arXiv:2503.20703*.

Chen, R., & Paschalidis, I. (2019). Selecting optimal decisions via distributionally robust nearest-neighbor regression. *Advances in Neural Information Processing Systems*, *32*.

Chen, T., Sun, Y., & Yin, W. (2021). Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, *69*, 4937–4948.

Chen, X., & Gao, X. (2019). Stochastic optimization with decisions truncated by positively dependent random variables. *Operations Research*, *67*(5), 1321–1327.

Chenreddy, A. R., Bandi, N., & Delage, E. (2022). Data-driven conditional robust optimization. *Advances in Neural Information Processing Systems*, *35*, 9525–9537.

Dapogny, C., Iutzeler, F., Meda, A., & Thibert, B. (2023). Entropy-regularized Wasserstein distributionally robust shape and topology optimization. *Structural and Multidisciplinary Optimization*, *66*(3), 42.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1–33.

Elmachtoub, A. N., & Grigas, P. (2022). Smart "predict, then optimize". *Management Science*, *68*(1), 9–26.

Elmachtoub, A. N., Liang, J. C. N., & McNellis, R. (2020). Decision trees for decision-making under the predict-then-optimize framework. *International Conference on Machine Learning*, 2858–2867.

Esteban-Pérez, A., & Morales, J. M. (2022). Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming*, *195*(1), 1069–1105.

Forel, A., Parmentier, A., & Vidal, T. (2023). Explainable data-driven optimization: From context to decision and back again. *International Conference on Machine Learning*, 10170–10187.

Frosst, N., & Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

Fu, M., Li, X., & Zhang, L. (2024). Distributionally robust newsvendor under stochastic dominance with a feature-based application. *Manufacturing & Service Operations Management*, *26*(5), 1962–1977.

Ghadimi, S., Lan, G., & Zhang, H. (2016). Mini-batch stochastic approximation methods for non-convex stochastic composite optimization. *Mathematical Programming*, *155*(1), 267–305.

Han, J., Hu, M., & Shen, G. (2025). Deep neural newsvendor. *Management Science*.

Hastie, T., Tibshirani, R., Friedman, J., et al. (2009). The elements of statistical learning.

Hu, Y., Chen, X., & He, N. (2020). Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, *30*(3), 2103–2133.

Hu, Y., Wang, J., Xie, Y., Krause, A., & Kuhn, D. (2023). Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, *36*, 78412–78434.

Jiang, G., & Mao, T. (2025). Sinkhorn distributionally robust conditional quantile prediction with fixed design. *Entropy*, *27*(6), 557.

Kallus, N., & Mao, X. (2023). Stochastic optimization forests. *Management Science*, *69*(4), 1975–1994.

Kannan, R., Bayraksan, G., & Luedtke, J. R. (2024). Residuals-based distributionally robust optimization with covariate information. *Mathematical Programming*, *207*(1), 369–425.

Liu, F., Chen, Z., Wang, R., & Wang, S. (2024). Newsvendor under mean-variance ambiguity and misspecification. *arXiv preprint arXiv:2405.07008*.

Liu, Z., Xu, Z., Liu, F., Gao, R., & Li, S. (2025). Neural decision rule for constrained contextual stochastic optimization. *NeurIPS 2025 Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making*.

Liyanage, L. H., & Shanthikumar, J. G. (2005). A practical inventory control policy using operational statistics. *Operations Research Letters*, *33*(4), 341–348.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Ma, C., Wu, L., & E, W. (2018). A priori estimates of the population risk for two-layer neural networks. *arXiv preprint arXiv:1810.06397*.

Nguyen, V. A., Zhang, F., Blanchet, J., Delage, E., & Ye, Y. (2020). Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems*, *33*, 15232–15242.

Nguyen, V. A., Zhang, F., Wang, S., Blanchet, J., Delage, E., & Ye, Y. (2025). Robustifying conditional portfolio decisions via optimal transport. *Operations Research*, *73*(5), 2801–2829.

Notz, P. M., & Pibernik, R. (2024). Explainable subgradient tree boosting for prescriptive analytics in operations management. *European Journal of Operational Research*, *312*(3), 1119–1133.

Oroojlooyjadid, A., Snyder, L. V., & Takáč, M. (2020). Applying deep learning to the newsvendor problem. *IISE Transactions*, *52*(4), 444–463.

Ouasfi, A., Jena, S., Marchand, E., & Boukhayma, A. (2025). Toward robust neural reconstruction from sparse point sets. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6552–6562.

Perakis, G., Sim, M., Tang, Q., & Xiong, P. (2023). Robust pricing and production with information partitioning and adaptation. *Management Science*, *69*(3), 1398–1419.

Poursoltani, M., Delage, E., & Georghiou, A. (2023). Robust data-driven prescriptiveness optimization. *arXiv preprint arXiv:2306.05937*.

Qi, M., Cao, Y., & Shen, Z.-J. (2022). Distributionally robust conditional quantile prediction with fixed design. *Management Science*, *68*(3), 1639–1658.

Qi, M., Grigas, P., & Shen, Z.-J. (2025). Integrated conditional estimation-optimization. *Operations Research*.

Qi, M., Shi, Y., Qi, Y., Ma, C., Yuan, R., Wu, D., & Shen, Z.-J. (2023). A practical end-to-end inventory management model with deep learning. *Management Science*, *69*(2), 759–773.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E., & Vidal, T. (2025). A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, *320*(2), 271–289.

Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2021). *Lectures on stochastic programming: Modeling and theory*. SIAM.

Shen, Y., Xu, P., & Zavlanos, M. M. (2023). Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *arXiv preprint arXiv:2309.08748*.

Sim, M., Tang, Q., Zhou, M., & Zhu, T. (2025). The analytics of robust satisficing: Predict, optimize, satisfice, then fortify. *Operations Research*, *73*(5), 2708–2728.

Song, J., He, N., Ding, L., & Zhao, C. (2024). Provably convergent policy optimization via metric-aware trust region methods. *Transactions on Machine Learning Research*.

Srivastava, P. R., Wang, Y., Hanasusanto, G. A., & Ho, C. P. (2021). On data-driven prescriptive analytics with side information: A regularized Nadaraya-Watson approach. *arXiv preprint arXiv:2110.04855*.

Wang, J., Gao, R., & Xie, Y. (2024). Non-convex robust hypothesis testing using Sinkhorn uncertainty sets. *arXiv preprint arXiv:2403.14822*.

Wang, J., Gao, R., & Xie, Y. (2025). Sinkhorn distributionally robust optimization. *Operations Research*.

Wang, J., & Xie, Y. (2022). A data-driven approach to robust hypothesis testing using Sinkhorn uncertainty sets. *arXiv preprint arXiv:2202.04258*.

Wang, M., Fang, E. X., & Liu, H. (2017). Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, *161*(1), 419–449.

Wang, T., Chen, N., & Wang, C. (2021). Distributionally robust prescriptive analytics with Wasserstein distance. *arXiv preprint arXiv:2106.05724*.

Yang, J., Zhang, L., Chen, N., Gao, R., & Hu, M. (2022). Decision-making with side information: A causal transport robust approach. *Optimization Online*.

Yang, S.-B., & Li, Z. (2023). Distributionally robust chance-constrained optimization with Sinkhorn ambiguity set. *AIChE Journal*, *69*(10), e18177.

Zhang, L., Yang, J., & Gao, R. (2024). Optimal robust policy for feature-based newsvendor. *Management Science*, *70*(4), 2315–2329.

Zhu, T., Xie, J., & Sim, M. (2022). Joint estimation and robustness optimization. *Management Science*, *68*(3), 1659–1677.

E-Companion

EC.1 Technical Analyses and Proofs

EC.1.1 Analysis for Condition 1

We present the following sufficient conditions to verify whether the Condition 1 holds.

Proposition EC.1. *Condition 1 holds if there exist $\lambda > 0$ and a constant $\alpha \in [0,1)$ so that for $\widehat{\mathbb{P}}\otimes\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}$-almost every $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})$, the following condition is satisfied*

$$\Psi(f(\boldsymbol{x}),\boldsymbol{y}) \le \alpha\cdot\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) + M(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}), \tag{EC.1}$$

where $M(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})$ is a measurable function, and satisfies $\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{P}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}[e^{M(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})/(\lambda\epsilon)}] < \infty$ for $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$-almost every $\widehat{\boldsymbol{x}}$.

Proof of Proposition EC.1. In Equation (3), we set $\boldsymbol{\xi}_1 = \boldsymbol{x}-\widehat{\boldsymbol{x}}$ and $\boldsymbol{\xi}_2 = \boldsymbol{y}-\widehat{\boldsymbol{y}}$. When relation (EC.1) holds, for $\widehat{\mathbb{P}}\otimes\nu_{\mathcal{X}}$-almost every $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right]\\
&\quad= \int_{\boldsymbol{y}\sim\nu_{\mathcal{Y}}}\left[\frac{e^{-\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^p/\epsilon}}{\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}-\widehat{\boldsymbol{y}}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})}\cdot\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\boldsymbol{y})}{\lambda\epsilon}\right)\right]\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\\
&\quad\le \int_{\boldsymbol{y}\sim\nu_{\mathcal{Y}}}\left[\frac{e^{-\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^p/\epsilon}}{\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}-\widehat{\boldsymbol{y}}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})}\cdot\exp\left(\frac{\alpha\cdot\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))+M(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}{\lambda\epsilon}\right)\right]\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\\
&\quad= \frac{e^{\alpha\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|^p/\epsilon+M(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})/\lambda\epsilon}}{\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}-\widehat{\boldsymbol{y}}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})}\cdot\int_{\boldsymbol{y}\sim\nu_{\mathcal{Y}}}\exp\left(-\frac{(1-\alpha)\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^p}{\epsilon}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\\
&\quad< \infty,
\end{aligned}$$

where the first inequality is due to (EC.1), and the second inequality is due to the assumption in Proposition EC.1 and Assumptions 1(III) and (IV), ensuring that the terms $e^{\alpha\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|^p/\epsilon} < \infty$, $e^{M(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})/\lambda\epsilon} < \infty$, the constant $\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}-\widehat{\boldsymbol{y}}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u}) < \infty$, and $\int_{\boldsymbol{y}\sim\nu_{\mathcal{Y}}}\exp\left(-\frac{(1-\alpha)\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^p}{\epsilon}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y}) < \infty$ as $1-\alpha > 0$. Hence, we have

$$g'(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda\epsilon\log\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right]\right] < \infty,$$

and it follows that

$$\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\exp\left(\frac{g'(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon}\right)\right] < \infty,$$

for $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$-almost every $\widehat{\boldsymbol{x}}$. □

EC.1.2 Proof of Theorem 1 in Section 3.1

To prove the strong duality in Theorem 1, we first develop the following Lemma EC.1 and a weak duality result Lemma EC.2.

Lemma EC.1. (Several Measurable Functions). *Assume that Assumption 1 holds, then the following results hold.*

(I) *Define the function* $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda):\mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times\mathbb{R}_+\to\mathbb{R}$ *as*

$$w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda):=\sup_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}\mathbb{E}_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}\Bigg[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\\ -\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))-\lambda\epsilon\log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y}\,|\,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right)\Bigg].$$

This function is jointly measurable with respect to $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})$ *regardless of the choice of* $\lambda\geq 0$.

(II) *Define the function* $g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda):\mathcal{X}\times\mathcal{X}\times\mathbb{R}_+\to\mathbb{R}$ *as*

$$\begin{aligned}g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)&:=\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\Big[w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda)\Big]\\&=\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda\epsilon\log\int_{\mathcal{Y}}\exp\left(\frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y})-\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right].\end{aligned}$$

This function is jointly measurable with respect to $(\widehat{\boldsymbol{x}},\boldsymbol{x})$ *regardless of the choice of* $\lambda\geq 0$.

(III) *Define the function* $w_2(\widehat{\boldsymbol{x}},\lambda):\mathcal{X}\times\mathbb{R}_+\to\mathbb{R}$ *as*

$$w_2(\widehat{\boldsymbol{x}},\lambda):=\sup_{\gamma\in\Gamma_c(\mathbb{P},\mathbb{Q})}\left\{\mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\left[g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)-\lambda\epsilon\log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x}\,|\,\widehat{\boldsymbol{x}})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right)\right]\right\}.$$

This function is measurable with respect to $\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$ *regardless of the choice of* $\lambda\geq 0$.

Proof of Lemma EC.1. We prove the measurability of the three functions in sequence.

(I) **For function** $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda)$, we consider the following two cases.

- When $\lambda=0$, according to Lemma EC.2 in the E-companion of Wang et al. (2025), it holds that
$$w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)=\operatorname*{ess\,sup}_{\nu_{\mathcal{Y}}}\quad\Psi(f(\boldsymbol{x}),\boldsymbol{y}).$$
By Assumption 1, the loss function $\Psi(f(\boldsymbol{x}),\boldsymbol{y})$ is measurable. Since the essential supremum of a measurable function with respect to one variable is a measurable function of the remaining variables (Blackwell and Ryll-Nardzewski, 1963), the function $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)$ is jointly measurable with respect to $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})$.
- When $\lambda>0$, using the Fenchel duality, we have
$$w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda)=\lambda\epsilon\log\int_{\mathcal{Y}}\exp\left(\frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y})-\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y}).$$
According to Assumption 1, function c_p is measurable. Since the difference of two measurable functions is measurable, the function $\Psi(f(\boldsymbol{x}),\boldsymbol{y})-\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))$ is measurable with respect to $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})$. As the composition of a measurable function with

a continuous function is measurable, the function $\exp\left(\frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y})-\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon}\right)$ is measurable. According to Tonelli's Theorem (Cohn, 2013, Proposition 5.2.1), integrating a non-negative, jointly measurable function with respect to one variable (here, $\boldsymbol{y}$) yields a function that is measurable with respect to the remaining variables. Therefore, the function $\int_{\mathcal{Y}} \exp\left(\frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y})-\lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})$ is measurable, and its compositional with the continuous function $\log(\cdot)$ is also measurable, i.e., the function $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda)$ is jointly measurable with respect to $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})$ when $\lambda > 0$.

Hence, function $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda)$ is jointly measurable with respect to $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})$ regardless of the choice of $\lambda \geq 0$.

(II) **For function** $g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)$, we also consider the following two cases.

- When $\lambda > 0$, w_1 is positive-valued function, which implies that w_1 is measurable. On a probability space, any finite-valued measurable function is integrable. Therefore, by the Tonelli's Theorem, the function $g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)$ is jointly measurable with respect to $(\widehat{\boldsymbol{x}},\boldsymbol{x})$.
- When $\lambda = 0$, for the optimization problem to be well-posed, the loss function Ψ must be bounded below, which implies that $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)$ is also bounded below. Let the lower bound of function $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)$ be $M \in \mathbb{R}$. We decompose the function $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)$ as two non-negative measurable functions

$$w_1^+(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0) = \max\{0, w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\} \in [0,+\infty],$$

and

$$w_1^-(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0) = -\min\{0, w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\} \in [0,-\min\{0,M\}],$$

such that

$$\begin{aligned} g(\widehat{\boldsymbol{x}},\boldsymbol{x},0) &= \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\left[w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\right] \\ &= \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\left[w_1^+(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\right] - \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\left[w_1^-(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\right]. \end{aligned}$$

Based on Tonelli's Theorem, both $\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\left[w_1^+(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\right]$ and $\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\left[w_1^-(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\right]$ are well-defined measurable functions of $(\widehat{\boldsymbol{x}},\boldsymbol{x})$. Since $\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\left[w_1^-(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},0)\right]$ is a finite measurable function, this subtraction is well-defined (it avoids the $\infty-\infty$ form). The difference of two measurable functions is measurable. Thus, function $g(\widehat{\boldsymbol{x}},\boldsymbol{x},0)$ is also jointly measurable with respect to $(\widehat{\boldsymbol{x}},\boldsymbol{x})$.

Hence, function $g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)$ is jointly measurable with respect to $(\widehat{\boldsymbol{x}},\boldsymbol{x})$ regardless of the choice of $\lambda \geq 0$.

(III) **For function** $w_2(\widehat{\boldsymbol{x}},\lambda)$, according to Lemma EC.3 in the E-companion of Wang et al. (2025), it's measurable with respect to $\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$ regardless of the choice of $\lambda \geq 0$.

This completes the proof. □

Lemma EC.2. (Weak Duality). *Under Assumption 1, $v_{\mathrm{P}} \le v_{\mathrm{D}}$.*

Proof of Lemma EC.2. For the primal problem (Causal-SDRO), let $\lambda \ge 0$ be the Lagrangian multiplier of the causal Sinkhorn ball constraint. Then, we have the following dual function:

$$\begin{aligned} q(\lambda) &:= \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \left[\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}[\Psi(f(\boldsymbol{x}),\boldsymbol{y})] - \lambda\left(R_p(\widehat{\mathbb{P}},\mathbb{P})^p - \rho^p\right)\right] \\ &= \lambda\rho^p + \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \left[\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}[\Psi(f(\boldsymbol{x}),\boldsymbol{y})] - \lambda R_p(\widehat{\mathbb{P}},\mathbb{P})^p\right]. \end{aligned}$$

Define

$$v_{\mathrm{P}}(\lambda) := \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \left[\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}[\Psi(f(\boldsymbol{x}),\boldsymbol{y})] - \lambda R_p(\widehat{\mathbb{P}},\mathbb{P})^p\right], \tag{EC.2}$$

then

$$q(\lambda) = \lambda\rho^p + v_{\mathrm{P}}(\lambda).$$

Based on Lagrangian weak duality, for the primal problem, we have

$$\begin{aligned} v_{\mathrm{P}} &= \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \left\{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\left[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\right] : R_p(\widehat{\mathbb{P}},\mathbb{P})^p \le \rho^p\right\} \\ &= \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \inf_{\lambda\ge 0} \left\{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\left[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\right] - \lambda\left(R_p(\widehat{\mathbb{P}},\mathbb{P})^p - \rho^p\right)\right\} \\ &\le \inf_{\lambda\ge 0} \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \left\{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\left[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\right] - \lambda\left(R_p(\widehat{\mathbb{P}},\mathbb{P})^p - \rho^p\right)\right\} \\ &= \inf_{\lambda\ge 0}\left\{\lambda\rho^p + v_{\mathrm{P}}(\lambda)\right\}, \end{aligned} \tag{EC.3}$$

where the inequality holds because of the min-max inequality.

We reformulate the term $v_{\mathrm{P}}(\lambda)$ to prove the weak duality. For the formulation of CSD

$$R_p(\widehat{\mathbb{P}},\mathbb{P})^p = \inf_{\gamma\in\Gamma_c(\widehat{\mathbb{P}},\mathbb{P})} \mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\left[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) + \epsilon H\left(\gamma \mid \mu\otimes\left(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}\right)\right)\right], \tag{EC.4}$$

where the second item can be reformulated as

$$\begin{aligned} H\left(\gamma \mid \mu\otimes\left(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}\right)\right) &= \log\left(\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \\ &= \log\left(\frac{\mathrm{d}\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}})\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x}\mid\widehat{\boldsymbol{x}})\mathrm{d}\gamma_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}},\boldsymbol{X}}(\widehat{\boldsymbol{y}}\mid\widehat{\boldsymbol{x}},\boldsymbol{x})\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y}\mid\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}})\mathrm{d}\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{y}}\mid\widehat{\boldsymbol{x}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \\ &= \log\left(\frac{\mathrm{d}\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}})\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x}\mid\widehat{\boldsymbol{x}})\mathrm{d}\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{y}}\mid\widehat{\boldsymbol{x}})\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y}\mid\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}})\mathrm{d}\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{y}}\mid\widehat{\boldsymbol{x}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \\ &= \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x}\mid\widehat{\boldsymbol{x}})\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y}\mid\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \\ &= \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x}\mid\widehat{\boldsymbol{x}})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right) + \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y}\mid\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right), \end{aligned} \tag{EC.5}$$

where the second equality is due to the chain rule decomposition of the densities for both the joint measure γ and the empirical distribution $\widehat{\mathbb{P}}$, and $\gamma_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}}) = \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{x}})$. The third equality in relation (EC.5) holds since $\gamma_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}},\boldsymbol{X}}(\widehat{\boldsymbol{y}} \mid \widehat{\boldsymbol{x}}, \boldsymbol{x}) = \gamma_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{y}} \mid \widehat{\boldsymbol{x}}) = \widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}(\widehat{\boldsymbol{y}} \mid \widehat{\boldsymbol{x}})$ under the causal optimal transport setting. By the tower property, we have

$$\mathbb{E}_{\gamma}[\cdot] = \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}[\cdot \mid \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \boldsymbol{X}] \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\right] \mid \widehat{\boldsymbol{X}}\right]\right]. \tag{EC.6}$$

From relations (EC.5) and (EC.6), the expectation term on the RHS of the Equation (EC.4) can be rewritten as

$$\begin{aligned}
&\mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\left[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) + \epsilon H\left(\gamma \mid \mu \otimes \left(\nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}}\right)\right)\right] \\
&= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) + \epsilon H\left(\gamma \mid \mu \otimes \left(\nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}}\right)\right) \mid \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \boldsymbol{X}] \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\right] \mid \widehat{\boldsymbol{X}}\right]\right] \\
&= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Bigg[\mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\Bigg[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\Bigg[\mathbb{E}_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) \\
&\qquad + \epsilon \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y} \mid \widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \mid \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \boldsymbol{X}] \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\Bigg] + \epsilon \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x} \mid \widehat{\boldsymbol{x}})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right) \mid \widehat{\boldsymbol{X}}\Bigg]\Bigg],
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
v_{\mathrm{P}}(\lambda) = \sup_{\gamma \in \Gamma_c(\mathbb{P},\mathbb{Q})} \Bigg\{ &\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Bigg[\mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\Bigg[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\Bigg[\mathbb{E}_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}\Bigg[\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) \\
&- \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y} \mid \widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \mid \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \boldsymbol{X}\Bigg] \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\Bigg] - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x} \mid \widehat{\boldsymbol{x}})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right) \mid \widehat{\boldsymbol{X}}\Bigg]\Bigg]\Bigg\}\Bigg\}.
\end{aligned}$$

Similar to relation (EC.6), the optimization for γ can be decomposed to optimize $\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}$ and $\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}$ (distributions $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$ and $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}$ are determined). Thus, using Jensen's inequality, we have

$$\begin{aligned}
v_{\mathrm{P}}(\lambda) &\le \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Bigg[\sup_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}} \mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\Bigg[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\Bigg[\sup_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}} \mathbb{E}_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}\Bigg[\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) \\
&\qquad - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}(\boldsymbol{y} \mid \widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})}{\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \mid \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \boldsymbol{X}\Bigg] \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\Bigg] - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x} \mid \widehat{\boldsymbol{x}})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right) \mid \widehat{\boldsymbol{X}}\Bigg]\Bigg] \\
&= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Bigg[\underbrace{\sup_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}} \mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\Bigg[\underbrace{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\Bigg[w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda) \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\Bigg]}_{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)} - \lambda\epsilon \log\left(\frac{\mathrm{d}\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}(\boldsymbol{x} \mid \widehat{\boldsymbol{x}})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right) \mid \widehat{\boldsymbol{X}}\Bigg]}_{w_2(\widehat{\boldsymbol{x}},\lambda)}\Bigg],
\end{aligned}$$

where functions $w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda)$, $g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)$, and $w_2(\widehat{\boldsymbol{x}},\lambda)$ are all measurable for any $\lambda \ge 0$ according to Lemma EC.1. By the Fenchel duality, we have

$$\sup_{\mathbb{P}}\left\{\mathbb{E}_{\boldsymbol{y}\sim\mathbb{P}}[f(\boldsymbol{y})] - \epsilon H(\mathbb{P}|\mathbb{Q})\right\} = \epsilon \log \int \exp\left(\frac{f(\boldsymbol{y})}{\epsilon}\right)\mathrm{d}\mathbb{Q}(\boldsymbol{y}),$$

and it follows that

$$v_{\mathrm{P}}(\lambda) \leq \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon \log \int_{\mathcal{X}} \exp\left(\frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right].$$

Hence, according to Equation (EC.3), we have

$$\begin{aligned} v_{\mathrm{P}} &\leq \inf_{\lambda\geq 0}\left\{\lambda\rho^p + v_{\mathrm{P}}(\lambda)\right\} \\ &\leq \inf_{\lambda\geq 0}\left\{\lambda\rho^p + \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon \log \int_{\mathcal{X}} \exp\left(\frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right]\right\} \\ &= v_{\mathrm{D}}. \end{aligned}$$

This completes the proof of the weak duality. □

In the following, we complete the proof of the strong duality theorem.

Proof of Theorem 1. To prove Theorem 1(I), we first rewrite the constraint of the primal problem (Causal-SDRO) as

$$\mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\left[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) + \epsilon \log\left(\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right)\right] \leq \rho^p. \tag{EC.7}$$

Based on Assumption 1(II), the relation (EC.7) can be reformulated as

$$\mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}}\left[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) + \epsilon \log\left(\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right)\right] \leq \rho^p. \tag{EC.8}$$

We define a kernel probability distribution $\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}(\boldsymbol{x},\boldsymbol{y})$ using the kernel distributions Q_ϵ and W_ϵ from Equations (3a) and (3b):

$$\mathrm{d}\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}(\boldsymbol{x},\boldsymbol{y}) := \mathrm{d}Q_\epsilon(\boldsymbol{x})\cdot\mathrm{d}W_\epsilon(\boldsymbol{y}) = \frac{e^{-c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))/\epsilon}}{Z(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y}),$$

where $Z(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}) = \int_{\mathbb{R}^{d_x}} e^{-\|\boldsymbol{u}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u})\cdot\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})$. Therefore, we have

$$\log\left(\frac{\mathrm{d}\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) = -\frac{c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\epsilon} - \log Z(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}).$$

We decompose the logarithm term in the constraint

$$\begin{aligned} \log\left(\frac{\mathrm{d}\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) &= \log\left(\frac{\mathrm{d}\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}(\boldsymbol{x},\boldsymbol{y})}\right) + \log\left(\frac{\mathrm{d}\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right) \\ &= \mathbb{D}_{\mathrm{KL}}\left(\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}\|\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}\right) - \frac{c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\epsilon} - \log Z(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}), \end{aligned}$$

where $\mathbb{D}_{\mathrm{KL}}\left(\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}\|\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}\right)$ represents the KL-divergence from distribution $\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}$ to $\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}$. Thus, we reformulate (EC.8) as the following equivalent constraint in terms of the KL-divergence

$$\epsilon\cdot\mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\left[\mathbb{D}_{\mathrm{KL}}\left(\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}\|\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}\right)\right] \leq \rho^p + \epsilon\cdot\mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\left[\log Z(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\right] = \bar{\rho}, \tag{EC.9}$$

which implies that the constraint of problem (Causal-SDRO) is equivalent to constraint (EC.9),

For Theorem 1(I), according to the constraint (EC.9), we first prove the "if" part. When $\bar{\rho} \geq 0$, the primal problem v_{P} has at least one feasible solution, i.e., the empirical distribution $\widehat{\mathbb{P}}$. Thus, the primal problem is feasible. The "if" part is thus completed.

Next, we prove the "only if" part by contradiction. Suppose that if the primal problem is feasible, then $\bar{\rho} < 0$. According to the definition of the KL-divergence, the constraint (EC.9) always satisfies $\mathbb{D}_{\mathrm{KL}}\left(\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})} \| \mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}\right) \geq 0$ for any distributions $\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})}$ and $\mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}$. Thus, if the primal problem is feasible, we have $0 \leq \epsilon \cdot \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\left[\mathbb{D}_{\mathrm{KL}}\left(\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})} \| \mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}\right)\right] \leq \bar{\rho}$, which contradicts the assumption that $\bar{\rho} < 0$. Therefore, the "only if" part is completed.

For Theorem 1(II), we first consider that there exists $\lambda > 0$ such that $\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\exp\left(\frac{g'(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon}\right)\right] < \infty$ for $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$-almost every $\widehat{\boldsymbol{x}}$, and $\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right] < \infty$ for $\widehat{\mathbb{P}}\otimes\nu_{\mathcal{X}}$-almost every $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x})$. According to Lemma EC.2, we already have $v_{\mathrm{P}} \leq v_{\mathrm{D}}$. Thus, we next prove $v_{\mathrm{P}} \geq v_{\mathrm{D}}$. Denote the optimal solution of v_{D} is λ^*, and the optimal distribution of v_{P} is $\mathbb{P}^*$. Suppose $\bar{\rho} \geq 0$ is bounded above such that the CSD constraint is binding, i.e., $R_p\left(\widehat{\mathbb{P}}, \mathbb{P}^*\right) = \rho$ and $\epsilon \cdot \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\left[\mathbb{D}_{\mathrm{KL}}\left(\gamma_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})} \| \mathcal{K}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),\epsilon}\right)\right] = \bar{\rho}$. Therefore, there always exists $\lambda^* > 0$.

Since the dual problem is convex in λ, the λ^* satisfies the following first-order optimality condition

$$\rho^p + \epsilon\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\log\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right] = \frac{1}{\lambda^*}\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}t(\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right], \tag{EC.10}$$

where

$$r(\widehat{\boldsymbol{x}},\boldsymbol{x}) = \frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda^*)}{\lambda^*\epsilon} = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\log\int_{\mathcal{Y}} e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right],$$

$$t(\widehat{\boldsymbol{x}},\boldsymbol{x}) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\frac{\int_{\mathcal{Y}} e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\cdot\Psi(f(\boldsymbol{x}),\boldsymbol{y})\cdot\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}{\int_{\mathcal{Y}} e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right],$$

and

$$s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y}) = \frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda^* c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda^*\epsilon}.$$

We next construct a distribution $\mathbb{P}_*$, which can be proved to be feasible for the primal problem. We take the transport mapping γ_* such that

$$\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \propto e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}\cdot e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})},$$

Let $\alpha_{\widehat{\boldsymbol{x}}} = \left(\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^{-1}$ and $\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}} = \left(\int_{\mathcal{Y}} e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right)^{-1}$, we have

$$\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \alpha_{\widehat{\boldsymbol{x}}}\cdot\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}}\cdot e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})+s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}.$$

We verify the feasibility of distribution $\mathbb{P}_*$ by the definition of the CSD, that is

$$
\begin{aligned}
R_p(\widehat{\mathbb{P}}, \mathbb{P}_*) &= \inf_{\gamma \in \Gamma_c(\widehat{\mathbb{P}}, \mathbb{P}_*)} \mathbb{E}_{((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \sim \gamma} \left[c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) + \epsilon \log \left(\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \right) \right] \\
&= \inf_{\gamma \in \Gamma_c(\widehat{\mathbb{P}}, \mathbb{P}_*)} \mathbb{E}_{((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \sim \gamma} \left[\epsilon \log \left(\frac{e^{c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))/\epsilon} \cdot \mathrm{d}\gamma((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \right) \right] \\
&\leq \mathbb{E}_{((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \sim \gamma_*} \left[\epsilon \log \left(\frac{e^{c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))/\epsilon} \cdot \mathrm{d}\gamma_*((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \right) \right] \\
&= \mathbb{E}_{((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \sim \gamma_*} \left[\epsilon \log \left(e^{c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))/\epsilon} \cdot \alpha_{\widehat{\boldsymbol{x}}} \cdot e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} \cdot \beta_{\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}} \cdot e^{s(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}, \boldsymbol{y})} \right) \right] \\
&= \mathbb{E}_{((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \sim \gamma_*} \left[\frac{1}{\lambda^*} \Psi(f(\boldsymbol{x}), \boldsymbol{y}) + \epsilon r(\widehat{\boldsymbol{x}}, \boldsymbol{x}) + \epsilon \log(\alpha_{\widehat{\boldsymbol{x}}}) + \epsilon \log \left(\beta_{\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}} \right) \right] \\
&= \frac{1}{\lambda^*} \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\frac{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} t(\widehat{\boldsymbol{x}}, \boldsymbol{x}) \cdot \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})} \right] - \epsilon \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\log \int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \right] \\
&= \rho^p,
\end{aligned}
$$

where the inequality relation is because γ_* is a feasible solution in $\Gamma_c(\widehat{\mathbb{P}}, \mathbb{P}_*)$, and the fourth and fifth equalities are by substituting the expression of γ_*, and the last equality is due to the first-order optimality condition (EC.10). Therefore, under Assumption 1, the distribution $\mathbb{P}_*$ is feasible for the primal problem. We show that the primal optimal value is lower bounded by the dual optimal value

$$
\begin{aligned}
v_{\mathrm{P}} &\geq \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathbb{P}_*} \left[\Psi(f(\boldsymbol{x}), \boldsymbol{y}) \right] \\
&= \mathbb{E}_{((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \sim \gamma_*} \left[\Psi(f(\boldsymbol{x}), \boldsymbol{y}) \right] \\
&= \int_{(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})} \Psi(f(\boldsymbol{x}), \boldsymbol{y}) \mathrm{d}\gamma((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \cdot \frac{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}{\mathrm{d}\gamma((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))} \cdot \frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \\
&= \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\frac{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} t(\widehat{\boldsymbol{x}}, \boldsymbol{x}) \cdot \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})} \right] \\
&= \lambda^* \rho^p + \lambda^* \epsilon \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\log \int_{\mathcal{X}} \exp \left(\frac{g(\widehat{\boldsymbol{x}}, \boldsymbol{x}, \lambda^*)}{\lambda^* \epsilon} \right) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \right] \\
&= v_{\mathrm{D}}.
\end{aligned}
\tag{EC.11}
$$

According to Lemma EC.2 and Equation (EC.11), we have $v_{\mathrm{P}} \geq v_{\mathrm{D}}$ and $v_{\mathrm{P}} \leq v_{\mathrm{D}}$. Thus $v_{\mathrm{P}} = v_{\mathrm{D}}$. The proof for Theorem 1(II) is completed.

If for any $\lambda > 0$, at least one between $\mathbb{E}_{\boldsymbol{\xi}_1 \sim Q_\epsilon} \left[e^{\left(\frac{g'(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1, \lambda)}{\lambda \epsilon} \right)} \right] = \infty$ and $\mathbb{E}_{\boldsymbol{\xi}_2 \sim W_\epsilon} \left[e^{\left(\frac{\Psi(f(\widehat{\boldsymbol{x}} + \boldsymbol{\xi}_1), \widehat{\boldsymbol{y}} + \boldsymbol{\xi}_2)}{\lambda \epsilon} \right)} \right] = \infty$ holds, then $q(\lambda) = \lambda \rho^p + \infty$. Therefore, in this case, we have $v_{\mathrm{P}} = v_{\mathrm{D}} = \infty$. □

EC.1.3 Analysis for Remark 1

For function g in Equation (2b), when $\epsilon \to 0$, we have

$$
\begin{aligned}
&\lim_{\epsilon\to 0} g(\widehat{\boldsymbol{x}}, \boldsymbol{x}, \lambda, \epsilon) \\
&= \lim_{\epsilon\to 0} \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda\epsilon \log \int_{\mathcal{Y}} \exp\left(\frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right] \\
&= \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda \lim_{\tau\to\infty} \frac{1}{\tau}\cdot \log \int_{\mathcal{Y}} \exp\left(\frac{\tau\Big(\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\Big)}{\lambda}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right] \\
&= \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda \lim_{\tau\to\infty} \nabla_\tau \log \int_{\mathcal{Y}} \exp\left(\frac{\tau\Big(\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\Big)}{\lambda}\right)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right] \\
&= \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lim_{\tau\to\infty} \frac{\int_{\mathcal{Y}} e^{\tau\left(\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\right)/\lambda}\Big(\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\Big)\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}{\int_{\mathcal{Y}} e^{\tau\left(\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\right)/\lambda}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right] \\
&= \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\sup_{\boldsymbol{y}\in\operatorname{supp}\nu_{\mathcal{Y}}}\Big\{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\Big\}\right],
\end{aligned}
$$

where the third equality is due to L'Hôpital's rule. Then, for the Equation (2a), similarly we have

$$
\begin{aligned}
&\lim_{\epsilon\to 0} \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon \log \int_{\mathcal{X}} \exp\left(\frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda,\epsilon)}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right] \\
=&\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lim_{\tau\to\infty} \frac{\lambda}{\tau}\log \int_{\mathcal{X}} \exp\left(\frac{\tau\cdot g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda,\frac{1}{\tau})}{\lambda}\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right] \\
=&\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\sup_{\boldsymbol{x}\in\operatorname{supp}\nu_{\mathcal{X}}}\left\{\lim_{\tau\to\infty} g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda,\frac{1}{\tau})\right\}\right] \\
=&\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\sup_{\boldsymbol{x}\in\operatorname{supp}\nu_{\mathcal{X}}}\left\{\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\sup_{\boldsymbol{y}\in\operatorname{supp}\nu_{\mathcal{Y}}}\Big\{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\Big\}\right]\right\}\right],
\end{aligned}
$$

where the second equality is due to the properties of the Log-Sum-Exp limit, cf. Laplace's method. When $\operatorname{supp}\nu_{\mathcal{X}} = \mathcal{X}$ and $\operatorname{supp}\nu_{\mathcal{Y}} = \mathcal{Y}$, the dual objective function of the problem (Causal-SDRO) converges into that of the problem (Causal-WDRO).

EC.1.4 Proof of Theorem 2 in Section 3.2

Proof of Theorem 2. In the proof of Theorem 1, we have derived a worst-case distribution of v_{P}, that is,

$$
\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \alpha_{\widehat{\boldsymbol{x}}}\cdot\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}}\cdot e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})+s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}, \tag{EC.12}
$$

where

$$
\alpha_{\widehat{\boldsymbol{x}}} = \left(\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^{-1}, \qquad \beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}} = \left(\int_{\mathcal{Y}} e^{s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right)^{-1},
$$

and function $r(\widehat{\boldsymbol{x}},\boldsymbol{x})$ and $s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})$ are defined in (EC.11). We now prove that λ^* is the unique optimal solution of the dual problem, which implies that the worst-case distribution is also unique.

Recall that $v(\lambda)$ denotes the objective function for the dual problem, then we have

$$\nabla_\lambda v(\lambda) = \rho^p + \epsilon \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\log \int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right] - \frac{1}{\lambda}\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right],$$

where

$$r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x}) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\log \int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right],$$

$$t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x}) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\frac{\int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\Psi(f(\boldsymbol{x}),\boldsymbol{y})\cdot \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}{\int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right],$$

and

$$s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y}) = \frac{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon}.$$

For its second-order derivative function, we have

$$\begin{aligned}
\nabla^2_\lambda v(\lambda) &= \nabla_\lambda\left[\epsilon\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\log\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right]\right] - \nabla_\lambda\left[\frac{1}{\lambda}\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r\lambda,(\widehat{\boldsymbol{x}},\boldsymbol{x})} t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right]\right] \\
&= -\frac{1}{\lambda^2}\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right] - \nabla_\lambda\left[\frac{1}{\lambda}\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right]\right] \\
&= -\frac{1}{\lambda}\cdot\nabla_\lambda\left[\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right]\right].
\end{aligned}$$

Here, we have

$$\begin{aligned}
&\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\right] \\
&\quad= -\frac{1}{\lambda^2\epsilon}\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\frac{\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\left(t^2(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x}) + u(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\cdot\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\left(\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^2}\right. \\
&\qquad\qquad\left. - \frac{\left(\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})} t(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^2}{\left(\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^2}\right],
\end{aligned}$$

where

$$\begin{aligned}
u(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x}) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\frac{\int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\Psi^2(f(\boldsymbol{x}),\boldsymbol{y})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\cdot\int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}{\left(\int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right)^2}\right. \\
\left. - \frac{\left(\int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\Psi(f(\boldsymbol{x}),\boldsymbol{y})\cdot\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right)^2}{\left(\int_{\mathcal{Y}} e^{s(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})\right)^2}\right].
\end{aligned}$$

According to the Cauchy-Schwarz inequality, we have $u(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x}) \geq 0$ for any $\lambda \geq 0$, $\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$ and $\boldsymbol{x} \sim \nu_{\mathcal{X}}$. Thus, we have

$$\mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[\int_{\mathcal{X}} u(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \cdot \int_{\mathcal{X}} e^{r(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\Big] \geq 0, \tag{EC.13}$$

and it follows that

$$\begin{aligned}
\nabla^2_\lambda v(\lambda) \geq \frac{1}{\lambda^3 \epsilon} \cdot \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Bigg[& \frac{\int_{\mathcal{X}} e^{r(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x})} t^2(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x}) \cdot \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \int_{\mathcal{X}} e^{r(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\left(\int_{\mathcal{X}} e^{r(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^2} \\
& - \frac{\left(\int_{\mathcal{X}} e^{r(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x})} t(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x}) \cdot \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^2}{\left(\int_{\mathcal{X}} e^{r(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right)^2} \Bigg] \\
\geq 0,
\end{aligned}$$

where the first inequality is due to relation (EC.13), and the second inequality is due to the Cauchy-Schwarz inequality.

Therefore, for any $\lambda > 0$, we have $\nabla^2_\lambda v(\lambda) \geq 0$, and the equality holds if and only if the function Ψ is a constant. Thus, the strict convexity of the dual problem (2a) holds for the dual objective, and it implies the uniqueness of λ^*. □

EC.1.5 Proof of Proposition 1 in Section 4.2

Proof of Proposition 1. For brevity, let $z_{i,t} := \boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}$ and let $s_{i,t} := \mathrm{S}(z_{i,t})$ be the Sigmoid activation at node i. For an internal node i on route l in tree t, the routing probability $\Omega_{i,t}(\boldsymbol{x})$ is defined as:

$$\Omega_{i,t}(\boldsymbol{x}) := \begin{cases} s_{i,t}, & \text{if } l \text{ goes left at } i; \\ 1 - s_{i,t}, & \text{if } l \text{ goes right at } i; \end{cases} \qquad \forall i \in \Lambda(l).$$

Then the route probability is given by $p_{l,t}(\boldsymbol{x}) = \prod_{i \in \Lambda(l)} \Omega_{i,t}(\boldsymbol{x})$. Taking the logarithm of both sides yields:

$$\ln p_{l,t}(\boldsymbol{x}) = \sum_{i \in \Lambda(l)} \ln \Omega_{i,t}(\boldsymbol{x}). \tag{EC.14}$$

Equation (EC.14) is well-defined since $p_{l,t}(\boldsymbol{x}) > 0$ and $\Omega_{i,t}(\boldsymbol{x}) > 0$ strictly hold for any $\boldsymbol{x} \in \mathcal{X}$.

Taking the partial derivative of both sides of Equation (EC.14) with respect to feature x_j:

$$\frac{\partial \ln p_{l,t}(\boldsymbol{x})}{\partial x_j} = \frac{1}{p_{l,t}(\boldsymbol{x})} \frac{\partial p_{l,t}(\boldsymbol{x})}{\partial x_j} = \sum_{i \in \Lambda(l)} \frac{1}{\Omega_{i,t}(\boldsymbol{x})} \frac{\partial \Omega_{i,t}(\boldsymbol{x})}{\partial x_j}.$$

Recall that the derivative of the Sigmoid function is $s'_{i,t} = s_{i,t}(1 - s_{i,t})$, for each node i on route l in tree t, we define:

$$\psi_{i,t} := \frac{1}{\Omega_{i,t}(\boldsymbol{x})} \frac{\partial \Omega_{i,t}(\boldsymbol{x})}{\partial z_{i,t}} = \begin{cases} \frac{s_{i,t}(1-s_{i,t})}{s_{i,t}} = 1 - s_{i,t}, & \text{if route } l \text{ goes left at } i; \\ \frac{-s_{i,t}(1-s_{i,t})}{1-s_{i,t}} = -s_{i,t}, & \text{if route } l \text{ goes right at } i. \end{cases}$$

Since $\frac{\partial \Omega_{i,t}}{\partial x_j} = \frac{\partial \Omega_{i,t}}{\partial z_{i,t}} \cdot [\boldsymbol{w}_{i,t}]_j$, we obtain the first-order derivative:

$$\frac{\partial p_{l,t}(\boldsymbol{x})}{\partial x_j} = p_{l,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_j. \tag{EC.15}$$

We next compute the second-order derivative $\frac{\partial^2 p_{l,t}(\boldsymbol{x})}{\partial x_j \partial x_k}$ by differentiating Equation (EC.15) with respect to x_k. Applying the product rule yields two terms:

$$\frac{\partial^2 p_{l,t}}{\partial x_j \partial x_k} = \underbrace{\frac{\partial p_{l,t}}{\partial x_k} \left(\sum_{i \in \Lambda(l)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_j \right)}_{\text{Term 1}} + \underbrace{p_{l,t} \frac{\partial}{\partial x_k} \left(\sum_{i \in \Lambda(l)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_j \right)}_{\text{Term 2}},$$

where Term 1 is equivalent to

$$\text{Term 1} = p_{l,t}(\boldsymbol{x}) \left(\sum_{i \in \Lambda(l)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_k \right) \left(\sum_{i \in \Lambda(l)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_j \right).$$

For Term 2, we note that $\frac{\partial \psi_{i,t}}{\partial x_k} = \frac{\partial \psi_{i,t}}{\partial z_{i,t}} \cdot [\boldsymbol{w}_{i,t}]_k$. For both directions (left or right), the derivative of $\psi_{i,t}$ with respect to $z_{i,t}$ is identical:

$$\frac{\partial \psi_{i,t}}{\partial z_{i,t}} = \begin{cases} \frac{\partial (1 - s_{i,t})}{\partial z_{i,t}} = -s_{i,t}(1 - s_{i,t}), & \text{if route } l \text{ goes left at } i; \\ \frac{\partial (-s_{i,t})}{\partial z_{i,t}} = -s_{i,t}(1 - s_{i,t}), & \text{if route } l \text{ goes left at } i. \end{cases}$$

Thus we have $\frac{\partial \psi_{i,t}}{\partial x_k} = -s_{i,t}(1 - s_{i,t})[\boldsymbol{w}_{i,t}]_k$, and it follows that

$$\text{Term 2} = -p_{l,t}(\boldsymbol{x}) \sum_{i \in \Lambda(l)} -s_{i,t}(1 - s_{i,t}) [\boldsymbol{w}_{i,t}]_j [\boldsymbol{w}_{i,t}]_k.$$

Combining both terms, the second-order derivative is explicitly characterized by:

$$\frac{\partial^2 p_{l,t}(\boldsymbol{x})}{\partial x_j \partial x_k} = p_{l,t}(\boldsymbol{x}) \left[\left(\sum_{i \in \Lambda(l)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_j \right) \left(\sum_{i \in \Lambda(l)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_k \right) - \sum_{i \in \Lambda(l)} s_{i,t}(1 - s_{i,t}) [\boldsymbol{w}_{i,t}]_j [\boldsymbol{w}_{i,t}]_k \right].$$

□

EC.1.6 Proof of Proposition 2 in Section 4.2

Proof of Proposition 2. Since the covariate space $\mathcal{X}$ is compact, and the SRF consists of smooth sigmoid compositions, $f_{\boldsymbol{\theta}}^{\text{SRF}}$ is continuously differentiable. Therefore, to prove Lipschitz continuity, it suffices to show that $\|\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})\|_2$ is bounded by L^{SRF}. Similarly, to establish Lipschitz smoothness, it suffices to show that the Lipschitz constant of the $\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})$ is bounded by S^{SRF}.

For Lipschitz continuity, the upper bound of $\|\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x})\|_2$ is given by:

$$\begin{aligned}\|\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x})\|_2 &= \left\| \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} \nabla_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x}) \cdot \boldsymbol{\pi}_{l,t}^{\top} \right\|_2 \\ &\leq \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} \left\| \nabla_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x}) \cdot \boldsymbol{\pi}_{l,t}^{\top} \right\|_2 \\ &= \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} \left\| \nabla_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x}) \right\|_2 \cdot \left\| \boldsymbol{\pi}_{l,t} \right\|_2,\end{aligned} \tag{EC.16}$$

where the inequality is due to the triangle inequality, and the second equality is due to the property of the spectral norm for outer products.

We next analyze the gradient of route probability $p_{l,t}(\boldsymbol{x})$. Let $W_{\max} := \max_{i,t} \|\boldsymbol{w}_{i,t}\|_2$ be the maximum norm of gating weights, $\Pi_{\max} := \max_{l,t} \|\boldsymbol{\pi}_{l,t}\|_2$ be the maximum norm of leaf vectors, and $D_{\max}$ be the maximum tree depth. According to Proposition 1, we have

$$\begin{aligned}\left\| \nabla_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x}) \right\|_2 &= \left\| p_{l,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l)} \psi_{i,t} \boldsymbol{w}_{i,t} \right\|_2 \\ &\leq p_{l,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l)} \left\| \psi_{i,t} \boldsymbol{w}_{i,t} \right\|_2 \\ &\leq p_{l,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l)} \left\| \boldsymbol{w}_{i,t} \right\|_2 \\ &\leq p_{l,t}(\boldsymbol{x}) \cdot (D_{\max} - 1) \cdot W_{\max}.\end{aligned} \tag{EC.17}$$

Here, the first inequality is due to the triangle inequality, and the second inequality is due to $\psi_{i,t} \in (-1, 1)$. The final inequality is because $\left|\Lambda(l)\right| \leq D_{\max} - 1$ for all $l \in [2^{D(t)-1}]$ and $t \in T$.

Based on Equation (EC.17), Equation (EC.16) can be bounded by

$$\begin{aligned}\|\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x})\|_2 &\leq \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} p_{l,t}(\boldsymbol{x}) \cdot (D_{\max} - 1) \cdot W_{\max} \cdot \Pi_{\max} \\ &= (D_{\max} - 1) \cdot W_{\max} \cdot \Pi_{\max} \\ &= L^{\mathrm{SRF}},\end{aligned}$$

where the equality is due to $\sum_{l=1}^{2^{D(t)}} p_{l,t}(\boldsymbol{x}) = 1$. Thus, $f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x})$ is Lipschitz continuous in $\boldsymbol{x}$.

For Lipschitz smoothness, we examine the spectral norm of the Hessian of the route probabil-

ity $\left\|\nabla^2_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x})\right\|_2$:

$$\begin{aligned}
\left\|\nabla^2_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x})\right\|_2 &= \Bigg\| p_{l,t}(\boldsymbol{x}) \cdot \Bigg[\Big(\sum_{i\in\Lambda(l)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big)\Big(\sum_{i\in\Lambda(l)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big)^\top - \\
&\qquad \sum_{i\in\Lambda(l)} \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big(1 - \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big)\boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}^\top \Bigg]\Bigg\|_2 \\
&\le p_{l,t}(\boldsymbol{x}) \cdot \Bigg[\Big\| \Big(\sum_{i\in\Lambda(l)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big)\Big(\sum_{i\in\Lambda(l)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big)^\top \Big\|_2 + \\
&\qquad \Big\| \sum_{i\in\Lambda(l)} \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big(1 - \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big)\boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}^\top \Big\|_2 \Bigg] \qquad \text{(EC.18)} \\
&\le p_{l,t}(\boldsymbol{x}) \cdot \Bigg[\Big\| \Big(\sum_{i\in\Lambda(l)} \boldsymbol{w}_{i,t} \Big)\Big(\sum_{i\in\Lambda(l)} \boldsymbol{w}_{i,t} \Big)^\top \Big\|_2 + \frac{1}{4} \Big\| \sum_{i\in\Lambda(l)} \boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}^\top \Big\|_2 \Bigg] \\
&\le p_{l,t}(\boldsymbol{x}) \cdot \Bigg[\sum_{i\in\Lambda(l)} \left\|\boldsymbol{w}_{i,t}\right\|_2 \cdot \sum_{i\in\Lambda(l)} \left\|\boldsymbol{w}_{i,t}\right\|_2 + \frac{1}{4} \sum_{i\in\Lambda(l)} \left\|\boldsymbol{w}_{i,t}\right\|_2 \left\|\boldsymbol{w}_{i,t}\right\|_2 \Bigg] \\
&\le p_{l,t}(\boldsymbol{x}) \cdot \Bigg[(D_{\max}-1)^2 \cdot W^2_{\max} + \frac{1}{4}(D_{\max}-1) \cdot W^2_{\max} \Bigg],
\end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality is due to the range of $\psi_{i,t}$ and fundamental inequality, and the third inequality is due to both the triangle inequality and the property of the spectral norm for outer products. Thus, similar to Equation (EC.16), the upper bound of the gradient of $\nabla_{\boldsymbol{x}} f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is given by

$$\begin{aligned}
\|\nabla^2_{\boldsymbol{x}} f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 &= \Big\| \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} \nabla^2_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x}) \cdot \boldsymbol{\pi}_{l,t}^\top \Big\|_2 \\
&\le \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} \left\|\nabla^2_{\boldsymbol{x}} p_{l,t}(\boldsymbol{x})\right\|_2 \cdot \left\|\boldsymbol{\pi}_{l,t}\right\|_2 \\
&\le \frac{1}{T} \sum_{t=1}^{T} \sum_{l=1}^{2^{D(t)}} p_{l,t}(\boldsymbol{x}) \cdot \Big[(D_{\max}-1)^2 \cdot W^2_{\max} + \frac{1}{4}(D_{\max}-1) \cdot W^2_{\max} \Big] \cdot \Pi_{\max} \\
&= (D_{\max}-1)(D_{\max} - \frac{3}{4}) \cdot W^2_{\max} \cdot \Pi_{\max} \\
&= S^{\mathrm{SRF}},
\end{aligned}$$

where the second inequality is due to the definition of $\Pi_{\max}$ and Equation (EC.18). Therefore, $f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is S^{SRF}-Lipschitz smoothness. □

EC.1.7 Proof of Proposition 3 in Section 5

Proof of Proposition 3. Under Assumption 2(III), the function Ψ is bounded by compact set $[0, B]$. Next, we analyze the properties of functions t_3, t_2, and t_1 in sequence.

- **For function** t_3, we first define a vector-valued function

$$\boldsymbol{u}(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)=\left[\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_1+\boldsymbol{\xi}_2),\cdots,\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_{n_{\widehat{\boldsymbol{x}}}}+\boldsymbol{\xi}_2)\right]^\top.$$

For any $\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta$, we write $\boldsymbol{u}:=\boldsymbol{u}(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)$ and $\boldsymbol{u}':=\boldsymbol{u}(\boldsymbol{\theta}';\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)$ for brevity.

For each $i\in[n_{\widehat{\boldsymbol{x}}}]$, we have

$$\Big[t_3(\boldsymbol{u})\Big]_i'=\frac{1}{\lambda\epsilon}\exp(u_i/\lambda\epsilon)\le\frac{1}{\lambda\epsilon}\exp(B/\lambda\epsilon),$$

which implies that

$$\left|\Big[t_3(\boldsymbol{u})\Big]_i-\Big[t_3(\boldsymbol{u}')\Big]_i\right|\le L_3'\left|u_i-u_i'\right|,$$

where $L_3'=\frac{1}{\lambda\epsilon}\exp(B/\lambda\epsilon)$ and u_i,u_i' are the i-th elements of $\boldsymbol{u}$ and $\boldsymbol{u}'$, respectively. Hence, we obtain

$$\|t_3(\boldsymbol{u})-t_3(\boldsymbol{u}')\|_2^2=\sum_{i\in[n_{\widehat{\boldsymbol{x}}}]}\left|\Big[t_3(\boldsymbol{u})\Big]_i-\Big[t_3(\boldsymbol{u}')\Big]_i\right|^2\le(L_3')^2\sum_{i\in[n_{\widehat{\boldsymbol{x}}}]}\left|u_i-u_i'\right|^2=L_3^2\|\boldsymbol{u}-\boldsymbol{u}'\|_2^2,$$

where $L_3=L_3'=\frac{1}{\lambda\epsilon}\exp(B/\lambda\epsilon)$. This result shows that the function t_3 is L_3-Lipschitz continuous.

According to the second-order derivation of the function t_3

$$\Big[t_3(u)\Big]_i''=\frac{1}{(\lambda\epsilon)^2}\exp\left(\frac{u}{\lambda\epsilon}\right)\le\frac{1}{(\lambda\epsilon)^2}\exp\left(\frac{B}{\lambda\epsilon}\right),$$

we obtain

$$\left|\Big[t_3(\boldsymbol{u})\Big]_i'-\Big[t_3(\boldsymbol{u}')\Big]_i'\right|\le S_3\left|u_i-u_i'\right|,$$

where $S_3=\frac{1}{(\lambda\epsilon)^2}\exp\left(\frac{B}{\lambda\epsilon}\right)$. Since the Jacobian matrix of function t_3 for any $\boldsymbol{u}$, i.e. $J_{t_3}(\boldsymbol{u})$, is a diagonal matrix, we have

$$\|J_{t_3}(\boldsymbol{u})-J_{t_3}(\boldsymbol{u}')\|_{\mathrm{op}}=\max_{i\in[n_{\widehat{\boldsymbol{x}}}]}\left|\Big[t_3(\boldsymbol{u})\Big]_i'-\Big[t_3(\boldsymbol{u}')\Big]_i'\right|\le\max_{i\in[n_{\widehat{\boldsymbol{x}}}]}S_3\left|u_i-u_i'\right|\le S_3\|\boldsymbol{u}-\boldsymbol{u}'\|_2,$$

Therefore, the function t_3 is S_3-Lipschitz smooth with $S_3=\frac{1}{(\lambda\epsilon)^2}\exp\left(\frac{B}{\lambda\epsilon}\right)$.

Using the chain rule, we have

$$\nabla[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)]_i=\Big[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\Big]_i\cdot\frac{1}{\lambda\epsilon}\cdot\nabla\left(\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_i+\boldsymbol{\xi}_2)\right).$$

According to Assumption 2(II) and 2(III), we obtain

$$\begin{aligned}\Big\|\nabla[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)]_i\Big\|_2&=\left\|\frac{1}{\lambda\epsilon}\cdot\Big[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\Big]_i\cdot\nabla\left(\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_i+\boldsymbol{\xi}_2)\right)\right\|_2\\&=\frac{1}{\lambda\epsilon}\cdot\Big[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\Big]_i\cdot\|\nabla L(\boldsymbol{\theta};\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}}_i+\boldsymbol{\xi}_2)\|_2\\&\le L_3\cdot L_{\boldsymbol{\theta}}\end{aligned}$$

where the inequality is due to $L_3 = \frac{1}{\lambda\epsilon}\exp(B/\lambda\epsilon)$ and $\|L(\boldsymbol{\theta}_1;\boldsymbol{x},\boldsymbol{y}) - L(\boldsymbol{\theta}_2;\boldsymbol{x},\boldsymbol{y})\|_2 \le L_{\boldsymbol{\theta}}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$. Therefore, we have

$$\begin{aligned}\mathbb{E}\Big[\|\nabla t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\|_2^2\Big] &= \mathbb{E}\Big[\sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}}\Big\|\nabla[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)]_i\Big\|_2^2\Big] \\ &\le \mathbb{E}\Big[n_{\widehat{\boldsymbol{x}}}\cdot L_3^2 L_{\boldsymbol{\theta}}^2\Big] \\ &\le L_3^2 L_{\boldsymbol{\theta}}^2\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[n_{\widehat{\boldsymbol{x}}}\Big].\end{aligned}$$

Let $C_3^2 = L_3^2 L_{\boldsymbol{\theta}}^2\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[n_{\widehat{\boldsymbol{x}}}\Big]$. Since the number of observations of $\widehat{\boldsymbol{y}}$, i.e., $\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[n_{\widehat{\boldsymbol{x}}}\Big]$, is finite, we show that the the stochastic gradients in expectation function t_3 is bounded, i.e., $\mathbb{E}\Big[\|\nabla t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\|_2^2\Big] \le C_3^2$.

According to the fundamental result in Casella and Berger (2024), for a bounded random variable (vector), its variance is finite. Thus, to prove the finite variance of the function t_3, it suffices to show that the norm of the function t_3 is bounded. According to Assumption 2(III), we have

$$\|t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\|_2^2 = \sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}}\Big(\Big[t_3(\boldsymbol{u})\Big]_i\Big)^2 \le n_{\widehat{\boldsymbol{x}}}\exp\Big(2B/\lambda\epsilon\Big),$$

and it follows that $\sigma_3^2 = \sup_{\boldsymbol{\theta}\in\Theta,\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}}}\mathbb{V}_{\boldsymbol{\xi}_2}\Big(t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2\Big)\Big) < \infty$.

- **For function** t_2, according to the analyses for function t_3, its domain is also bounded by $\boldsymbol{v}\in\Big[1,\exp(B/\lambda\epsilon)\Big]^{n_{\widehat{\boldsymbol{x}}}}$. For brevity, we denote $t_2(\boldsymbol{v};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)$ as $t_2(\boldsymbol{v})$, and denote $\widehat{p}(\widehat{\boldsymbol{y}}_i\mid\widehat{\boldsymbol{x}})$ as p_i for each $i\in[n_{\widehat{\boldsymbol{x}}}]$. Since $\sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}} p_i = 1$, we have

$$t_2(\boldsymbol{v}) = \exp\Big(\sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}} p_i\cdot\log(v_i)\Big)\in\Big[1,\exp(B/\lambda\epsilon)\Big].$$

Therefore, we obtain

$$\begin{aligned}\|\nabla t_2(\boldsymbol{v})\|_2^2 &= \sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}}\Big(\frac{\partial t_2(\boldsymbol{v})}{\partial v_i}\Big)^2 \\ &= \sum_{i=1}^{n_{\widehat{\boldsymbol{x}}}}\Big(t_2(\boldsymbol{v})\cdot\frac{p_i}{v_i}\Big)^2 \\ &\le \exp\Big(2B/\lambda\epsilon\Big).\end{aligned} \tag{EC.19}$$

where the inequality is due to the domain and range of the function t_2 $\sum_{j=1}^{n_{\widehat{\boldsymbol{x}}}} p_j^2 \le 1$. Let $L_2 = C_2 = \exp\Big(B/\lambda\epsilon\Big)$. From Equation (EC.19), the function t_2 is $L2$-Lipschitz continuous, and have bounded stochastic gradients in expectation, i.e., $\mathbb{E}\Big[\|\nabla t_2(\boldsymbol{v};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)\|_2^2\Big] \le C_2^2$, and its variance is also finite, i.e., $\sigma_2^2 = \sup_{\boldsymbol{\theta}\in\Theta,\widehat{\boldsymbol{x}}}\mathbb{V}_{\boldsymbol{\xi}_1}\Big(t_2\Big(\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\Big[t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2\Big)\Big];\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1\Big)\Big) < \infty$.

We now show that the differentiable function t_2 is Lipschitz smooth. By Taylor's theorem, for any $\boldsymbol{v}_1, \boldsymbol{v}_2$ in the domain of t_2, there exists a point $\boldsymbol{c}$ on the line segment connecting $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ such that

$$t_2(\boldsymbol{v}_1) = t_2(\boldsymbol{v}_2) + \nabla t_2(\boldsymbol{v}_2)^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2) + \frac{1}{2}(\boldsymbol{v}_1 - \boldsymbol{v}_2)^\top \nabla^2 t_2(\boldsymbol{c})(\boldsymbol{v}_1 - \boldsymbol{v}_2).$$

Thus, we have

$$\begin{aligned} \left| t_2(\boldsymbol{v}_1) - t_2(\boldsymbol{v}_2) - \nabla t_2(\boldsymbol{v}_2)^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2) \right| &= \left| \frac{1}{2} \cdot (\boldsymbol{v}_1 - \boldsymbol{v}_2)^\top \nabla^2 t_2(\boldsymbol{c})(\boldsymbol{v}_1 - \boldsymbol{v}_2) \right| \\ &\leq \frac{1}{2} \|\nabla^2 t_2(\boldsymbol{c})\|_{\text{op}} \cdot \|(\boldsymbol{v}_1 - \boldsymbol{v}_2)\|_2^2, \end{aligned} \tag{EC.20}$$

where the inequality is due to the Cauchy-Schwarz inequality. According to the definition of Lipschitz smoothness, if $\|\nabla^2 t_2(\boldsymbol{c})\|_{\text{op}}$ is bounded by a constant S_2, i.e., $\|\nabla^2 t_2(\boldsymbol{c})\|_{\text{op}} \leq S_2$, then the function t_2 is S_2-Lipschitz smooth. Denote the j-th element in vector $\boldsymbol{c}$ as c_j for any $j \in [n_{\widehat{\boldsymbol{x}}}]$. For any $j, k \in [n_{\widehat{\boldsymbol{x}}}]$, the (j,k)-th element of the Hessian Matrix $\nabla^2 t_2(\boldsymbol{c})$ is given by

$$(\nabla^2 t_2(\boldsymbol{c}))_{jk} = \frac{\partial^2 t_2(\boldsymbol{c})}{\partial c_k \partial c_j} = \begin{cases} t_2(\boldsymbol{c}) \cdot \frac{p_j p_k}{c_j c_k}, & \text{if } j \neq k; \\ t_2(\boldsymbol{c}) \cdot \left(\frac{p_j^2}{c_j^2} - \frac{p_j}{c_j^2} \right), & \text{if } j = k. \end{cases}$$

Therefore, we have

$$\begin{aligned} \|\nabla^2 t_2(\boldsymbol{c})\|_{\text{op}}^2 &\leq \|\nabla^2 t_2(\boldsymbol{c})\|_{\text{F}}^2 \\ &= \sum_{j=1}^{n_{\widehat{\boldsymbol{x}}}} \left(t_2(\boldsymbol{c}) \left(\frac{p_j^2}{c_j^2} - \frac{p_j}{c_j^2} \right) \right)^2 + \sum_{j \neq k} \left(t_2(\boldsymbol{c}) \frac{p_j p_k}{c_j c_k} \right)^2 \\ &\leq \sum_{j=1}^{n_{\widehat{\boldsymbol{x}}}} \left(e^{\frac{B}{\lambda \epsilon}} \left(\frac{p_j - p_j^2}{1} \right) \right)^2 + \sum_{j \neq k} \left(e^{\frac{B}{\lambda \epsilon}} \frac{p_j p_k}{1} \right)^2 \\ &\leq e^{\frac{2B}{\lambda \epsilon}} \left[\sum_{j=1}^{n_{\widehat{\boldsymbol{x}}}} p_j^2 + \left(\sum_{j=1}^{n_{\widehat{\boldsymbol{x}}}} p_j^2 \right)^2 \right] \\ &\leq 2 \cdot e^{\frac{2B}{\lambda \epsilon}}, \end{aligned}$$

where the first inequality is because the Frobenius norm $\|\cdot\|_{\text{F}}$ is a upper bound of the operator norm for a matrix, the second inequality is due to the domain and range of function t_2, the third inequality is due to $p_j \in [0,1]$ for any $j \in [n_{\widehat{\boldsymbol{x}}}]$ and the Cauchy-Schwarz inequality, and the last inequality is due to $\sum_{j=1}^{n_{\widehat{\boldsymbol{x}}}} p_j^2 \leq 1$. Let $S_2 = \sqrt{2}\exp(B/\lambda\epsilon)$, based on relation (EC.20), we have

$$\left| t_2(\boldsymbol{v}_1) - t_2(\boldsymbol{v}_2) - \nabla t_2(\boldsymbol{v}_2)^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2) \right| \leq \frac{1}{2} \|\nabla^2 t_2(\boldsymbol{c})\|_{\text{op}} \cdot \|(\boldsymbol{v}_1 - \boldsymbol{v}_2)\|_2^2 \leq \frac{S_2}{2} \cdot \|(\boldsymbol{v}_1 - \boldsymbol{v}_2)\|_2^2,$$

which implies that the function t_2 is S_2-Lipschitz smooth.

- **For function** t_1, its domain is the range of function t_2, i.e., $\left[1, \exp(B/\lambda\epsilon)\right]$. Thus, the range of function t_1 is $\left[0, B/\lambda\epsilon\right]$. For brevity, we denote $t_1(z; \widehat{\boldsymbol{x}})$ as $t_1(z)$. The first and second

derivatives of t_1 with respect to z are

$$t_1'(z) = \frac{1}{z} \quad \text{and} \quad t_1''(z) = -\frac{1}{z^2}.$$

Over the domain $z \in \left[1, \exp(B/\lambda\epsilon)\right]$, we can bound the absolute values of these derivatives:

$$|t_1'(z)| = \frac{1}{z} \le 1,$$

$$|t_1''(z)| = \frac{1}{z^2} \le 1.$$

The first bound implies that t_1 is L_1-Lipschitz continuous with $\boldsymbol{L_1 = 1}$. The second bound implies that t_1 is S_1-Lipschitz smooth with $\boldsymbol{S_1 = 1}$. Let $C_1 = 1$; the expected squared norm of the gradient is also bounded

$$\mathbb{E}\left[|\nabla t_1(z;\widehat{\boldsymbol{x}})|^2\right] = \mathbb{E}\left[\left(t_1'(z)\right)^2\right] \le C_1^2.$$

Since the value of the function t_1 is bounded, its variance is finite, i.e.,

$$\sigma_1^2 = \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{V}_{\widehat{\boldsymbol{x}}}\left(t_1\left(\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[t_3\left(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2\right)\right];\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1\right)\right];\widehat{\boldsymbol{x}}\right)\right) < \infty.$$

This completes the proof. □

EC.1.8 Proof of Theorem 3 in Section 5.2

As an essential part of sample complexity analysis, we first introduce the following lemmas. Based on the Cramér's large deviations theorem, we introduce the following Lemma EC.3.

Lemma EC.3. (Cramér's Inequality, Kleywegt et al., 2002). *Let $X_1,\ldots,X_n$ be i.i.d. samples of a zero-mean random variable X with finite variance σ^2. For any $\delta > 0$, it holds*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \ge \delta\right) \le \exp(-nI(\delta)),$$

where $I(\delta) := \sup_{t\in\mathbb{R}}\{t\delta - \log M(t)\}$ is the rate function of random variable X, and $M(t) := \mathbb{E}e^{tX}$ is the moment generating function of X. For any $\kappa > 0$, there exists $\delta_1 > 0$, for any $\delta \in (0,\delta_1)$, $I(\delta) \ge \frac{\delta^2}{(2+\kappa)\sigma^2}$.

Y. Hu et al. (2020) extend the Cramér's Inequality from random variables to random vectors, as shown in the following Lemma EC.4.

Lemma EC.4. (Concentration Inequality, Y. Hu et al., 2020). *Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N$ be i.i.d. samples of a zero-mean random vector $\boldsymbol{X} \in \mathbb{R}^k$ with finite variance $\mathbb{E}\|\boldsymbol{X}\|_2^2 = \sigma^2 < \infty$. Then for any $\kappa > 0$, there exists $\delta_1 > 0$ such that for any $\delta \in (0,\delta_1)$, it holds that*

$$\Pr\left(\left\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\right\|_2 \ge \delta\right) \le 2k\exp\left(-\frac{N\delta^2}{(2+\kappa)\sigma^2}\right).$$

Using Lemma EC.4, we present the following Lemma EC.5.

Lemma EC.5. *Under Assumption 2, for any $\kappa > 0$, there exists an $\delta_1 > 0$ such that for any $\delta \in (0, \delta_1)$, it holds that*

$$\begin{aligned}
&\Pr\left(\sup_{\boldsymbol{\theta}\in\Theta}\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right| > \delta\right) \\
\leq &\mathcal{O}(1)\left(\frac{4L_1L_2L_3D_\Theta}{\delta}\right)^{d_\theta}\left(N_1N_2n_{\widehat{\boldsymbol{x}}}\exp\left(-\frac{N_3\delta^2}{36(2+\kappa)\lambda^2\epsilon^2L_1^2L_2^2\sigma_3^2}\right)\right. \\
&\left.+N_1\exp\left(-\frac{N_2\delta^2}{36(2+\kappa)\lambda^2\epsilon^2L_1^2\sigma_2^2}\right)+\exp\left(-\frac{N_1\delta^2}{36(2+\kappa)\lambda^2\epsilon^2\sigma_1^2}\right)\right).
\end{aligned}$$

Proof of Lemma EC.5. For $v \in (0,1)$, the set $\{\boldsymbol{x}_l\}_{l=1}^Q$ is said to be a v-net of $\mathcal{X}$, if $\boldsymbol{x}_l \in \mathcal{X}$, $\forall l = 1,\ldots,Q$, and the following holds: $\forall \boldsymbol{x} \in \mathcal{X}, \exists l(\boldsymbol{x}) \in \{1,\ldots,Q\}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_{l(\boldsymbol{x})}\|_2 \leq v$. We construct a v-net to get rid of the supremum over $\boldsymbol{\theta}$ and use a concentration inequality to bound the probability. First, we pick a v-net $\{\boldsymbol{\theta}_l\}_{l=1}^Q$ on the decision set $\Theta \in \mathbb{R}^{d_\theta}$, such that $L_1L_2L_3v = \delta/4$. Under Assumption 2(I), Θ has a finite diameter D_Θ, for any $v \in (0,1)$, there exists a v-net of Θ, and the size of the v-net is bounded, $Q \leq \mathcal{O}((D_\times/v)^{d_\theta})$ (Shapiro et al., 2021). By definition of v-net, we have $\forall \boldsymbol{\theta} \in \Theta, \exists l(\boldsymbol{\theta}) \in \{1,2,\ldots,Q\}$, s.t.

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_{l(\boldsymbol{\theta})}\|_2 \leq v = \frac{\delta}{4L_1L_2L_3}.$$

Based on Proposition 3, we have

$$\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}) - \widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right)\right| \leq L_1L_2L_3\|\boldsymbol{\theta} - \boldsymbol{\theta}_{l(\boldsymbol{\theta})}\|_2 \leq \frac{\delta}{4},$$

and

$$\left|F\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right) - F(\boldsymbol{\theta})\right| \leq L_1L_2L_3\|\boldsymbol{\theta} - \boldsymbol{\theta}_{l(\boldsymbol{\theta})}\|_2 \leq \frac{\delta}{4}.$$

Thus, for any $\boldsymbol{\theta} \in \Theta$, we have

$$\begin{aligned}
&\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\right| \\
\leq &\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}) - \widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right)\right| + \left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right) - F\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right)\right| + \left|F\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right) - F(\boldsymbol{\theta})\right| \\
\leq &\frac{\delta}{2} + \left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right) - F\left(\boldsymbol{\theta}_{l(\boldsymbol{\theta})}\right)\right| \\
\leq &\frac{\delta}{2} + \max_{l=1,\cdots,Q}\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}_l) - F(\boldsymbol{\theta}_l)\right| \\
\leq &\frac{\delta}{2} + \sum_{l=1}^{Q}\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}_l) - F(\boldsymbol{\theta}_l)\right|
\end{aligned}$$

It follows that

$$
\begin{aligned}
&\Pr\left(\sup_{\boldsymbol{\theta}\in\Theta}\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta})-F(\boldsymbol{\theta})\right|>\delta\right)\\
\le&\Pr\left(\sum_{l=1}^{Q}\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}_l)-F(\boldsymbol{\theta}_l)\right|>\frac{\delta}{2}\right)\\
\le&\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}_l)-F(\boldsymbol{\theta}_l)\right|>\frac{\delta}{2}\right)\\
\le&\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}_l)-\widehat{F}_{N_1,N_2}(\boldsymbol{\theta}_l)\right|+\left|\widehat{F}_{N_1,N_2}(\boldsymbol{\theta}_l)-\widehat{F}_{N_1}(\boldsymbol{\theta}_l)\right|+\left|\widehat{F}_{N_1}(\boldsymbol{\theta}_l)-F(\boldsymbol{\theta}_l)\right|>\frac{\delta}{2}\right)\\
\le&\underbrace{\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}_l)-\widehat{F}_{N_1,N_2}(\boldsymbol{\theta}_l)\right|>\frac{\delta}{6}\right)}_{\Delta_1}+\underbrace{\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2}(\boldsymbol{\theta}_l)-\widehat{F}_{N_1}(\boldsymbol{\theta}_l)\right|>\frac{\delta}{6}\right)}_{\Delta_2}\\
&+\underbrace{\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1}(\boldsymbol{\theta}_l)-F(\boldsymbol{\theta}_l)\right|>\frac{\delta}{6}\right)}_{\Delta_3}.
\end{aligned}
\tag{EC.21}
$$

For the term Δ_1 in Equation (EC.21), we have

$$
\begin{aligned}
\Delta_1=&\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}_l)-\widehat{F}_{N_1,N_2}(\boldsymbol{\theta}_l)\right|>\frac{\delta}{6}\right)\\
=&\sum_{l=1}^{Q}\Pr\Bigg(\Bigg|\frac{\lambda\epsilon}{N_1}\sum_{i=1}^{N_1}t_1\Bigg(\frac{1}{N_2}\sum_{j=1}^{N_2}t_2\Bigg(\frac{1}{N_3}\sum_{k=1}^{N_3}t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\widehat{\boldsymbol{y}}^i,\boldsymbol{\xi}_2^k\Big);\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j\Bigg)\Bigg)\\
&\qquad-\frac{\lambda\epsilon}{N_1}\sum_{i=1}^{N_1}t_1\Bigg(\frac{1}{N_2}\sum_{j=1}^{N_2}t_2\Big(\mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\widehat{\boldsymbol{y}}^i,\boldsymbol{\xi}_2^k\Big)\Big];\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j\Big)\Bigg)(\boldsymbol{\theta}_l)\Bigg|>\frac{\delta}{6}\Bigg)\\
\le&\sum_{l=1}^{Q}\Pr\left(\max_{i=1,\cdots,N_1;j=1,\cdots,N_2}\left|L_1L_2\|\frac{1}{N_3}\sum_{k=1}^{N_3}t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\widehat{\boldsymbol{y}}^i,\boldsymbol{\xi}_2^k\Big)-\mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\widehat{\boldsymbol{y}}^i,\boldsymbol{\xi}_2^k\Big)\Big]\|_2\right|>\frac{\delta}{6\lambda\epsilon}\right)\\
\le&\sum_{l=1}^{Q}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\Pr\left(\|\frac{1}{N_3}\sum_{k=1}^{N_3}t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\widehat{\boldsymbol{y}}^i,\boldsymbol{\xi}_2^k\Big)-\mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\widehat{\boldsymbol{y}}^i,\boldsymbol{\xi}_2^k\Big)\Big]\|_2>\frac{\delta}{6\lambda\epsilon L_1L_2}\right)\\
\le&QN_1N_2\cdot 2n_{\widehat{\boldsymbol{x}}}\exp\Big(-\frac{N_3\delta^2}{36(2+\kappa)\lambda^2\epsilon^2L_1^2L_2^2\sigma_3^2}\Big),
\end{aligned}
\tag{EC.22}
$$

where the first inequality is due to the Lipschitz continuity, and the last inequality is due to Lemma EC.4. Similarly, we obtain

$$
\Delta_2\le QN_1\cdot 2\exp\Big(-\frac{N_2\delta^2}{36(2+\kappa)\lambda^2\epsilon^2L_1^2\sigma_2^2}\Big),
\tag{EC.23}
$$

and

$$\Delta_3 \le Q \cdot 2\exp\Big(-\frac{N_1\delta^2}{36(2+\kappa)\lambda^2\epsilon^2\sigma_1^2}\Big). \tag{EC.24}$$

Combining with Equations (EC.21)- (EC.24), and the fact that $Q \le \mathcal{O}(1)(4L_1L_2L_3D_\Theta/\delta)^{d_\theta}$, we can obtain the desired result of Lemma EC.5. □

In the following, we prove the results in Theorem 3.

Proof of Theorem 3. (I) For $\Pr\Big(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-F(\boldsymbol{\theta}^*)\le\delta\Big)$ in Theorem 3(I), we have

$$\begin{aligned}
&\Pr\Big(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-F(\boldsymbol{\theta}^*)>\delta\Big)\\
=&\Pr\Big(\Big[F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-\widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)\Big]\\
&\quad+\Big[\widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*)\Big]+\Big[\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*)-F(\boldsymbol{\theta}^*)\Big]>\delta\Big)\\
\le&\Pr\Big(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-\widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)>\frac{\delta}{2}\Big)+\Pr\Big(\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*)-F(\boldsymbol{\theta}^*)>\frac{\delta}{2}\Big)\\
\le&\Pr\Big(\Big|F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-\widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)\Big|>\frac{\delta}{2}\Big)+\Pr\Big(\Big|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*)-F(\boldsymbol{\theta}^*)\Big|>\frac{\delta}{2}\Big)
\end{aligned} \tag{EC.25}$$

where the first inequality is due to $\widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*)\le 0$. Using the result of Lemma EC.5, we can obtain the desired result of Theorem 3(I) using Equation (EC.25).

(II) For Theorem 3(II), to analyze

$$\Pr\Big(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-F(\boldsymbol{\theta}^*)\le\delta\Big)\ge 1-\alpha,$$

it suffices to study

$$\Pr\Big(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)-F(\boldsymbol{\theta}^*)>\delta\Big)<\alpha.$$

Let each of the three terms on the right-hand side (RHS) of the inequality in Theorem 3(I) be no more than $\alpha/3$. This leads to

$$\mathcal{O}(1)\Big(\frac{8L_1L_2L_3D_\Theta}{\delta}\Big)^{d_\theta}\exp\Big(-\frac{N_1\delta^2}{144(2+\kappa)\lambda^2\epsilon^2\sigma_1^2}\Big)\Big)<\frac{\alpha}{3},$$

and then we obtain the necessary sample size from distribution $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$

$$N_1>\frac{\mathcal{O}(1)\sigma_1^2}{\delta^2}\Big[d_\theta\log\Big(\frac{8L_1L_2L_3D_\Theta}{\delta}\Big)+\log\Big(\frac{1}{\alpha}\Big)\Big].$$

Similarly, we can obtain the desired result of N_2 and N_3.

Ignoring the log factors, the required sample sizes N_1, N_2, and N_3 are all of order $\mathcal{O}\Big(d_\theta/\delta^2\Big)$. Therefore, the total sample complexity of problem (SAA) for achieving a δ-optimal solution is $T=N_1+N_2+N_3=\mathcal{O}\Big(d_\theta/\delta^2\Big)$.

□

EC.1.9 Proof of Theorem 4 in Section 5.3

Proof of Theorem 4. According to the Theorem 3 in Chen et al. (2021), Theorem 4(I) holds.

For Theorem 4(II), to ensure that $\mathbb{E}\left[\|\nabla F(\widehat{\boldsymbol{\theta}})\|^2\right] \leq \varepsilon^2$, it follows that (according to Chen et al., 2021)

$$\frac{\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\boldsymbol{\theta}^k)\|^2\right]}{K} \leq \frac{C_{\text{const}}}{\sqrt{K}} \leq \varepsilon^2,$$

where C_{const} is a constant that depends on the initial setting of the algorithm and constants $C_1, C_2, C_3, S_1, S_2, S_3$. This implies that the number of iterations required satisfies

$$K \geq \frac{C_{\text{const}}^2}{\varepsilon^4} = \mathcal{O}(\varepsilon^{-4}).$$

Since one sample is drawn from each of the three distributions in each iteration, the total number of samples is $3 \cdot K$, which implies that the sampling complexity of the SCSC method is also at the order of $\mathcal{O}(\varepsilon^{-4})$.

In each iteration, we perform one gradient calculation on each of the functions t_1, t_2, and t_3. Thus, each function performs a total of K gradient calculations, which implies that their gradient complexities are the same at the order of $\mathcal{O}(\varepsilon^{-4})$.

For the classical stochastic nonconvex optimization problem, the complexity bounds of SCSC match the existing lower bounds by Arjevani et al. (2023), i.e., $\mathcal{O}(\varepsilon^{-4})$. □

EC.2 Worst-case Distributions for Compared DRO Models

In this section, we show the models and worst-case distribution formulations of Sinkhorn DRO (SDRO), causal Wasserstein DRO (Causal-WDRO), and KL-divergence-based DRO (KL-DRO) in soft-constrained and contextual settings.

Based on Wang et al. (2025), the soft-constrained SDRO without causal consideration for contextual DRO is defined as

$$\inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\left[\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda\cdot\mathcal{W}_p(\widehat{\mathbb{P}},\mathbb{P})^p\right], \tag{SDRO}$$

where

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q}) := \left(\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\left[c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\right] + \epsilon\cdot H\left(\gamma \mid \mu\otimes\left(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}\right)\right)\right)^{1/p}.$$

We present the worst-case distribution of problem (SDRO) in the following Lemma EC.6 by extending the results in Wang et al. (2025).

Lemma EC.6. (Worst-case Distribution of the SDRO Problem, Wang et al., 2025). *Under Assumption 1, the density of worst-case distribution $\mathbb{P}^*_{\lambda,\text{SDRO}}$ of the inner problem of (SDRO) for any λ is given by*

$$\frac{\mathrm{d}\mathbb{P}^*_{\lambda,\text{SDRO}}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\left[\widetilde{\alpha}_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}}(\lambda)\cdot e^{s'(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\right], \tag{EC.26}$$

where $\widetilde{\alpha}_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}}(\lambda) = \left(\int_{\mathcal{X}\times\mathcal{Y}} e^{s'(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\cdot\mathrm{d}\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}(\boldsymbol{x},\boldsymbol{y})\right)^{-1}$.

According to Lemma EC.6, $\mathbb{P}^*_{\lambda,\text{SDRO}}$ is a mixture of Gibbs distributions. Compared with the worst-case distribution of problem (Causal-SDRO) in Theorem 2, $\mathbb{P}^*_{\lambda,\text{SDRO}}$ has a simpler density function structure.

Without considering the entropy regularization, the Causal-WDRO model is defined as Yang et al. (2022):

$$\inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda\cdot C_p(\widehat{\mathbb{P}},\mathbb{P})^p\Big], \tag{Causal-WDRO}$$

where the causal transport distance $C_p(\widehat{\mathbb{P}},\mathbb{P})$ is defined in Definition 1. Yang et al. (2022) characterize the worst-case distribution of the Causal-WDRO problem. In the following, suppose that the empirical distribution $\widehat{\mathbb{P}}$ is grouped into K distinct covariates $\widehat{\boldsymbol{x}}_k$ for any $k\in[K]$. For each covariate, there are n_k observations of the uncertain parameter, denoted by $\widehat{\boldsymbol{y}}_{ki}$ for any $i\in[n_k]$. Let $\widehat{p}_{ki}$ be the probability mass of the data point $(\widehat{\boldsymbol{x}}_k,\widehat{\boldsymbol{y}}_{ki})$. Then, according to Yang et al. (2022), the following lemma holds.

Lemma EC.7. (Worst-case Distribution of the Causal-WDRO Problem, Yang et al., 2022). *If the worst-case distribution of Causal-WDRO problem exists for given $\lambda>0$, then it has the following form*

$$\mathbb{P}^*_{\lambda,\text{Causal-WDRO}} = \sum_{k\neq k_0}\sum_{i=1}^{n_k}\widehat{p}_{ki}\widehat{\mathbb{P}}_{(\boldsymbol{x}^*_k(\lambda),\boldsymbol{y}^*_{ki}(\lambda))} + \sum_{i=1}^{n_{k_0}}\widehat{p}_{k_0 i}\left(q\widehat{\mathbb{P}}_{(\bar{\boldsymbol{x}}_{k_0}(\lambda),\bar{\boldsymbol{y}}_{k_0 i}(\lambda))} + (1-q)\widehat{\mathbb{P}}_{(\underline{\boldsymbol{x}}_{k_0}(\lambda),\underline{\boldsymbol{y}}_{k_0 i}(\lambda))}\right), \tag{EC.27}$$

where $1\le k_0\le K$, $0\le q\le 1$, $(\boldsymbol{x}^*_k(\lambda),\boldsymbol{y}^*_{ki}(\lambda)) = (\bar{\boldsymbol{x}}_k(\lambda),\bar{\boldsymbol{y}}_{ki}(\lambda))$, and for every k and i,

$$\bar{\boldsymbol{x}}_k(\lambda),\underline{\boldsymbol{x}}_k(\lambda)\in\arg\max_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\boldsymbol{Y}|\boldsymbol{X}}}\left[\sup_{\boldsymbol{y}\in\mathcal{Y}}\left\{\Psi(f(\boldsymbol{x}),\boldsymbol{y})-\lambda\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^p\right\}\mid\boldsymbol{X}=\widehat{\boldsymbol{x}}_k\right]-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_k\|^p\right\},$$

and

$$\bar{\boldsymbol{y}}_{ki}(\lambda)\in\arg\max_{\boldsymbol{y}\in\mathcal{Y}}\left\{\Psi(f(\bar{\boldsymbol{x}}_k),\boldsymbol{y})-\lambda\|\boldsymbol{y}-\widehat{\boldsymbol{y}}_{ki}\|^p\right\},\quad \underline{\boldsymbol{y}}_{ki}(\lambda)\in\arg\max_{\boldsymbol{y}\in\mathcal{Y}}\left\{\Psi(f(\underline{\boldsymbol{x}}_k),\boldsymbol{y})-\lambda\|\boldsymbol{y}-\widehat{\boldsymbol{y}}_{ki}\|^p\right\}.$$

The worst-case distribution of problem (Causal-WDRO) in Equation (EC.27) is discrete, while that of (Causal-SDRO) in Equation (9) is continuous. This shows that the introduction of CSD allows a more realistic and smoother representation of the underlying distribution.

The contextual KL-divergence-based DRO (KL-DRO) model is defined as

$$\inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda\cdot\mathbb{D}_{\text{KL}}\Big(\mathbb{P}\|\widehat{\mathbb{P}}\Big)\Big]. \tag{KL-DRO}$$

For problem (KL-DRO), its worst-case distribution is given by the following Lemma EC.8.

Lemma EC.8. (Worst-case Distribution of the KL-DRO Problem, Z. Hu and Hong, 2013). *If the worst-case distribution of KL-DRO problem exists for given $\lambda>0$, then it has the following form*

$$\mathbb{P}^*_{\lambda,\text{KL-DRO}} = \sum_{i=1}^{N}\frac{\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}_i),\widehat{\boldsymbol{y}}_i)}{\lambda}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}_j),\widehat{\boldsymbol{y}}_j)}{\lambda}\right)}\cdot\widehat{\mathbb{P}}_{(\widehat{\boldsymbol{x}}_i,\widehat{\boldsymbol{y}}_i)}, \tag{EC.28}$$

where N is the number of historical observations.

In fact, KL-DRO finds the worst-case distribution by changing the likelihood ratios of the empirical distribution rather than changing its support. Therefore, the worst-case distribution of (KL-DRO) is still discrete.

EC.3 Interpretability Measures

In this section, we analyze the intrinsic interpretability of the SRT by introducing global and local interpretation measures in Sections EC.3.1 and EC.3.2, respectively.

EC.3.1 Global Interpretation Measure

Global interpretability provides a holistic view of the model by quantifying the contribution of each feature to the overall decision-making process (Dwivedi et al., 2023). For tree-based methods, standard techniques include impurity-based and permutation-based importance measures (Hastie et al., 2009; Kallus and Mao, 2023). However, impurity-based measures are designed for trees with hard splits and are inapplicable to SRTs, while permutation-based methods are often computationally expensive. Therefore, leveraging the differentiability of the SRT, we define feature importance based on the average marginal sensitivity of the decision with respect to each feature over the training set:

$$\mathcal{C}_j := \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x}^i)}{\partial x_j^i}\right\|_1 = \frac{1}{N\cdot T}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{2^{D(t)}}\left\|\frac{\partial p_{l,t}(\boldsymbol{x}^i)}{\partial x_j^i}\boldsymbol{\pi}_{l,t}\right\|_1, \quad \forall j\in[d_x], \tag{EC.29}$$

where $\boldsymbol{x}^i$ is the i-th training sample and x_j^i is its j-th element, and the partial derivatives of $p_{l,t}(\boldsymbol{x}^i)$ are computed following Proposition 1. To show the relative importance of features, we normalize the importance scores $\mathcal{C}_j$ for each $j\in[d_x]$ such that they sum to 1, i.e., the relative feature importance is given by

$$\bar{\mathcal{C}}_j := \frac{\mathcal{C}_j}{\sum_{k=1}^{d_x}\mathcal{C}_k}, \quad \forall j\in[d_x].$$

EC.3.2 Local Interpretation Measure

Local interpretability demonstrates how an individual decision is derived, clarifying the contribution of specific features and their interactions (Dwivedi et al., 2023; Notz and Pibernik, 2024). Lundberg et al. (2020) propose SHAP (SHapley Additive exPlanations) as a post-hoc local explainer to characterize feature contributions. While SHAP enhances the transparency of inherently uninterpretable models, it can be used only after the decision is made and does not exploit the model's structure (i.e., model-agnostic).

In contrast, given that the SRF is intrinsically interpretable and differentiable, we propose a novel metric, the Empirical Integrated Gradient (EIG). Adapted from the Integrated Gradients (IG) method of Sundararajan et al. (2017), EIG derives feature contributions directly from the model's structure. We validate the consistency of our intrinsic interpretability by comparing EIG with SHAP values.

According to Ancona et al. (2017) and Notz and Pibernik (2024), given input covariate $\boldsymbol{x}$, the k-th decision of the SRF can be decomposed into the sum of feature contributions relative to a baseline:

$$\left[f_{\boldsymbol{\theta}}^{\mathrm{SRF}}(\boldsymbol{x})\right]_k - \mathrm{Baseline}_k = \sum_{j=1}^{d_x}\varphi_{j,k}(\boldsymbol{\theta},\boldsymbol{x}).$$

Distinct from Sundararajan et al. (2017) and Ancona et al. (2017), which typically use a zero baseline, we employ the average decision over the training set as a robust baseline. Then, we define the EIG for feature j and decision k as

$$\varphi^{\mathrm{EIG}}_{j,k}(\boldsymbol{\theta},\boldsymbol{x}) := \frac{1}{N}\cdot\sum_{i=1}^{N}(x_j - x_j^i)\cdot\int_{\alpha=0}^{1}\frac{\partial\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}\left(\boldsymbol{x}^i+\alpha(\boldsymbol{x}-\boldsymbol{x}^i)\right)\right]_k}{\partial x_j}\mathrm{d}\alpha,\quad \forall j\in[d_x],k\in[d_z],$$

where $\boldsymbol{x}^i$ denotes the i-th training sample. Proposition EC.2 shows that the k-th output of SRF can be decomposed into the sum of EIG contributions from each feature, using as a baseline the average value of the k-th output over the dataset, i.e.,

$$\mathrm{Baseline}_k = \frac{1}{N}\sum_{i=1}^{N}\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x}^i)\right]_k.$$

Proposition EC.2. *Since the decision rule $f^{\mathrm{SRF}}_{\boldsymbol{\theta}}:\mathbb{R}^{d_x}\to\mathbb{R}^{d_z}$ is differentiable, given an input covariate $\boldsymbol{x}$, the EIG value $\varphi^{\mathrm{EIG}}_{j,k}(\boldsymbol{\theta},\boldsymbol{x})$ exactly quantifies the contribution of feature j to the deviation of the decision from the average training decision baseline, i.e.,*

$$\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x})\right]_k - \frac{1}{N}\sum_{i=1}^{N}\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x}^i)\right]_k = \sum_{j=1}^{d_x}\varphi^{\mathrm{EIG}}_{j,k}(\boldsymbol{\theta},\boldsymbol{x}),\quad \forall k\in[d_z].$$

Proof of Proposition EC.2. According to the Proposition 1 in Sundararajan et al. (Sundararajan et al., 2017), for any $\boldsymbol{x}'\in\mathbb{R}^{d_x}$, we have

$$\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x})\right]_k - \left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x}')\right]_k = \sum_{j=1}^{d_x}(x_j - x_j')\cdot\int_{\alpha=0}^{1}\frac{\partial\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}\left(\boldsymbol{x}'+\alpha(\boldsymbol{x}-\boldsymbol{x}')\right)\right]_k}{\partial x_j}\mathrm{d}\alpha,\quad \forall k\in[d_z].$$

By setting the baseline as the average decision over the training set, we obtain

$$\begin{aligned}\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x})\right]_k - \frac{1}{N}\sum_{i=1}^{N}\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}(\boldsymbol{x}^i)\right]_k &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d_x}(x_j - x_j^i)\cdot\int_{\alpha=0}^{1}\frac{\partial\left[f^{\mathrm{SRF}}_{\boldsymbol{\theta}}\left(\boldsymbol{x}^i+\alpha(\boldsymbol{x}-\boldsymbol{x}^i)\right)\right]_k}{\partial x_j}\mathrm{d}\alpha\\ &= \sum_{j=1}^{d_x}\varphi^{\mathrm{EIG}}_{j,k}(\boldsymbol{\theta},\boldsymbol{x}),\qquad \forall k\in[d_z].\end{aligned}$$

□

Regarding local feature interactions, the differentiable nature of the SRT allows us to explicitly characterize interaction effects by computing the Hessian matrix, as detailed in Proposition 1.

EC.4 Equivalent Reformulation for Feature-based Inventory Substitution Problem in Section 6.2

The feature-based inventory substitution problem with soft CSD constraint is given by

$$\inf_{f\in\mathcal{F}}\max_{\mathbb{P}\in\mathfrak{P}(\widehat{\mathbb{P}})}\quad \boldsymbol{c}^\top f(\boldsymbol{x}) + \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi_1(f(\boldsymbol{x}),\boldsymbol{y})\Big] - \lambda\cdot R_p(\widehat{\mathbb{P}},\mathbb{P})^p \tag{EC.30}$$

where

$$\begin{aligned}\Psi_{\mathrm{I}}(\boldsymbol{z},\boldsymbol{y}) = \min \quad & \sum_{j=1}^{d_y}\sum_{i=1}^{j} s_{i,j}w_{i,j} + \sum_{i=1}^{d_z} h_i u_i + \sum_{j=1}^{d_y} b_j u'_j \\ \text{s.t.} \quad & \sum_{j=i}^{d_y} w_{i,j} + u_i = z_i, && \forall i \in [d_z], \\ & \sum_{i=1}^{j} w_{i,j} + u'_j = y_j, && \forall j \in \left[d_y\right], \\ & u_i, u'_j, w_{i,j} \geq 0, && \forall i \in [d_z], j \in \left[d_y\right].\end{aligned} \tag{EC.31}$$

The dual problem of the linear programming problem $\Psi_{\mathrm{I}}(f(\boldsymbol{x}),\boldsymbol{y})$ is given by

$$\begin{aligned}\max_{\boldsymbol{\eta}\in\mathbb{R}^{d_z},\boldsymbol{\upsilon}\in\mathbb{R}^{d_y}} \quad & \sum_{i=1}^{d_z} z_i\eta_i + \sum_{j=1}^{d_y} y_j\upsilon_j \\ \text{s.t.} \quad & \eta_i \leq h_i, && \forall i \in [d_z], \\ & \upsilon_j \leq b_j, && \forall j \in \left[d_y\right], \\ & \eta_i + \upsilon_j \leq s_{i,j}, && \forall j \in \left\{i, i+1, \cdots, d_y\right\}, i \in [d_z].\end{aligned} \tag{EC.32}$$

where $\eta_i \in \mathbb{R}$ for each $i \in [d_z]$ represents the dual variable of the i-th constraint in the first constraint set of problem (EC.31), while $\upsilon_j \in \mathbb{R}$ for each $j \in \left[d_y\right]$ represents the dual variable of the j-th constraint in the second constraint set of problem (EC.31). According to the duality theorem, the strong duality holds.

Define that

$$\begin{aligned}\Psi_{\mathrm{I}}^*(f(\boldsymbol{x}),\boldsymbol{y}) := \max_{\boldsymbol{\eta}\in\mathbb{R}^{d_z},\boldsymbol{\upsilon}\in\mathbb{R}^{d_y}} \Bigg\{ & \sum_{i=1}^{d_z}[f(\boldsymbol{x})]_i(\eta_i + c_i) + \sum_{j=1}^{d_y} y_j\upsilon_j \quad \Bigg| \\ & \eta_i \leq h_i, && \forall i \in [d_z], \\ & \upsilon_j \leq b_j, && \forall j \in [d_y], \\ & \eta_i + \upsilon_j \leq s_{i,j}, && \forall j \in \left\{i, i+1, \cdots, d_y\right\}, i \in [d_z] \Bigg\},\end{aligned}$$

and then the problem (EC.30) can be rewritten as

$$\inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathfrak{p}(\widehat{\mathbb{P}})} \quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\left[\Psi_{\mathrm{I}}^*(f(\boldsymbol{x}),\boldsymbol{y})\right] - \lambda \cdot R_p(\widehat{\mathbb{P}},\mathbb{P})^p,$$

which is the same as (Soft-Causal-SDRO). Therefore, following the same reformulation processes as in Theorem 1, its dual formulation is given by

$$\inf_{f\in\mathcal{F}} \quad \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\exp\left(\frac{g^*(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon}\right)\right]\right],$$

where

$$g^*(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda) := \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi_{\mathrm{I}}^*(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right]\right].$$

For a decision rule parameterized by $\boldsymbol{\theta} \in \Theta$, following the same reformulation processes as in Section 5.1, this problem can be solved as the following stochastic compositional optimization problem

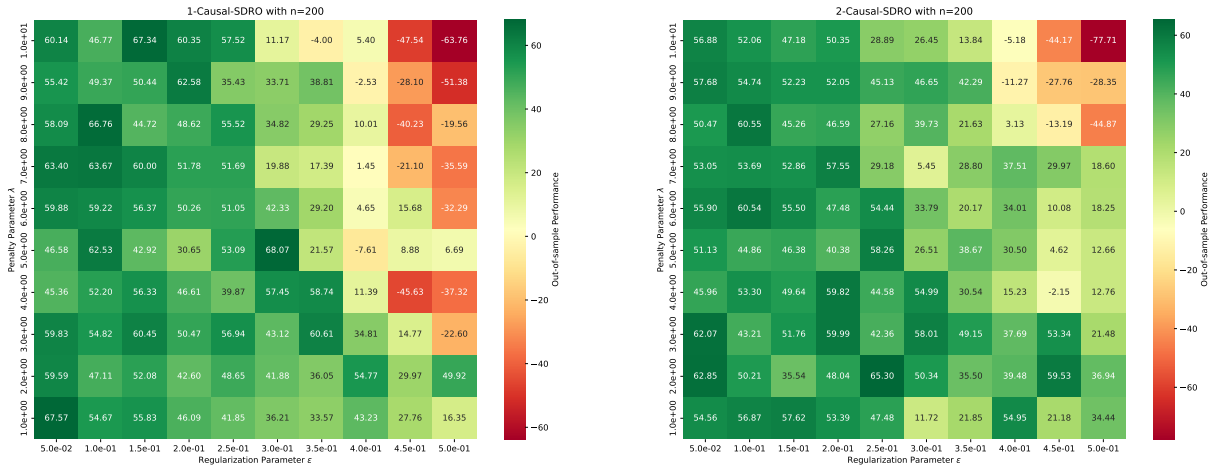
$$\min_{\boldsymbol{\theta}\in\Theta} \quad F(\boldsymbol{\theta}) = \lambda\epsilon \cdot \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[t_1\left(\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[t_3'\left(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2\right)\right];\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1\right)\right];\widehat{\boldsymbol{x}}\right)\right]$$

where

$$\left[t_3'(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\widehat{\boldsymbol{y}},\boldsymbol{\xi}_2)\right]_i = \exp\left(\frac{\Psi_1^*(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_i+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right), \quad \forall i \in [n_{\widehat{\boldsymbol{x}}}].$$

EC.5 Out-of-sample Performance across Different Parameter Combinations

For the feature-based inventory substitution problem, the out-of-sample performance of the proposed method across different parameter combinations is shown in Figure EC.1.



(a) 1-Causal-SDRO

(b) 2-Causal-SDRO

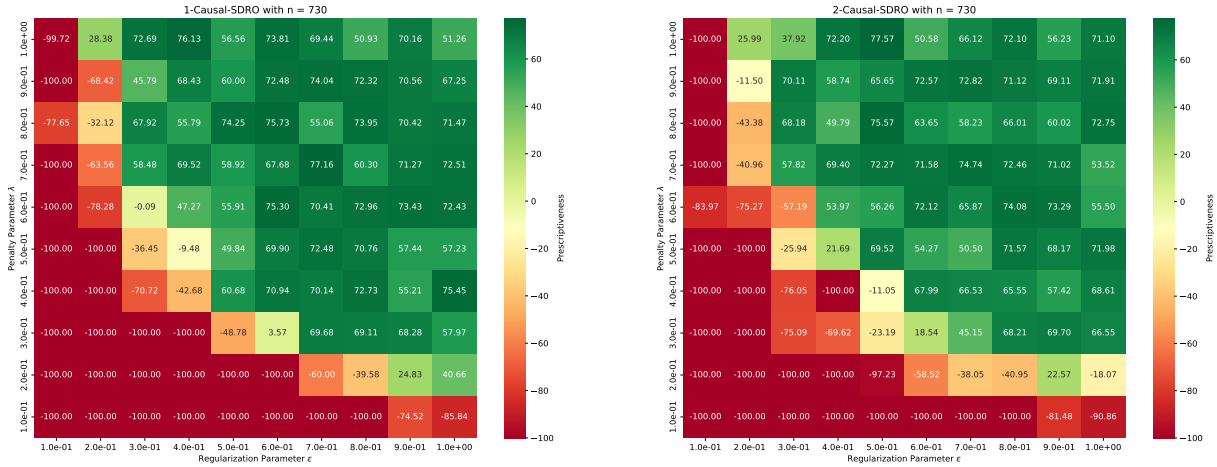
Figure EC.1. Out-of-sample Performance of the inventory substitution problem with different parameters ($N = 200, d_x = 10$)

For the data-driven portfolio selection problem, the out-of-sample performance of the proposed method across different parameter combinations is shown in Figure EC.2.

References for E-Companion

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., & Woodworth, B. (2023). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, *199*(1), 165–214.



(a) 1-Causal-SDRO (b) 2-Causal-SDRO

Figure EC.2. Out-of-sample Performance of the portfolio problem with different parameters ($\eta = 5$)

Blackwell, D., & Ryll-Nardzewski, C. (1963). Non-existence of everywhere proper conditional distributions. *The Annals of Mathematical Statistics*, *34*(1), 223–225.

Casella, G., & Berger, R. (2024). *Statistical inference*. Chapman; Hall/CRC.

Chen, T., Sun, Y., & Yin, W. (2021). Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, *69*, 4937–4948.

Cohn, D. L. (2013). *Measure theory* (Vol. 1). Springer.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1–33.

Hastie, T., Tibshirani, R., Friedman, J., et al. (2009). The elements of statistical learning.

Hu, Y., Chen, X., & He, N. (2020). Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, *30*(3), 2103–2133.

Hu, Z., & Hong, L. J. (2013). Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, *1*(2), 9.

Kallus, N., & Mao, X. (2023). Stochastic optimization forests. *Management Science*, *69*(4), 1975–1994.

Kleywegt, A. J., Shapiro, A., & Homem-de-Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*(2), 479–502.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Notz, P. M., & Pibernik, R. (2024). Explainable subgradient tree boosting for prescriptive analytics in operations management. *European Journal of Operational Research*, *312*(3), 1119–1133.

Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2021). *Lectures on stochastic programming: Modeling and theory*. SIAM.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328.

Wang, J., Gao, R., & Xie, Y. (2025). Sinkhorn distributionally robust optimization. *Operations Research*.

Yang, J., Zhang, L., Chen, N., Gao, R., & Hu, M. (2022). Decision-making with side information: A causal transport robust approach. *Optimization Online*.