CONTEXTUAL DISTRIBUTIONALLY ROBUST OPTIMIZATION WITH CAUSAL AND CONTINUOUS STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a framework for contextual distributionally robust optimization (DRO) that considers the causal and continuous structure of the underlying distribution, and we develop interpretable and tractable decision rules. We first introduce the causal Sinkhorn discrepancy (CSD), an entropy-regularized causal Wasserstein distance that encourages continuous transport plans while preserving 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 obtain an (infinite-dimensional) optimal policy, we propose a soft regression forest (SRF) decision rule: it 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. Finally, we validate our method through numerical experiments on synthetic and real-world datasets, demonstrating its superior performance and interpretability.

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

To hedge against distributional shifts in practice, the contextual distributionally robust optimization (DRO) method has received much attention. This approach seeks optimal robust decision rules that minimize the worst-case risk over an *ambiguity set* containing all plausible joint distributions of features and uncertain parameters. Extensive work constructs ambiguity sets based on Wasserstein distance (Kannan et al., 2024; Nguyen et al., 2020; 2025; Yang et al., 2026). However, these result in discrete worst-case distributions, which may lead to overly conservative or unrealistic decisions since the underlying distribution is typically continuous. Another critical yet often overlooked aspect in contextual DRO is the *causal information structure*: future covariates are conditionally independent of observed uncertain parameters given historical covariates. It thus remains an open question how to develop a contextual DRO model that simultaneously captures the causal relationship and the absolute continuity of the worst-case distribution. Wang et al. (2025) develop a Sinkhorn DRO framework based on the Sinkhorn discrepancy to characterize the continuity of underlying distributions. This framework incorporates entropic regularization into the Wasserstein distance, thereby excluding discrete distributions from the ambiguity set (Blanchet et al., 2023; Zhang et al., 2026).

While the Sinkhorn DRO can provide continuous worst-case distributions, the infinite-dimensional nature of policy optimization poses a computational challenge. To address this, recent literature has developed both parametric and non-parametric decision rules to seek effective approximate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

policies. For parametric rules, Ban & Rudin (2019) consider computationally efficient but locally optimal affine decision rules. Later, Han et al. (2025) and Qi et al. (2023) propose deep learning-based decision rules for general problems, but these rules are difficult to interpret. Alternatively, non-parametric rules developed in Nguyen et al. (2025) and Zhang et al. (2024) offer greater interpretability but are applicable only to specific problems. These lead to the following open question.

How to develop computationally tractable and interpretable decision rules for general contextual DRO problems?

The tree-based family, including decision trees and random forests, has been developed for general CSO models to estimate conditional distributions of uncertain parameters and improve interpretability (Bertsimas & Kallus, 2020; Elmachtoub et al., 2020; Kallus & Mao, 2023). However, few studies employ these 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 (Notz & Pibernik, 2024; Aghaei et al., 2025). It is desirable to explore a computationally tractable, tree-based model for decision-rule optimization that preserves interpretability.

In this paper, we propose a framework for contextual DRO that preserves the causal consistency and continuity of worst-case distributions and includes a tractable and interpretable decision rule. Our main contributions are summarized as follows.

(I) First, we develop a Causal Sinkhorn DRO (Causal-SDRO) model with an ambiguity set based on causal and entropy-regularized Wasserstein distance, combining Yang et al. (2026) and Wang et al. (2025). We further derive the strong dual reformulation and the worst-case distribution for the inner problem of Causal-SDRO under regularity assumptions.

(II) To optimize Causal-SDRO, we propose a parametric, differentiable, and Lipschitz-smooth Soft Regression Forest (SRF) decision rule that approximates optimal policies in a measurable, continuous, and differentiable function space while retaining the intrinsic interpretability of non-differentiable decision trees.

(III) We reformulate Causal-SDRO with parametric decision rules as a multilevel stochastic compositional optimization problem and develop a stochastic compositional gradient algorithm that converges to an ε-stationary point at a rate of $\mathcal{O}(\varepsilon^{-4})$, matching the order of standard stochastic gradient descent (Ghadimi et al., 2016). Numerical results demonstrate the superior performance of our method.

The remainder of this paper is organized as follows. Section 2 formulates the Causal-SDRO model. Section 3 derives the strong dual formulation of Causal-SDRO given a fixed decision rule. Section 4 proposes the SRF decision rule. Section 5 presents an algorithm for solving Causal-SDRO. Section 6 reports representative numerical results. Section 7 concludes the paper. Proofs, analyses, additional discussions, and detailed numerical results are provided in the appendix to this paper.

Notation. For integer $K \in \mathbb{Z}_+$, define $[K] \triangleq \{1, \cdots, K\}$. We denote random vectors by bold uppercase letters (for example, $\boldsymbol{X}, \boldsymbol{Y}$) and their realizations by bold lowercase letters (for example, $\boldsymbol{x}, \boldsymbol{y}$). 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 $\mathcal{Y}$. 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 $\log$ be taken 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 $\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. A function $f : \mathcal{X} \to \mathbb{R}^n$ 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) $t(\cdot; \boldsymbol{\xi})$. 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 decision rule $f : \mathcal{X} \to \mathcal{Z}$ maps a covariate vector $\boldsymbol{x}$ from a covariate space $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ to a decision $\boldsymbol{z}$ from a decision space $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$. Let N be the number of observed data points and K be the number of different values of observed covariates. For each covariate $\widehat{\boldsymbol{x}}_k$, there are n_k observations of the uncertain parameter, denoted by $\widehat{\boldsymbol{y}}_{k,i}$ for any $i \in [n_k]$, such that $N = \sum_{k=1}^{K} n_k$. Let $\widehat{p}_{k,i}$ be the probability mass of the data point

$(\widehat{\boldsymbol{x}}_k, \widehat{\boldsymbol{y}}_{k,i})$. Then, the empirical joint distribution is defined as

$$\widehat{\mathbb{P}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{p}_{k,i} \delta_{(\widehat{\boldsymbol{x}}_k, \widehat{\boldsymbol{y}}_{k,i})}, \quad \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{p}_{k,i} = 1. \tag{1}$$

In the following, we define $\hat{p}_{k,i} := 1/N$. The contextual DRO problem assumes that the unknown true distribution $\mathbb{P}$ lies within an ambiguity set $\Re(\widehat{\mathbb{P}})$ constructed around $\widehat{\mathbb{P}}$, and seeks an optimal decision rule that minimizes the worst-case expected loss over all distributions in the ambiguity set:

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

where $\mathcal{F}$ is a space of measurable functions, and $\Psi : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ is a measurable loss function. To construct our ambiguity set, we first recall the causal transport distance (Yang et al., 2026), which incorporates the causal structure into the optimal transport framework.

Definition 1. (Causal Transport Distance, Yang et al., 2026). *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 causal transport distance-based ambiguity set includes both discrete and continuous distributions, whereas it typically hedges against a discrete one, which may not be realistic in practice. We introduce the *Causal Sinkhorn Discrepancy (CSD)* below by incorporating entropy regularization. When the reference measures $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$ are chosen as continuous measures, $\mathbb{Q} \ll \nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}}$ restricts distributions in the CSD-based ambiguity set to be absolutely continuous.

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

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

where $\Gamma_c(\mathbb{P}, \mathbb{Q})$ is the causal transport plan set, and the relative entropy (also known as Kullback–Leibler divergence) of γ with respect to the measure $\mathbb{P} \otimes (\nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}})$ is given by

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

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

In what follows, we use the **Lebesgue measure** as our reference measure. 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 a CSD-based ambiguity set:

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

where

$$\Re(\widehat{\mathbb{P}}) = \Big\{ \mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p \leq \rho^p \Big\}.$$

This Causal Sinkhorn Distributionally Robust Optimization (Causal-SDRO) model hedges against a continuous worst-case distribution (see Section 3.2) while preserving the causal information structure and thereby avoiding overly conservative and causally inconsistent decisions. We provide detailed discussions on CSD and the causal transport plans in Appendix B.

As an example, we introduce an application of Causal-SDRO. We also evaluate Causal-SDRO on a two-stage inventory substitution problem and a portfolio selection problem; see Appendix F.

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, respectively. Let $(\cdot)^+$ denote the component-wise positive part operator. Consider features $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ (e.g., weather) and uncertain demand $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y} (d_y = d_z)$, and then the problem is given by*

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

This problem will be revisited in Sections 3.2 and 6. ♣

3 Duality Reformulation for Causal-SDRO

In this section, we establish strong duality and characterize the worst-case distribution for the inner maximization problem in (Causal-SDRO), assuming a fixed decision rule $f \in \mathcal{F}$ and a regularization parameter $\epsilon > 0$. We first define the primal problem as

$$v_{\mathrm{P}} := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \Big\{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\Big] : R_\epsilon(\widehat{\mathbb{P}},\mathbb{P})^p \le \rho^p\Big\}. \tag{2}$$

We derive the corresponding dual problem v_{D} as follows (see Appendices B.2 and G.2)

$$v_{\mathrm{D}} = \inf_{\lambda\ge 0}\left\{\lambda\bar{\rho} + \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon\log\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\exp\left(\frac{h(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon}\right)\right]\right]\right\}, \tag{3a}$$

where

$$h(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda) := \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda\epsilon\log\mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right]\right], \tag{3b}$$

$$\mathrm{d}Q_\epsilon(\boldsymbol{\xi}_1) := \frac{e^{-\|\boldsymbol{\xi}_1\|^p/\epsilon}\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{\xi}_1)}{\int_{\mathbb{R}^{d_x}} e^{-\|\boldsymbol{u}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u})}, \quad \mathrm{d}W_\epsilon(\boldsymbol{\xi}_2) := \frac{e^{-\|\boldsymbol{\xi}_2\|^p/\epsilon}\cdot\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{\xi}_2)}{\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})}, \tag{4}$$

and constant

$$\bar{\rho} := \rho^p + \epsilon\cdot\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\log\int_{\mathbb{R}^{d_x}} e^{-\|\boldsymbol{u}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u})\right] + \epsilon\cdot\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{x}}}}\left[\log\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}\|^p/\epsilon}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})\right]. \tag{5}$$

In the following section, Section 3.1 presents the main result on strong duality. Section 3.2 presents the worst-case distribution of the soft-constrained Causal-SDRO.

3.1 Strong Duality

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

Assumption 1. (I) *Both $\mathcal{X}$ and $\mathcal{Z}$ are measurable sets. The loss function $\Psi : \mathcal{Z}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ is measurable and bounded from below by a finite constant, and the decision rule $f : \mathcal{X} \to \mathcal{Z}$ is 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) *For any $\delta > 0$, it holds that $\int_{\mathbb{R}^{d_x}} e^{-\delta\|\boldsymbol{u}\|^p}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u}) < \infty$ and $\int_{\mathbb{R}^{d_y}} e^{-\delta\|\boldsymbol{u}\|^p}\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u}) < \infty$.*

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 several 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 function Ψ 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 Appendix G.1.

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

We call the constraint $R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p \leq \rho^p$ in the primal problem (2) the CSD constraint in the following. Based on Assumption 1 and Condition 1, the following strong duality theorem holds.

Theorem 3. (Strong Duality). *Under Assumption 1, the following results hold.*

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

(II) *Additionally, assume that $\bar{\rho} \geq 0$ 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 3 is provided in Appendix G.2. We further provide discussions, including a comparison with causal Wasserstein DRO and the worst-case distribution, in Appendix B.3.

Remark 4. (Soft-constrained Causal-SDRO). *In (Causal-SDRO), the radius ρ is a hyperparameter to be tuned while λ is the dual variable. In the following soft-constrained Causal-SDRO problem, it suffices to tune a hyperparameter $\lambda > 0$:*

$$\inf_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathbb{P}} \Big[\Psi(f(\boldsymbol{x}), \boldsymbol{y}) \Big] - \lambda \cdot R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p. \tag{Soft-Causal-SDRO}$$

By the Fenchel duality, the dual formulation of (Soft-Causal-SDRO) is given by

$$\inf_{f \in \mathcal{F}} \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \left[\lambda \epsilon \log \mathbb{E}_{\boldsymbol{\xi}_1 \sim Q_\epsilon} \left[\exp \left(\frac{h(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1, \lambda)}{\lambda \epsilon} \right) \right] \right], \tag{6}$$

where the definition of function h is the same as Eq. (3b). Under Assumption 1, the strong duality of the inner problem in (Soft-Causal-SDRO) holds (by a proof similar to that of Theorem 3). ♣

3.2 Worst-Case Distribution

Based on Theorem 3, Theorem 5 derive the worst-case distribution expression for the inner problem of (Soft-Causal-SDRO), see Appendices B.4 and G.3 for details.

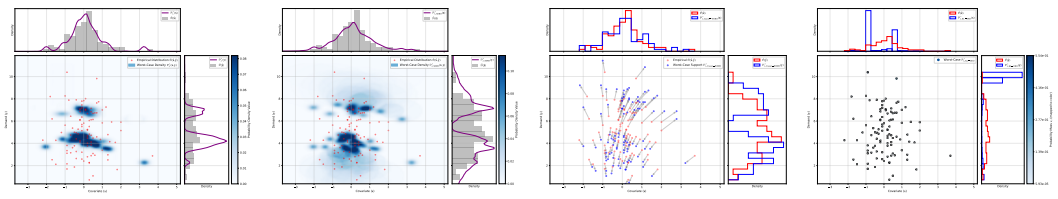
Theorem 5. (Worst-case Distribution of the Soft-constrained Causal-SDRO). *Under Assumption 1, the density of worst-case distribution $\mathbb{P}^*_\lambda$ of the inner problem of (Soft-Causal-SDRO) for any $\lambda > 0$ is given by*

$$\frac{\mathrm{d}\mathbb{P}^*_\lambda(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \mathbb{E}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \sim \widehat{\mathbb{P}}} \Big[\alpha_{\widehat{\boldsymbol{x}}}(\lambda) \cdot \beta_{\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}}(\lambda) \cdot e^{r'(\lambda, \widehat{\boldsymbol{x}}, \boldsymbol{x}) + s'(\lambda, \widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}, \boldsymbol{y})} \Big], \tag{7}$$

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

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

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



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

Figure 1: Structure of distributions $\widehat{\mathbb{P}}$ (red points in 1a-1c), $\mathbb{P}^*_\lambda$, $\mathbb{P}^*_{\lambda,\text{SDRO}}$, $\mathbb{P}^*_{\lambda,\text{Causal-WDRO}}$, and $\mathbb{P}^*_{\lambda,\text{KL-DRO}}$.($p = 2$, $\lambda = 0.5$, $\epsilon = 0.05$, and $N = 100$)

Theorem 5 reveals that the worst-case distribution $\mathbb{P}^*_\lambda$ is a mixture of Gibbs distributions. 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 Appendix C.

Example 1. (Revisited). *Consider a single-product feature-based newsvendor problem with one covariate (see Example 1, $d_x = d_y = d_z = 1$). We choose the Lebesgue measure as the reference measure. Consider a true decision rule $f = f_{true}$, in Figure 1, 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. Comparing Figure 1(a-b) with Figure 1(c-d), we see that entropic regularization yields continuous worst-case distributions. Comparing Figure 1(a) with Figure 1(b), the probability density of $\mathbb{P}^*_\lambda$ is more concentrated than that of $\mathbb{P}^*_{\lambda,\text{SDRO}}$, as Causal-SDRO 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 space's infinite dimensionality. Instead, we consider a parametric decision rule $f : \mathcal{X} \to \mathcal{Z}$ that approximates the optimal mapping from features to decisions. In this section, we propose a parametric and interpretable Soft Regression Forest (SRF) decision rule.

Unlike uninterpretable deep-learning-based methods, we propose a soft regression forest (SRF) decision rule based on the principles of soft decision trees (Frosst & Hinton, 2017) and ensemble learning to capture the 'if-then' decision-making structure. Compared with the traditional decision-tree-based methods, SRF is parametric and differentiable while retaining intrinsic interpretability.

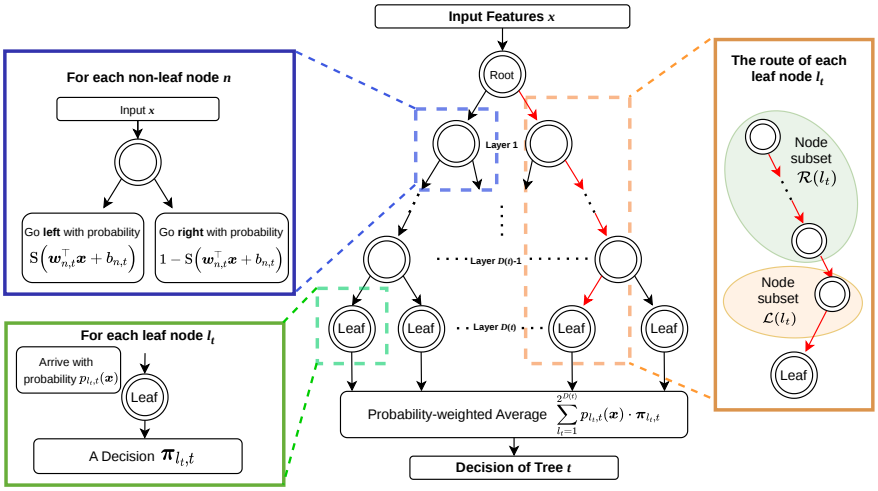


Figure 2: Structure of a soft regression tree t with depth $D(t)$.

The SRF consists of T full binary Soft Regression Trees (SRTs). For the t-th tree ($t \in [T]$) with depth $D(t) \geq 2$, each leaf node $l_t \in \left[2^{D(t)}\right]$ corresponds to a decision $\boldsymbol{\pi}_{l_t,t} \in \mathcal{Z}$ and a unique route from the root node in the tree. Unlike a hard decision tree, which selects only one child node at each branch, a Sigmoid function $\mathrm{S}(\boldsymbol{w}_{j,t}^\top \boldsymbol{x} + b_{j,t})$ determines the probability of directing the input $\boldsymbol{x}$ to the left child at each internal node $j \in [2^{D(t)} - 1]$ for $\boldsymbol{w}_{j,t} \in \mathbb{R}^{d_x}$ and $b_{j,t} \in \mathbb{R}$. Let $\mathcal{L}(l_t)$ and $\mathcal{R}(l_t)$ denote the left- and right-hand-side node sets on the route of leaf node l_t in tree t, respectively, and $\Lambda(l_t) := \mathcal{L}(l_t) \cup \mathcal{R}(l_t)$. Consequently, the probability that a given input covariate $\boldsymbol{x}$ reaches leaf node l (corresponds to the decision $\boldsymbol{\pi}_{l_t,t}$) in tree t is given by

$$p_{l_t,t}(\boldsymbol{x}) = \prod_{i \in \mathcal{L}(l_t)} \mathrm{S}(\boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}) \cdot \prod_{j \in \mathcal{R}(l_t)} \Big(1 - \mathrm{S}(\boldsymbol{w}_{j,t}^\top \boldsymbol{x} + b_{j,t})\Big).$$

Assume that $\mathcal{Z}$ is a convex set. The final output of the SRF is the ensemble average of the expected decisions from all trees, and thus the SRF decision rule $f_{\boldsymbol{\theta}}^{\text{SRF}} : \mathcal{X} \to \mathcal{Z}$ is explicitly defined as

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

where $[\boldsymbol{\pi}_{l_t,t}]_k$ represents the k-th decision for any $k \in [d_z]$, the vector $\boldsymbol{\theta} \in \Theta$ is the collection of all individual parameters $\{\boldsymbol{w}_{i,t}, \boldsymbol{b}_{i,t}, \boldsymbol{\pi}_{l_t,t}\}$ for each non-leaf node $i \in \Lambda(l_t)$, leaf node $l_t \in \left[2^{D(t)}\right]$ and tree $t \in [T]$. This decision rule has $(d_x + 1) \cdot \sum_{t=1}^{T}(2^{D(t)} - 1) + d_z \cdot \sum_{t=1}^{T} 2^{D(t)}$ parameters.

Recall that Bertsimas & Kallus (2020) and Kallus & Mao (2023) introduce tree-based models for solving contextual optimization. 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.

Figure 2 shows the structure of a single SRT 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 the node subsets $\mathcal{L}(l_t)$ and $\mathcal{R}(l_t)$ for each leaf node. We further demonstrate the intrinsic interpretability of SRF, including its Lipschitz smoothness, consistency with traditional decision trees, and both global and local interpretation, in Appendix D. We also examine the interpretability of SRF in one of our applications in Appendix F.3.3.

5 Solving Causal-SDRO

In this section, we propose an algorithm to solve the Causal-SDRO problem. Section 5.1 reformulates (6) into a three-level stochastic compositional optimization problem. For this tractable formulation, Section 5.2 develops a gradient-based algorithm.

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:

$$\min_{\boldsymbol{\theta}\in\Theta} \quad F(\boldsymbol{\theta}) = \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\exp\left(\frac{h(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\lambda)}{\lambda\epsilon}\right)\right]\right] \tag{8a}$$

where

$$h(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) = \mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}}\left[\lambda\epsilon \log \mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}\left[\exp\left(\frac{\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right)\right]\right]. \tag{8b}$$

We next explicitly define the empirical conditional probability weights $p(\widehat{\boldsymbol{y}}_j \mid \widehat{\boldsymbol{x}})$ associated with $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}$ in Eq. (8b). For any covariate $\widehat{\boldsymbol{x}}$ in historical observations $\{(\widehat{\boldsymbol{x}}_j,\widehat{\boldsymbol{y}}_j)\}_{j=1}^N$, we denote it as corresponding to $n(\widehat{\boldsymbol{x}})$ uncertain parameters, and define $\mathcal{I}(\widehat{\boldsymbol{x}}) := \{j \in [N] : \widehat{\boldsymbol{x}}_j = \widehat{\boldsymbol{x}}\}$ such that $n(\widehat{\boldsymbol{x}}) = |\mathcal{I}(\widehat{\boldsymbol{x}})|$. Let $p(\widehat{\boldsymbol{y}}_j \mid \widehat{\boldsymbol{x}}) := \mathbb{1}_{\{j\in\mathcal{I}(\widehat{\boldsymbol{x}})\}}/n(\widehat{\boldsymbol{x}})$ for given $\widehat{\boldsymbol{x}}$ and any $j \in [N]$, which satisfies $\sum_{j=1}^N p(\widehat{\boldsymbol{y}}_j \mid \widehat{\boldsymbol{x}}) = 1$, where $\mathbb{1}_{\{j\in\mathcal{I}(\widehat{\boldsymbol{x}})\}}$ equals 1 when $j \in \mathcal{I}(\widehat{\boldsymbol{x}})$ and 0 otherwise.

Therefore, by rewriting the outer expectation in Eq. (8b) as a finite-sample summation, we reformulate the problem (8) as a three-level Stochastic Compositional optimization (SCO) problem, driven by the three independent random vectors $\widehat{\boldsymbol{x}}$, $\boldsymbol{\xi}_1$, and $\boldsymbol{\xi}_2$:

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

where

$$\begin{aligned}
&t_1 : \mathbb{R}_+ \to \mathbb{R}, && t_1(z;\widehat{\boldsymbol{x}}) = \log(z),\\
&t_2 : \mathbb{R}^N \to \mathbb{R}_+, && t_2(\boldsymbol{v};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) = \exp\Big(\sum_{j=1}^N p(\widehat{\boldsymbol{y}}_j \mid \widehat{\boldsymbol{x}})\cdot\log(\boldsymbol{v}_j)\Big),\\
&t_3 : \Theta \to \mathbb{R}^N, && \Big[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\Big]_j = \exp\left(\frac{\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_1),\widehat{\boldsymbol{y}}_j+\boldsymbol{\xi}_2)}{\lambda\epsilon}\right), \forall j \in [N],
\end{aligned} \tag{9}$$

where v_j represents the j-th element of vector $\boldsymbol{v} \in \mathbb{R}^N$. We introduce several assumptions for this problem in Assumption 2, which are commonly used by related literature, e.g., Hu et al. (2020).

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 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \le D_\Theta$.*

(II) *(Lipschitz Continuity and Smoothness). For any $\boldsymbol{x}$ and $\boldsymbol{y}$ and given parametric decision rule, the loss function $L(\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}) := \Psi(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{y})$ is: (1) differentiable with respect to $\boldsymbol{\theta}$ on an open neighborhood of Θ; (2) $L_{\boldsymbol{\theta}}$-Lipschitz continuous and $S_{\boldsymbol{\theta}}$-Lipschitz smooth 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) regarding the diameter of the decision space is used in the sample complexity analysis. Assumption 2(II) is crucial for deriving the convergence rate of the gradient-based algorithms.

For this problem, we prove its properties under Assumption 2 and provide the sample complexity of the SAA method in Appendix E, which motivates the development of an efficient algorithm.

5.2 Stochastic Compositional Algorithm

In this subsection, we introduce a gradient-based algorithm for the problem (SCO). Since the bias of the stochastic gradient descent (SGD) method for (SCO) is uncontrollable, it cannot be used to solve it directly (see Appendix E.3). Therefore, we provide a Stochastically Corrected Stochastic Compositional gradient method (SCSC, Chen et al., 2021) to solve problem (SCO), which controls bias using momentum-based gradient updates.

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

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

The SCSC 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 $\boldsymbol{\xi}_2^k$, the expectations of t_3 with respect to $\boldsymbol{\xi}_2$ and of t_2 with respect to $\boldsymbol{\xi}_1$ are estimated by $y_1^k \in \mathbb{R}$ and $\boldsymbol{y}_2^k \in \mathbb{R}^N$, respectively, and thereby the parameters are updated by

$$\boldsymbol{\theta}^{k+1} := \boldsymbol{\theta}^k - \alpha_k \cdot \nabla t_1^k(y_1^{k+1}) \cdot \nabla t_2^k(\boldsymbol{y}_2^{k+1}) \cdot \nabla t_3^k(\boldsymbol{\theta}^k), \tag{10}$$

where

$$\boldsymbol{y}_2^{k+1} = (1-\beta_k) \cdot \Big(\boldsymbol{y}_2^k + t_3^k(\boldsymbol{\theta}^k) - t_3^k(\boldsymbol{\theta}^{k-1})\Big) + \beta_k \cdot t_3^k(\boldsymbol{\theta}^k), \tag{11}$$

and

$$y_1^{k+1} = (1-\beta_k) \cdot \Big(y_1^k + t_2^k(\boldsymbol{y}_2^{k+1}) - t_2^k(\boldsymbol{y}_2^k)\Big) + \beta_k \cdot t_2^k(\boldsymbol{y}_2^{k+1}). \tag{12}$$

The pseudo-code of the SCSC is provided in Algorithm 1. Steps 5 and 7 are based on Eqs. (11) and (12), respectively.

Algorithm 1 SCSC for the problem (SCO)

1: Initialize $\boldsymbol{\theta}^{-1} = \boldsymbol{\theta}^0 = \boldsymbol{\theta}^{\text{prev}}$, $y_1^0 = 1$, $\boldsymbol{y}_2^0 = \boldsymbol{y}_2^{\text{prev}} = \mathbf{1}$, and stepsizes α_0, β_0;
2: **for** $k = 0, \cdots, K-1$, **do**
3: Sample $(\widehat{\boldsymbol{x}}^k, \widehat{\boldsymbol{y}}^k)$, $\boldsymbol{\xi}_1^k$, and $\boldsymbol{\xi}_2^k$;
4: Compute $t_3^k(\boldsymbol{\theta}^k)$ and $t_3^k(\boldsymbol{\theta}^{\text{prev}})$;
5: Update $\boldsymbol{y}_2^{k+1} \leftarrow (1-\beta_k) \cdot \Big(\boldsymbol{y}_2^k + t_3^k(\boldsymbol{\theta}^k) - t_3^k(\boldsymbol{\theta}^{\text{prev}})\Big) + \beta_k \cdot t_3^k(\boldsymbol{\theta}^k)$;
6: Compute $t_2^k(\boldsymbol{y}_2^{k+1})$ and $t_2^k(\boldsymbol{y}_2^{\text{prev}})$;
7: Update $y_1^{k+1} \leftarrow (1-\beta_k) \cdot \Big(y_1^k + t_2^k(\boldsymbol{y}_2^{k+1}) - t_2^k(\boldsymbol{y}_2^{\text{prev}})\Big) + \beta_k \cdot t_2^k(\boldsymbol{y}_2^{k+1})$;
8: Update $\boldsymbol{\theta}^{k+1}$ by Eq. (10), $\boldsymbol{\theta}^{\text{prev}} \leftarrow \boldsymbol{\theta}^k$, $\boldsymbol{y}_2^{\text{prev}} \leftarrow \boldsymbol{y}_2^{k+1}$;
9: **end for**

For the convergence analysis, we assume that at each iteration k, the samples $\boldsymbol{\xi}_2^k \sim W_\epsilon$, $\boldsymbol{\xi}_1^k \sim Q_\epsilon$, and $\widehat{\boldsymbol{x}}^k \sim \widehat{\mathbb{P}}_{\boldsymbol{X}}$ are mutually independent and independent of all samples generated before iteration k. The convergence of SCSC for SCO is given as Theorem 6.

Theorem 6. (Convergence of SCSC for the Problem SCO). *Under Assumption 2, 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:*

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

where $A_1, A_2, C_{\text{const}}$ *are constants that depend on the setting of the algorithm and problem;*

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

The proof of Theorem 6 is provided in Appendix G.4. Theorem 6 also shows that the convergence rate of SCSC is $\mathcal{O}(K^{-1/2})$, which is of the same order as SGD's rate for the stochastic non-compositional nonconvex problems.

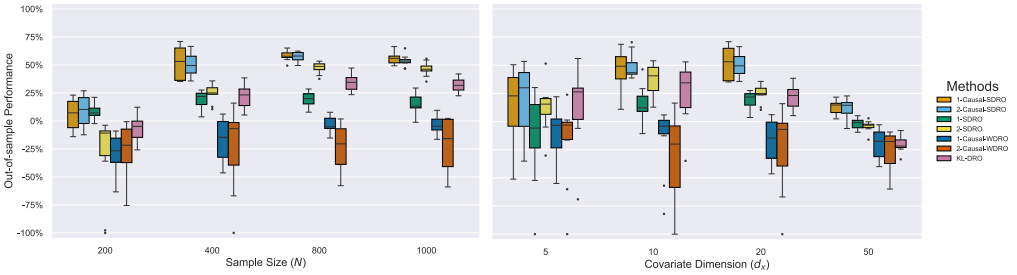
6 Numerical Results on the Feature-based Newsvendor Problem

In this section, we validate the efficiency of the proposed approach for the feature-based newsvendor problem (see Example 1). We also apply our approach to the inventory substitution and portfolio selection problems using real data in Appendices F.2 and F.3, respectively. Source code and results associated with this work are uploaded as part of the supplementary material.

In all experiments, we train the decision rule by solving the soft-constrained Causal-SDRO model. Following Remark 4, we treat the penalty coefficient λ as a tunable hyperparameter instead of the radius $\bar{\rho}$. We use the ℓ_p-norm as the transport cost in the causal Sinkhorn discrepancy, with $p \in \{1, 2\}$. For brevity, we denote the resulting model as *p-Causal-SDRO* and solve it using the proposed SCSC algorithm. We evaluate the out-of-sample performance using the coefficient of prescriptiveness (Bertsimas & Kallus, 2020), which is defined as

$$\text{Out-of-sample Performance}(\boldsymbol{\theta}) := \left(1 - \frac{\mathcal{H}(\boldsymbol{\theta}) - \mathcal{H}^*}{\mathcal{H}(\boldsymbol{\theta}^{\text{ERM}}) - \mathcal{H}^*}\right) \times 100\%, \tag{13}$$

where $\mathcal{H}(\boldsymbol{\theta})$ is the average out-of-sample loss of the trained decision rule $f_{\boldsymbol{\theta}}$, $\mathcal{H}^*$ represents the oracle loss under perfect information, and $\boldsymbol{\theta}^{\text{ERM}}$ denotes the parameter trained by the empirical risk minimization (ERM) model. A higher value indicates a better out-of-sample performance of decision rules. We place other necessary detailed settings in Appendix F and parameters of the feature-based newsvendor problem in Appendix F.1.



(a) $d_x = 20$, $N \in \{200, 400, 800, 1000\}$ (b) $N = 400$, $d_x \in \{5, 10, 20, 50\}$

Figure 3: Comparison of different DRO models on the newsvendor problem

Figure 3 compares the out-of-sample performance of the proposed Causal-SDRO models and existing DRO models equipped with the SRF decision rule, including SDRO, Causal-WDRO, and KL-DRO, across different sample sizes and feature dimensions, as shown in Appendix C. Figure 3(a) reports the scenario where the feature dimension $d_x = 20$ is fixed and sample size $N \in \{200, 400, 800, 1000\}$, and Figure 3(b) reports the scenario where $N = 400$ and $d_x \in \{5, 10, 20, 50\}$. The parameters of these DRO models are determined by cross-validation. As illustrated, the Causal-SDRO models (orange and blue bars) achieve a better average out-of-sample performance than other DRO models and a clear improvement over the ERM model.

In Appendix F.1, we further present additional experimental results, including comparisons with a decision-rule benchmark and an analysis of parameter sensitivity.

7 Conclusion

In this paper, we consider the causal and continuous structure of the underlying distribution for contextual DRO. We develop a new contextual DRO framework, Causal-SDRO, that constructs the ambiguity set using the entropy-regularized causal Wasserstein distance, thereby excluding discrete and causally implausible distributions. We then develop a soft regression forest decision rule that approaches optimal policies over measurable, continuous, and differentiable function spaces while preserving interpretability and computational tractability. To solve the resulting model, we present a gradient-based algorithm that achieves a near-optimal convergence rate of $\mathcal{O}(\varepsilon^{-4})$. Numerical results demonstrate the superior performance of our method.

Future work may focus on designing more informative ambiguity sets and extending the framework to practical applications requiring robustness and interpretability.

AI USE STATEMENT

In this work, we used generative AI tools to polish and streamline the writing. We also used generative AI to assist us to check the correctness of our theoretical results and in writing code for generating plots. In particular, we used generative AI to assist with the proof of Theorem 6. All AI-assisted content, including the proof, was carefully reviewed and verified by the authors. We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Source code and results associated with this work are uploaded as part of the supplementary material. The data-generation program or the real dataset was also included in the supplementary materials we submitted. For our theoretical results, we provide the assumptions in the manuscript and the proof and analyses in the appendix.

REFERENCES

Sina Aghaei, Andrés Gómez, and Phebe Vayanos. Strong optimal classification trees. *Operations Research*, 73(4):2223–2241, 2025.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

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

Dimitris Bertsimas and Bartolomeo Stellato. The voice of optimization. *Machine Learning*, 110(2): 249–277, 2021.

Dimitris Bertsimas, Arthur Delarue, Patrick Jaillet, and Sebastien Martin. The price of interpretability. *arXiv preprint arXiv:1907.03419*, 2019.

David Blackwell and Czesław Ryll-Nardzewski. Non-existence of everywhere proper conditional distributions. *The Annals of Mathematical Statistics*, 34(1):223–225, 1963.

Jose Blanchet, Daniel Kuhn, Jiajin Li, and Bahar Taskesen. Unifying distributionally robust optimization via optimal transport theory. *arXiv preprint arXiv:2308.05414*, 2023.

George Casella and Roger Berger. *Statistical inference*. Chapman and Hall/CRC, 2024.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

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

Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. Data-driven conditional robust optimization. *Advances in Neural Information Processing Systems*, 35:9525–9537, 2022.

Donald L Cohn. *Measure theory*, volume 1. Springer, 2013.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

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

Adam N Elmachtoub, Jason Cheuk Nam Liang, and Ryan McNellis. Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning*, pp. 2858–2867. PMLR, 2020.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7):620–631, 2021.

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

Alexandre Forel, Axel Parmentier, and Thibaut Vidal. Explainable data-driven optimization: from context to decision and back again. In *International Conference on Machine Learning*, pp. 10170–10187. PMLR, 2023.

Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

Mingyang Fu, Xiaobo Li, and Lianmin Zhang. Distributionally robust newsvendor under stochastic dominance with a feature-based application. *Manufacturing & Service Operations Management*, 26(5):1962–1977, 2024.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

Jinhui Han, Ming Hu, and Guohao Shen. Deep neural newsvendor. *Management Science*, 2025.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

Yifan Hu, Xin Chen, and Niao He. Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, 30(3):2103–2133, 2020.

Yifan Hu, Jie Wang, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36:78412–78434, 2023.

Zhaolin Hu and L Jeff Hong. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

Nathan Kallus and Xiaojie Mao. Stochastic optimization forests. *Management Science*, 69(4): 1975–1994, 2023.

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Residuals-based distributionally robust optimization with covariate information. *Mathematical Programming*, 207(1):369–425, 2024.

Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Liwan H Liyanage and J George Shanthikumar. A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4):341–348, 2005.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

Chao Ma, Lei Wu, and Weinan E. A priori estimates of the population risk for two-layer neural networks. *arXiv preprint arXiv:1810.06397*, 2018.

Viet Anh Nguyen, Fan Zhang, Jose Blanchet, Erick Delage, and Yinyu Ye. Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems*, 33:15232–15242, 2020.

Viet Anh Nguyen, Fan Zhang, Shanshan Wang, Jose Blanchet, Erick Delage, and Yinyu Ye. Robustifying conditional portfolio decisions via optimal transport. *Operations Research*, 73(5): 2801–2829, 2025.

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

Afshin Oroojlooyjadid, Lawrence V Snyder, and Martin Takáč. Applying deep learning to the newsvendor problem. *IISE Transactions*, 52(4):444–463, 2020.

Georgia Perakis, Melvyn Sim, Qinshen Tang, and Peng Xiong. Robust pricing and production with information partitioning and adaptation. *Management Science*, 69(3):1398–1419, 2023.

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

Mehran Poursoltani, Erick Delage, and Angelos Georghiou. Robust data-driven prescriptiveness optimization. *arXiv preprint arXiv:2306.05937*, 2023.

Meng Qi, Ying Cao, and Zuo-Jun Shen. Distributionally robust conditional quantile prediction with fixed design. *Management Science*, 68(3):1639–1658, 2022.

Meng Qi, Yuanyuan Shi, Yongzhi Qi, Chenxin Ma, Rong Yuan, Di Wu, and Zuo-Jun Shen. A practical end-to-end inventory management model with deep learning. *Management Science*, 69 (2):759–773, 2023.

Meng Qi, Paul Grigas, and Zuo-Jun Shen. Integrated conditional estimation-optimization. *Operations Research*, 2025.

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

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Melvyn Sim, Qinshen Tang, Minglong Zhou, and Taozeng Zhu. The analytics of robust satisficing: predict, optimize, satisfice, then fortify. *Operations Research*, 73(5):2708–2728, 2025.

Prateek R Srivastava, Yijie Wang, Grani A Hanasusanto, and Chin Pang Ho. On data-driven prescriptive analytics with side information: A regularized Nadaraya-Watson approach. *arXiv preprint arXiv:2110.04855*, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *Operations Research*, 2025.

Jincheng Yang, Luhao Zhang, Ningyuan Chen, Rui Gao, and Ming Hu. Decision making with side information: A causal transport robust approach. *Operations Research*, 2026.

Fenglin Zhang, Teyan Liu, and Jie Wang. Generative neural networks for Sinkhorn distributionally robust hypothesis testing. *arXiv preprint arXiv:2608.22746*, 2026.

Luhao Zhang, Jincheng Yang, and Rui Gao. Optimal robust policy for feature-based newsvendor. *Management Science*, 70(4):2315–2329, 2024.

APPENDIX

A RELATED LITERATURE

In this section, 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 uncertainty (Bertsimas & Kallus, 2020; Sim et al., 2025). This approach, which leverages rich covariates to improve decision-making under parameter uncertainty, is called Contextual Stochastic Optimization (Sadana et al., 2025). To address this problem, existing studies have developed various frameworks. Bertsimas & Kallus (2020) propose a data-driven framework based on weighted sample average approximation. This method estimates the conditional distribution by generating data-driven weights using machine learning models (for example, the k-nearest neighbors method and decision trees) and makes decisions by minimizing the reweighted empirical cost. Building on this, Kallus & Mao (2023) propose a stochastic optimization forest method that calculates weights by optimizing downstream decision quality rather than prediction accuracy. Elmachtoub & Grigas (2022) develop a smart predict-then-optimize framework that integrates learning and optimization through a decision-oriented loss function. Qi et al. (2025) present an integrated conditional estimation-optimization framework based on the downstream objective. Unlike the aforementioned approaches, which still involve an intermediate estimation step, our work focuses on the decision-rule approach, which directly maps covariates to final decisions (Liyanage & Shanthikumar, 2005). Notable examples include the linear and kernel-based decision rules for newsvendor problems by Ban & 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 (Srivastava et al., 2021; Poursoltani et al., 2023) and moment-based (Perakis et al., 2023) ambiguity sets, primarily due to their computational tractability. In recent literature, the optimal-transport-based ambiguity sets are widely used for contextual DRO modeling (Qi et al., 2022; Esteban-Pérez & Morales, 2022; Zhang et al., 2024; Kannan et al., 2024; Nguyen et al., 2025), due to their data-driven nature and satisfactory out-of-sample guarantees. Yang et al. (2026) and Wang et al. (2025) proposed variants of the Wasserstein distance to model worst-case distributions with causal or continuous structures. Our work accounts for the structure of the contextual DRO model and develops an efficient, interpretable decision-rule approach.

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. Yang et al. (2026) 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. In such a case, Hu et al. (2023) reformulate the problem as a large-scale bilevel program, which remains computationally challenging. For non-parametric decision rules, finite-dimensional convex reformulations of contextual DRO have been proposed (Yang et al., 2026; Esteban-Pérez & Morales, 2022; Fu et al., 2024; Zhang et al., 2024; Nguyen et al., 2025) for specific problem structures or for particular choices of the Wasserstein distance and its variants. In this paper, we propose a Soft Regression Forest (SRF) decision rule to approximate an optimal policy in measurable, continuous, and differentiable function spaces, while balancing 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 (Doshi-Velez & Kim, 2017; Bertsimas et al., 2019). As mentioned above, kernel- or deep-learning-based methods for contextual optimization perform well but lack trustworthiness and interpretability (Oroojlooyjadid et al., 2020; Bertsimas & Koduri, 2022). Therefore, in high-stakes applications, such as healthcare and finance, those methods may pose risks (Rudin, 2019; Forel et al., 2023). Instead of explaining deep learning models (Lundberg et al., 2020), extensive literature is dedicated 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 (Bertsimas & Dunn, 2017; Bertsimas & Stellato,

2021; Aghaei et al., 2025). These methods have also been applied in CSO (Elmachtoub et al., 2020; Bertsimas & Kallus, 2020; Kallus & Mao, 2023; Notz & Pibernik, 2024). Unlike existing tree-based methods, which rely on greedy heuristics due to non-differentiability and NP-hardness, the proposed SRF decision rule is differentiable and can be efficiently trained with gradient-based methods while maintaining intrinsic interpretability.

B Discussion on the Causal Sinkhorn DRO

In this section, we provide more detailed discussions on the causal Sinkhorn discrepancy (CSD), causal transport plans, the CSD-based ambiguity set, and the worst-case distributions of the Causal-SDRO.

B.1 Discussion on CSD and Causal Transport Plans

In this subsection, we discuss the proposed CSD and the resulting causal transport plans.

Remark 7. *Note that the entropy regularization term $\mathbb{D}_{\mathrm{KL}}\left(\gamma\|\mathbb{P}\otimes(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}})\right)$ in the causal Sinkhorn discrepancy in Definition 2 is defined with respect to general reference measures $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$. By contrast, the relative entropy in entropically regularized optimal transport (Peyré & Cuturi, 2019) is typically defined as*

$$\mathbb{D}'_{\mathrm{KL}}\left(\gamma\|\mathbb{P}\otimes\mathbb{Q}\right):=\mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\left[\log\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\mathbb{P}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\mathbb{Q}(\boldsymbol{x},\boldsymbol{y})}\right].$$

If $\mathbb{Q}\ll\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}$, then $\mathrm{d}\mathbb{Q}(\boldsymbol{x},\boldsymbol{y})=\frac{\mathrm{d}\mathbb{Q}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})$, and therefore

$$\mathbb{D}_{\mathrm{KL}}\left(\gamma\|\mathbb{P}\otimes(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}})\right)=\mathbb{D}'_{\mathrm{KL}}\left(\gamma\|\mathbb{P}\otimes\mathbb{Q}\right)+\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{Q}}\left[\log\frac{\mathrm{d}\mathbb{Q}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}\right].$$

In particular, if $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$ are the Lebesgue measures and $\mathbb{Q}$ admits a joint density $q(\boldsymbol{x},\boldsymbol{y})$, then $\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})=\mathrm{d}\boldsymbol{x}$, $\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})=\mathrm{d}\boldsymbol{y}$, and

$$\mathbb{D}_{\mathrm{KL}}\left(\gamma\|\mathbb{P}\otimes(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}})\right)=\mathbb{D}'_{\mathrm{KL}}\left(\gamma\|\mathbb{P}\otimes\mathbb{Q}\right)+\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{Q}}\left[\log q(\boldsymbol{x},\boldsymbol{y})\right].$$ ♣

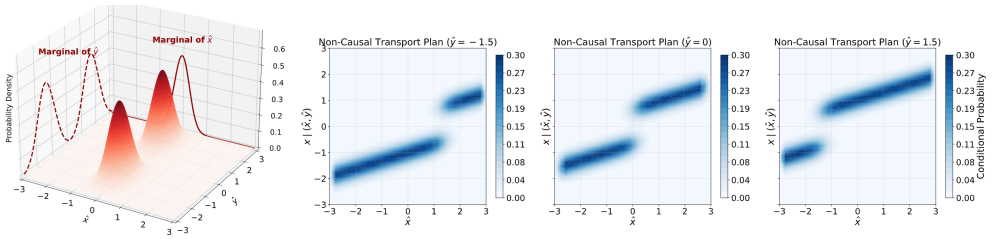
We use the Lebesgue measure as our reference measure. Unlike deterministic transport plans derived from the Wasserstein distance, Sinkhorn transport plans are probabilistic. Specifically, entropy regularization ($\epsilon>0$) penalizes deterministic transport plans, yielding smooth plans that map each source point to a probability distribution over the target space rather than a single point.

Figure 4 visualizes and compares the causal and non-causal transport plans for the illustrative source $(\widehat{x},\widehat{y})$ and target (x,y) distributions supported on $[-3,3]^2$. The source shown in Figure 4(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 4(b) is a mixture of two bivariate Gaussians centered at $(-1,-1)$ and $(1,1)$ with positive correlation. Figure 4(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 4(d), middle panel) enforces conditional independence and, consequently, appears as an aggregate of the non-causal plans. Figure 4(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 the marginal distributions of $\widehat{x}$ and x.

B.2 Reformulation for the Dual Causal-SDRO

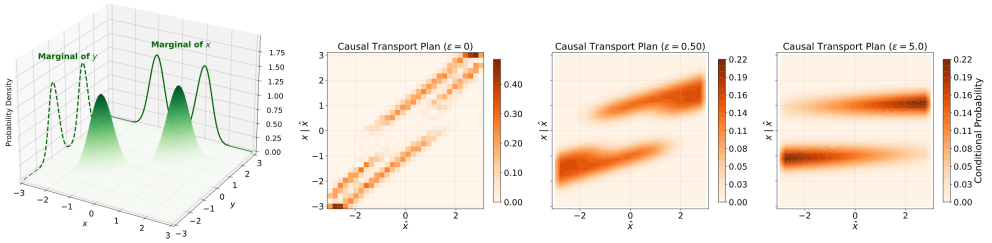
We derive the dual problem v_{D} of v_{P} in Eq. (2) as (see Appendix G.2):

$$v_{\mathrm{D}}:=\inf_{\lambda\geq0}\left\{\lambda\rho^p+\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\lambda\epsilon\log\int_{\mathcal{X}}\exp\left(\frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda)}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right]\right\},\tag{14a}$$



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

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



(b) Target Distribution (x, y)

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

Figure 4: Visualization for causal and non-causal Sinkhorn transport plans.

where

$$g(\widehat{\boldsymbol{x}}, \boldsymbol{x}, \lambda) = \mathbb{E}_{\widehat{\boldsymbol{y}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}} \mid \widehat{\boldsymbol{X}} = \widehat{\boldsymbol{x}}}} \left[\lambda \epsilon \log \int_{\mathcal{Y}} \exp \left(\frac{\Psi(f(\boldsymbol{x}), \boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))}{\lambda \epsilon} \right) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y}) \right]. \quad \text{(14b)}$$

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

$$\mathrm{d}Q_\epsilon(\boldsymbol{\xi}_1) := \frac{e^{-\|\boldsymbol{\xi}_1\|^p/\epsilon} \cdot \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{\xi}_1)}{\int_{\mathbb{R}^{d_x}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{u})}, \quad \mathrm{d}W_\epsilon(\boldsymbol{\xi}_2) := \frac{e^{-\|\boldsymbol{\xi}_2\|^p/\epsilon} \cdot \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{\xi}_2)}{\int_{\mathbb{R}^{d_y}} e^{-\|\boldsymbol{u}\|^p/\epsilon} \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{u})}, \quad (15)$$

and a constant

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

Note that when $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$ are taken as the Lebesgue measures, the kernel distributions Q_ϵ and W_ϵ of $\boldsymbol{\xi}_1 := \boldsymbol{x} - \widehat{\boldsymbol{x}}$ and $\boldsymbol{\xi}_2 := \boldsymbol{y} - \widehat{\boldsymbol{y}}$, respectively, are independent of the historical observations due to the translation invariance of the Lebesgue measure.

As $(\widehat{\boldsymbol{x}} + \boldsymbol{\xi}_1) \in \mathcal{X}$ and $(\widehat{\boldsymbol{y}} + \boldsymbol{\xi}_2) \in \mathcal{Y}$, the problem (14) can be reformulated as

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

where

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

which is the same as (3).

B.3 Discussion on Causal Sinkhorn DRO

In this subsection, we discuss the CSD-based ambiguity set and Causal-SDRO.

Remark 8. (Connection with KL-Divergence-based DRO). *When $\epsilon > 0$, the constraint $R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p \leq \rho^p$ in (Causal-SDRO) can be rewritten as (see Appendix G.2, Proof of Theorem 3 for details)*

$$\mathbb{E}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \sim \widehat{\mathbb{P}}} \left[\mathbb{D}_{\mathrm{KL}} \left(\gamma_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})} || \mathcal{K}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), \epsilon} \right) \right] \leq \frac{\bar{\rho}}{\epsilon}, \quad (17)$$

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 Eq. (4):

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

Compared with the KL-divergence-based DRO problem (Ben-Tal et al., 2013; Blanchet et al., 2023) with constraint $\mathbb{D}_{\mathrm{KL}}(\mathbb{P}||\widehat{\mathbb{P}}) \leq \rho$, in constraint (17), 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 $\widehat{\mathbb{P}}$. ♣

The following remark shows the relationship between the Causal-SDRO and causal Wasserstein DRO proposed in Yang et al. (2026).

Remark 9. (Comparison with Causal Wasserstein DRO). *Suppose additionally that, for each $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})$, the function $\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))$ is lower semicontinuous on $\operatorname{supp}\nu_{\mathcal{X}} \times \operatorname{supp}\nu_{\mathcal{Y}}$. If $\epsilon \to 0$, then the dual objective of the problem (14) converges to (see Appendix G.5 for detailed proof)*

$$\lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\sup_{\boldsymbol{x}\in supp\,\nu_{\mathcal{X}}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\left[\sup_{\boldsymbol{y}\in supp\,\nu_{\mathcal{Y}}}\left\{\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\right\} \mid \widehat{\boldsymbol{X}}\right]\right\}\right]. \quad (18)$$

When $\mathcal{X} = \mathbb{R}^{d_x}$ and $\mathcal{Y} = \mathbb{R}^{d_y}$, and $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$ are chosen as the corresponding Lebesgue measures, we have $\operatorname{supp}\nu_{\mathcal{X}} = \mathcal{X}$ and $\operatorname{supp}\nu_{\mathcal{Y}} = \mathcal{Y}$. Therefore, Eq. (18) coincides with the dual formulation of Causal Wasserstein DRO (Causal-WDRO) in Yang et al. (2026). The optimization for Causal-WDRO is computationally challenging due to the nested inner supremums within expectations in (18), which leads to a non-smooth stochastic min-max structure. Existing algorithms for solving Causal-WDRO typically focus on special cases. For example, Yang et al. (2026) assume a specific structure for the loss function, and Hu et al. (2023) assume that λ is sufficiently large, thereby converting the problem into a contextual stochastic bilevel optimization problem with a strongly convex lower-level problem. ♣

B.4 Discussion on the Worst-case Distribution of Causal-SDRO

As demonstrated in the proof of Theorem 3 (see Appendix G.2 for details), Theorem 10 characterizes the worst-case distribution of the primal problem v_{P}.

Theorem 10. (Worst-case Distribution of Causal-SDRO). *Under Assumption 1, suppose the dual problem v_{D} has an optimal solution $\lambda^* > 0$ and $\Psi(f(\cdot),\cdot)$ is not a constant function. Then the dual optimal solution λ^* is unique, and the density of the worst-case distribution $\mathbb{P}^*$ of the primal problem v_{P} is given by*

$$\frac{\mathrm{d}\mathbb{P}^*(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\Big[\alpha_{\widehat{\boldsymbol{x}}} \cdot \beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}} \cdot e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x}) + s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\Big], \quad (19)$$

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 10 in Appendix G.3. Based on this, Theorem 5 shows the worst-case distribution of soft-constrained Causal-SDRO and can be seen as a corollary of Theorem 10.

C Compared DRO Models

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

C.1 SDRO

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

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

where

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

We present the worst-case distribution for problem (SDRO) in the following Lemma 11, extending the results of Wang et al. (2025).

Lemma 11. (Worst-case Distribution of the SDRO Problem, Wang et al., 2025). *Under Assumption 1, the density of worst-case distribution $\mathbb{P}^*_{\lambda,\mathrm{SDRO}}$ of the inner problem of (SDRO) for any λ is given by*

$$\frac{\mathrm{d}\mathbb{P}^*_{\lambda,\mathrm{SDRO}}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\Big[\tilde{\alpha}_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}}(\lambda)\cdot e^{s'(\lambda,\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})}\Big], \tag{20}$$

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

According to Lemma 11, $\mathbb{P}^*_{\lambda,\mathrm{SDRO}}$ is a mixture of Gibbs distributions. Compared with the worst-case distribution of the problem (Causal-SDRO) in Theorem 10, $\mathbb{P}^*_{\lambda,\mathrm{SDRO}}$ has a simpler density function structure.

According to Wang et al. (2025), the strong dual formulation of the problem (SDRO) can be reformulated as follows when a parametric decision rule $f_{\boldsymbol{\theta}}$ is considered:

$$\min_{\boldsymbol{\theta}\in\Theta} \quad \mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\left[\lambda\epsilon\log\int_{\mathcal{X}\times\mathcal{Y}}\exp\left(\frac{\Psi(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\lambda\epsilon}\right)\mathrm{d}\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}(\boldsymbol{x},\boldsymbol{y})\right]. \tag{21}$$

We define

$$\mathrm{d}U_\epsilon(\boldsymbol{\xi}_3,\boldsymbol{\xi}_4) := \frac{e^{-\frac{\|\boldsymbol{\xi}_3\|^p+\|\boldsymbol{\xi}_4\|^p}{\epsilon}}}{\mathbb{E}_{(\boldsymbol{\xi}_3,\boldsymbol{\xi}_4)\sim\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}}\left[e^{-\frac{\|\boldsymbol{\xi}_3\|^p+\|\boldsymbol{\xi}_4\|^p}{\epsilon}}\right]}\cdot\mathrm{d}\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}(\boldsymbol{\xi}_3,\boldsymbol{\xi}_4),$$

and then Eq. (21) is equivalent to

$$\min_{\boldsymbol{\theta}\in\Theta} \quad C\cdot\mathbb{E}_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\sim\widehat{\mathbb{P}}}\Big[\lambda\epsilon\cdot\log\,\mathbb{E}_{(\boldsymbol{\xi}_3,\boldsymbol{\xi}_4)\sim U_\epsilon}\Big[\exp\left(\frac{\Psi(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}+\boldsymbol{\xi}_3),\widehat{\boldsymbol{y}}+\boldsymbol{\xi}_4)}{\lambda\epsilon}\right)\Big]\Big] \tag{22}$$

where $C = \mathbb{E}_{(\boldsymbol{\xi}_3,\boldsymbol{\xi}_4)\sim\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}}}\Big[e^{-(\|\boldsymbol{\xi}_3\|^p+\|\boldsymbol{\xi}_4\|^p)/\epsilon}\Big]$. The problem (22) is a stochastic compositional optimization problem and can be efficiently solved by the proposed SCSC algorithm.

C.2 CAUSAL-WDRO

Without considering the entropy regularization, the Causal-WDRO model is defined as Yang et al. (2026):

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

where the causal transport distance $C_p(\widehat{\mathbb{P}},\mathbb{P})$ is defined in Definition 1. Yang et al. (2026) 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 $\widehat{\boldsymbol{y}}_{k,i}$ for any $i\in[n_k]$. Let $\widehat{p}_{k,i}$ be the probability mass of the data point $(\widehat{\boldsymbol{x}}_k,\widehat{\boldsymbol{y}}_{k,i})$. Then, according to Yang et al. (2026), the following lemma holds.

Lemma 12. (Worst-case Distribution of the Causal-WDRO Problem, Yang et al., 2026). *If the worst-case distribution of Causal-WDRO problem exists for given $\lambda > 0$, then it has the following form*

$$\mathbb{P}^*_{\lambda,\text{Causal-WDRO}} = \sum_{k\neq k_0}\sum_{i=1}^{n_k} \widehat{p}_{k,i}\widehat{\mathbb{P}}_{(\boldsymbol{x}^*_k(\lambda),\boldsymbol{y}^*_{k,i}(\lambda))} + \sum_{i=1}^{n_{k_0}} \widehat{p}_{k_0 i}\left(q\widehat{\mathbb{P}}_{(\bar{\boldsymbol{x}}_{k_0}(\lambda),\bar{\boldsymbol{y}}_{k_0 i}(\lambda))} + (1-q)\widehat{\mathbb{P}}_{(\underline{\boldsymbol{x}}_{k_0}(\lambda),\underline{\boldsymbol{y}}_{k_0 i}(\lambda))}\right), \tag{23}$$

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

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

and

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

The worst-case distribution of the problem (Causal-WDRO) in Eq. (23) is discrete, while that of Causal-SDRO in Eq. (19) is continuous. This shows that the introduction of CSD allows a more realistic and smoother representation of the underlying distribution.

We solve the problem (Causal-WDRO) as a contextual stochastic bilevel optimization problem by the Random Truncated Multilevel Monte Carlo (RT-MLMC) approach proposed by Hu et al. (2023).

C.3 KL-DRO

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

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

For problem (KL-DRO), the worst-case distribution is given by Lemma 13.

Lemma 13. (Worst-case Distribution of the KL-DRO Problem, Hu & Hong, 2013). *If the worst-case distribution of KL-DRO problem exists for given $\lambda > 0$, then it has the following form*

$$\mathbb{P}^*_{\lambda,\text{KL-DRO}} = \sum_{i=1}^{N} \frac{\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}_i),\widehat{\boldsymbol{y}}_i)}{\lambda}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\Psi(f(\widehat{\boldsymbol{x}}_j),\widehat{\boldsymbol{y}}_j)}{\lambda}\right)} \cdot \widehat{\mathbb{P}}_{(\widehat{\boldsymbol{x}}_i,\widehat{\boldsymbol{y}}_i)}, \tag{24}$$

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 in (KL-DRO) remains discrete.

When a parametric decision rule $f_{\boldsymbol{\theta}}$ is considered, according to Fenchel duality, the problem (KL-DRO) is equivalent to

$$\min_{\boldsymbol{\theta}\in\Theta} \quad \lambda\cdot\log\,\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\widehat{\mathbb{P}}}\Big[\exp\Big(\frac{\Psi(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{y})}{\lambda}\Big)\Big], \tag{25}$$

which can be solved by the standard SGD method.

D Interpretability of the Soft Regression Forest

In this section, we demonstrate the intrinsic interpretability of SRF through its transparent structure and stability for decision-making (i.e., Lipschitz continuity and smoothness) in Appendix D.1, and we introduce both global and local intrinsic interpretation measures for SRF that depend only on its structure and avoid post hoc explanation analyses in Appendix D.2.

D.1 Interpretability as a Tree-based Model

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

Remark 14. (Asymptotic Consistency to Hard Regression Forest). *Adding a scaling parameter τ to the linear transformation part at each internal node in all SRTs, that is, $S\Big((\boldsymbol{w}^\top \boldsymbol{x} + b)/\tau\Big)$ for all internal nodes $j_t \in [2^{D(t)} - 1]$, we obtain a variant of SRF decision rule termed $f_{\theta,\tau}^{SRF}(\boldsymbol{x})$ where each leaf node $l_t \in [2^{D(t)}]$ in tree t can be reached with probability $p_{l_t,t,\tau}$. The proposed SRF $f_\theta^{SRF}(\boldsymbol{x})$ shown in Section 4 is a special case with $\tau = 1$. For any input covariate $\boldsymbol{x}$ not lying on any decision boundary, that is, $\{\boldsymbol{x} \in \mathcal{X} \mid \boldsymbol{w}_{j_t,t}^\top \boldsymbol{x} + \boldsymbol{b}_{j_t,t} \neq 0, \forall j_t, t\}$, 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, that is, $\sum_{i=1}^{2^{D(t)}} p_{i,t,\tau} = 1$ 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 every node, thereby improving accuracy and interpretability by reducing tree depth (Bertsimas & Dunn, 2017; Bertsimas & 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 Appendix F.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 15. (Traceability of Decisions in SRF). *In SRF, given an input covariate $\boldsymbol{x} \in \mathbb{R}^{d_x}$, for each route selected with probability $p_{l_t,t}(\boldsymbol{x})$, the marginal contribution of each feature and the interaction effects among features along that route are explicitly characterized by*

$$\frac{\partial p_{l_t,t}(\boldsymbol{x})}{\partial x_j} = p_{l_t,t}(\boldsymbol{x}) \cdot \sum_{i\in\Lambda(l_t)} \psi_{i,t}\Big[\boldsymbol{w}_{i,t}\Big]_j,$$

$$\frac{\partial^2 p_{l_t,t}(\boldsymbol{x})}{\partial x_j \partial x_k} = p_{l_t,t}(\boldsymbol{x}) \cdot \Bigg[\Big(\sum_{i\in\Lambda(l_t)} \psi_{i,t}\Big[\boldsymbol{w}_{i,t}\Big]_j\Big)\Big(\sum_{i\in\Lambda(l_t)} \psi_{i,t}\Big[\boldsymbol{w}_{i,t}\Big]_k\Big) - \sum_{i\in\Lambda(l_t)} S\Big(\boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}\Big)\Big(1 - S\Big(\boldsymbol{w}_{i,t}^\top \boldsymbol{x} + b_{i,t}\Big)\Big)\Big[\boldsymbol{w}_{i,t}\Big]_j\Big[\boldsymbol{w}_{i,t}\Big]_k\Bigg],$$

$$\forall j, k \in [d_x],\ l_t \in [2^{D(t)}],\ t \in T,$$

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

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

We provide the proof of Proposition 15 in Appendix D.1. 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. Define $W_{\max} := \max_{i,t} \|\boldsymbol{w}_{i,t}\|$ as the maximum norm of the internal node weights, $\Pi_{\max} := \max_{l_t,t} \|\boldsymbol{\pi}_{l_t,t}\|$ as the maximum norm of leaf vectors, and $D_{\max}$ as the maximum tree depth ($D_{\max} \geq 2$). Then, the following proposition holds.

Proposition 16. (Lipschitz Continuity and Smoothness of SRF). *The SRF decision rule* $f_{\boldsymbol{\theta}}^{\text{SRF}} : \mathcal{X} \to \mathcal{Z}$ *is globally* L_{SRF}*-Lipschitz continuous and* S_{SRF}*-Lipschitz smooth, where*

$$L^{\text{SRF}} = D_{\max} W_{\max} \Pi_{\max}, \quad S^{\text{SRF}} = D_{\max}(D_{\max} + \frac{1}{4}) W_{\max}^2 \Pi_{\max}.$$

We provide the proof of Proposition 16 in Appendix G.7. 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 & Jaakkola, 2018).

All analyses above demonstrate the structural transparency, stability, and robustness of the SRF decision rule. In Appendix D.2, 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 Appendix F.3, we confirm the practical interpretability of SRT, based on its structure and the proposed intrinsic interpretation measures for the portfolio problem using real data.

D.2 Interpretability Measures

In this subsection, we analyze the intrinsic interpretability of the SRT by introducing global and local interpretation measures, as detailed in Appendices D.2.1 and D.2.2, respectively.

D.2.1 Global Interpretation Measure

Global interpretability provides a holistic view of the model by quantifying each feature's contribution 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 & 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 each decision with respect to each feature over the training set:

$$\mathcal{C}_{j,k} := \frac{1}{N} \sum_{i=1}^{N} \left\| \frac{\partial \left[f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x}^i) \right]_k}{\partial x_j^i} \right\|_1 = \frac{1}{N \cdot T} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{l_t=1}^{2^{D(t)}} \left\| \frac{\partial p_{l_t,t}(\boldsymbol{x}^i)}{\partial x_j^i} [\boldsymbol{\pi}_{l_t,t}]_k \right\|_1, \quad \forall j \in [d_x], k \in [d_z],$$

where $\boldsymbol{x}^i$ is the i-th training sample and x_j^i is its j-th element, and the partial derivatives of $p_{l_t,t}(\boldsymbol{x}^i)$ are computed following Proposition 15. To show the relative importance of each feature to the k-th decision component, we normalize the importance scores $\mathcal{C}_{j,k}$ across all features $j \in [d_x]$ for each $k \in [d_z]$ such that they sum to one. The relative feature importance is then defined as

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

D.2.2 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 & 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 only be used after the decision is made and does not exploit the model's structure.

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 proposed by 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 & 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:

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

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, similar to Erion et al. (2021), we define the EIG for feature j and decision k as

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

where $\boldsymbol{x}^i$ denotes the i-th training sample. Proposition 17 shows that the k-th output of SRF can be decomposed into the sum of EIG contributions from each feature, with the baseline defined as the average value of the k-th output over the data set, that is,

$$\text{Baseline}_k = \frac{1}{N} \sum_{i=1}^{N} \Big[f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x}^i) \Big]_k.$$

Proposition 17. *Since the decision rule $f_{\boldsymbol{\theta}}^{\text{SRF}} : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$ is differentiable, given an input covariate $\boldsymbol{x}$, the EIG value $\varphi_{j,k}^{\text{EIG}}(\boldsymbol{\theta}, \boldsymbol{x})$ exactly quantifies the contribution of feature j to the deviation of the decision from the average training decision baseline:*

$$\Big[f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x}) \Big]_k - \frac{1}{N} \sum_{i=1}^{N} \Big[f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x}^i) \Big]_k = \sum_{j=1}^{d_x} \varphi_{j,k}^{\text{EIG}}(\boldsymbol{\theta}, \boldsymbol{x}), \quad \forall k \in [d_z].$$

Proof of Proposition 17. According to the Proposition 1 in Sundararajan et al. (2017), for any $\boldsymbol{x}' \in \mathbb{R}^{d_x}$, we have

$$\begin{aligned} &\Big[f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x}) \Big]_k - \Big[f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x}') \Big]_k \\ &\qquad = \sum_{j=1}^{d_x} (x_j - x_j') \cdot \int_{\alpha=0}^{1} \frac{\partial \Big[f_{\boldsymbol{\theta}}^{\text{SRF}} \Big(\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}') \Big) \Big]_k}{\partial x_j} \mathrm{d}\alpha, \quad \forall k \in [d_z]. \end{aligned}$$

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

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

□

Regarding local feature interactions, the SRT's differentiability allows us to explicitly characterize interaction effects by computing the Hessian matrix, as detailed in Proposition 15.

E Analyses for the Stochastic Compositional Optimization and Algorithm

In this section, we analyze the properties of (SCO) in Appendix E.1, establish the sample complexity of the sample average approximation method in Appendix E.2, and provide the motivation for developing a new gradient-based algorithm to solve (SCO) in Appendix E.3.

E.1 PROPERTIES OF SCO

For brevity, we denote functions

$$t_1(z) := t_1(z; \widehat{\boldsymbol{x}}), \quad t_2(\boldsymbol{v}) := t_2(\boldsymbol{v}; \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1), \quad t_3(\boldsymbol{\theta}) := t_3(\boldsymbol{\theta}; \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2), \tag{26}$$

and define

$$\phi^{(0)}(\boldsymbol{\theta}) := \mathbb{E}_{\widehat{\boldsymbol{x}}}\Big[t_1(\phi^{(1)}(\boldsymbol{\theta}))\Big], \quad \phi^{(1)}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_1}\Big[t_2(\phi^{(2)}(\boldsymbol{\theta}))\Big], \quad \phi^{(2)}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3(\boldsymbol{\theta})\Big]. \tag{27}$$

Under Assumption 2, for the problem (SCO), we have the following Proposition 18.

Proposition 18. (Properties of the Problem SCO). *Under Assumption 2, functions t_1, t_2, and t_3 in Eq. (9) are restricted to $[1, \exp(B/\lambda\epsilon)]$, $[1, \exp(B/\lambda\epsilon)]^N$, and Θ, respectively. On these domains, the following properties hold.*

(I) *t_1, t_2, and t_3 are respectively L_1-, L_2-, and L_3-Lipschitz continuous;*

(II) *t_1, t_2, and t_3 are respectively S_1-, S_2-, and S_3-Lipschitz smooth;*

(III) *Their first-order derivatives have uniformly bounded second moments, i.e.,*

$$\mathbb{E}\Big[|\nabla t_1(z)|^2\Big] \le C_1^2, \ \mathbb{E}\Big[\|\nabla t_2(\boldsymbol{v})\|^2\Big] \le C_2^2, \ \mathbb{E}\Big[\|J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta})\|_F^2\Big] \le C_3^2;$$

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

Here, $J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta})$ denotes the Jacobian of function t_3 with respect to $\boldsymbol{\theta}$, and

$$L_1 = S_1 = C_1 = 1; \qquad L_2 = C_2 = \exp\Big(\frac{B}{\lambda\epsilon}\Big), \quad S_2 = \sqrt{2} L_2;$$
$$L_3 = C_3 = \frac{\sqrt{N} L_{\boldsymbol{\theta}}}{\lambda\epsilon} \exp\Big(\frac{B}{\lambda\epsilon}\Big), \quad S_3 = \sqrt{N} e^{\frac{B}{\lambda\epsilon}} \Big(\frac{S_{\boldsymbol{\theta}}}{\lambda\epsilon} + \frac{L_{\boldsymbol{\theta}}^2}{\lambda^2\epsilon^2}\Big).$$

We provide the proof of Proposition 18 in Appendix G.8. These properties are used for the analyses in Appendices E.2 and E.3

E.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 empirical average based on finite samples generated via Monte Carlo sampling. In this section, we analyze the sample and computational complexity of SAA on our problem.

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):

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

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, that is, $\Pr\Big(F\Big(\widehat{\boldsymbol{\theta}}_{N_1, N_2, N_3}\Big) - F(\boldsymbol{\theta}^*) \le \delta\Big) \ge 1 - \alpha$ for any $\delta > 0$ and $\alpha \in (0, 1)$.

Based on Assumption 2, we derive the sample complexity of the SAA method for this problem in Theorem 19.

Theorem 19. (Sample Complexity for the 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*

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

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

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

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

and

$$N_3 > \frac{\mathcal{O}\left(1\right)L_1^2L_2^2\sigma_3^2}{\delta^2}\Big[d_\theta\log\left(\frac{8L_1L_2L_3D_\Theta}{\delta}\right)+\log\left(\frac{1}{\alpha}\right)+\log\left(N_1^2N_2\right)\Big].$$

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

The proof of Theorem 19 is provided in Appendix G.9. For problem (SAA), the computational complexity of gradient descent (GD) is at least $\mathcal{O}\Big(d_\theta^3/\delta^6\Big)$, 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 $\mathcal{O}\Big(d_\theta^2/\delta^4\Big)$, which is still computationally challenging.

E.3 The Motivation of SCSC

In this subsection, we show another challenge of applying the standard SGD algorithm to problem (SCO). In the k-th iteration, the true gradient of F at $\boldsymbol{\theta}^k$ is given by

$$\nabla F(\boldsymbol{\theta}^k) = \nabla t_1\Big(\phi^{(1)}(\boldsymbol{\theta}^k)\Big)\cdot\nabla t_2\Big(\phi^{(2)}(\boldsymbol{\theta}^k)\Big)\cdot\nabla t_3(\boldsymbol{\theta}^k),$$

where functions t_1, t_2, t_3, $\phi^{(1)}$, and $\phi^{(2)}$ are defined in Eqs. (26)and (27). 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

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

The SGD method replaces $t_2(\phi^{(2)}(\boldsymbol{\theta}^k))$ by $t_2^k(t_3^k(\boldsymbol{\theta}^k))$, and $t_1(\phi^{(1)}(\boldsymbol{\theta}^k))$ by $t_1^k(t_2^k(t_3^k(\boldsymbol{\theta}^k)))$ 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 generally non-linear, the stochastic gradient of the SGD method, denoted as $\Big(\nabla F(\boldsymbol{\theta}^k)\Big)_{\text{SGD}}$, is biased, that is,

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

Since the bias of standard SGD is uncontrollable, it cannot be used to solve problem (SCO) directly. Therefore, we present a new gradient-based method for solving (SCO) in Section 5.2 that controls bias using a momentum-based gradient update.

F Detailed Numerical Results

As a supplement to Section 6, in this section, we first provide details of our experimental settings and present detailed numerical results for three feature-based applications: the newsvendor problem

in Appendix F.1, the inventory substitution problem in Appendix F.2, and the portfolio selection problem with real data in Appendix F.3. Additionally, we demonstrate the interpretability of the SRF decision rule for the portfolio problem in Appendix F.3.3.

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). We take $T = 20, D(t) = \lceil \log_2 d_x \rceil + 1$ for each $t \in [T]$ for the SRF decision rule.

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 $f^{\text{2NN}}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Z}$, 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

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

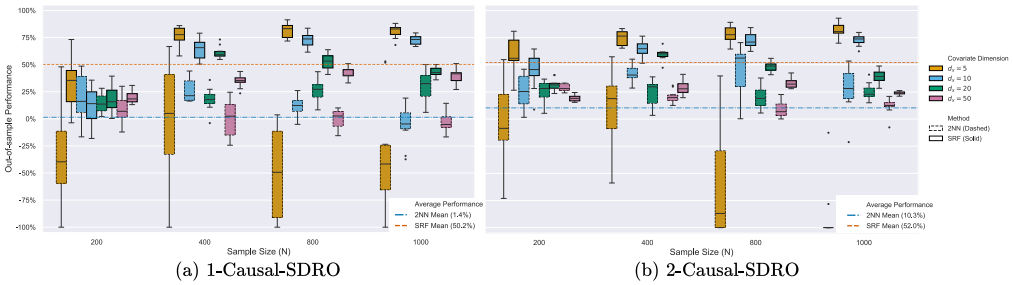
where $\text{ReLU} : \mathbb{R} \to \mathbb{R}_+$ represents the ReLU activation function, while a_i^k and b_i^k represent the i-th element in vectors $\boldsymbol{a}^k$ and $\boldsymbol{b}^k$, respectively, for any $i \in [m]$. This decision rule requires training a total of $m \cdot d_z(d_x + 2)$ parameters. 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.

F.1 Feature-based Newsvendor Problem

In this subsection, we provide more details and results for the feature-based newsvendor problem, which adopts an experimental setup similar to that of Yang et al. (2026), 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 using the acceptance-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.

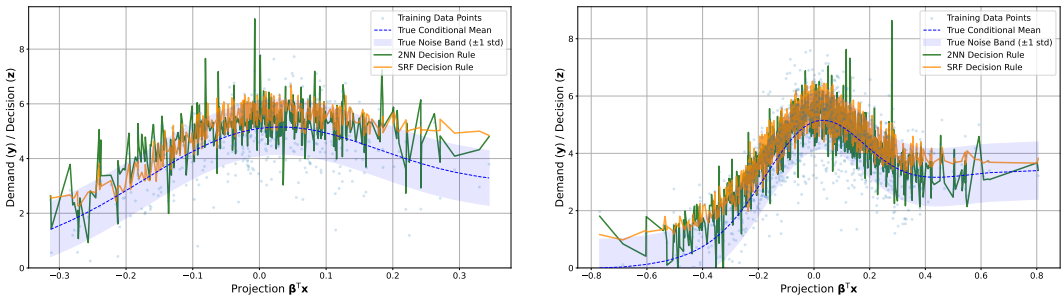


(a) 1-Causal-SDRO (b) 2-Causal-SDRO

Figure 5: Out-of-sample performance of the decision rules on the newsvendor problem

Figure 5 compares the out-of-sample performance of SRF and 2NN decision rules 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

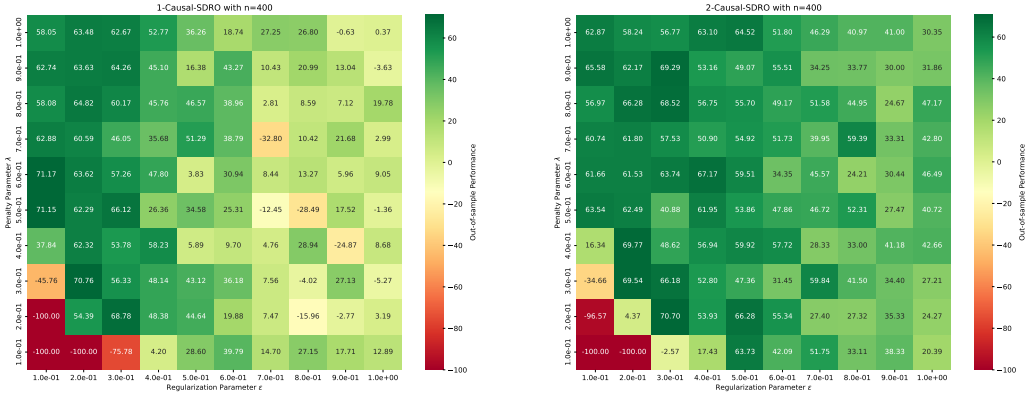
prescriptiveness scores of 50.2% (1-Causal-SDRO) and 52.0% (2-Causal-SDRO), marking a significant advantage over the 2NN, which yields only 1.4% and 10.3%, respectively. Beyond average performance, the box plots reveal that the SRF exhibits significantly lower variance (as indicated by shorter interquartile ranges) than 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) $N = 400$ (b) $N = 1000$

Figure 6: True distribution vs. Trained decision rules for 2-Causal-SDRO ($d_x = 10$)

Figure 6 visualizes the fitted 2NN and SRF decision rules against the true conditional mean (blue dashed line) for sample sizes $N \in \{400, 1000\}$. 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.



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

Figure 7: Out-of-sample performance of the newsvendor problem with different parameters ($N = 400, d_x = 10$)

Figure 7 reports the out-of-sample performance of the proposed method for the Causal-SDRO model across different parameter combinations, including the penalty parameter λ, the regularization parameter ϵ, and the norm p, using instances with $N = 400$ and $d_x = 10$. As shown in these plots, although both models achieve positive out-of-sample performance on almost all instances, the 2-Causal-SDRO exhibits higher out-of-sample performance across most parameter combinations. These results indicate that performance improves with a moderate increase in λ and a decrease in 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.

F.2 Feature-based Inventory Substitution Problem

In this subsection, we conduct a numerical study on the feature-based inventory substitution problem. This application serves as a representative two-stage contextual DRO problem. Appendix F.2.1 introduces the model and its tractable reformulation, and Appendix F.2.2 provides the experimental setting and numerical results.

F.2.1 Model and Reformulation for Feature-based Inventory Substitution Problem

This problem, as a variant of the supply chain substitution problem in Chen & Gao (2019), is a two-stage optimization problem. Consider a firm selling d_z types of products to satisfy customer demand, indexed by $i \in [d_z]$, where a lower index denotes 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 "here-and-now" 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 "wait-and-see" decision is to allocate the inventory to each demand class, targeting the lowest total cost. Let the decision variable $w_{i,j} \geq 0$ denote the quality of product i as a substitute for j, while $h_i \geq 0$ and $b_j \geq 0$ denote the unit holding cost of product i and the shortage cost for demand j, respectively. Taking features $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ (for example, demographic data) into account, this two-stage contextual DRO problem is given by

$$\inf_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathfrak{R}(\widehat{\mathbb{P}})} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\boldsymbol{c}^\top f(\boldsymbol{x}) + \Psi_{\text{Inventory}}(f(\boldsymbol{x}), \boldsymbol{y})\Big]$$

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

$$\begin{aligned}
\Psi_{\text{Inventory}}(\boldsymbol{z}, \boldsymbol{y}) = \min \quad & \sum_{j=1}^{d_y} \sum_{i=1}^{j} s_{i,j} w_{i,j} + \sum_{i=1}^{d_z} h_i u_i + \sum_{j=1}^{d_y} b_j u'_j & \\
\text{s.t.} \quad & \sum_{j=i}^{d_y} w_{i,j} + u_i = z_i, & \forall i \in [d_z], \\
& \sum_{i=1}^{j} w_{i,j} + u'_j = y_j, & \forall j \in [d_y], \\
& u_i, u'_j, w_{i,j} \geq 0, & \forall i \in [d_z], j \in [d_y].
\end{aligned} \tag{28}$$

where u_i represents the leftover inventory of product i, u'_j represents the shortage of demand j, and the three terms in $\Psi_{\text{Inventory}}(\boldsymbol{z}, \boldsymbol{y})$ represent the total purchasing cost, total holding cost, and total shortage cost, respectively.

We next focus on the feature-based inventory substitution problem with a soft CSD constraint, which is given by

$$\inf_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathfrak{R}(\widehat{\mathbb{P}})} \boldsymbol{c}^\top f(\boldsymbol{x}) + \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi_{\text{Inventory}}(f(\boldsymbol{x}), \boldsymbol{y})\Big] - \lambda \cdot R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p. \tag{29}$$

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

$$\begin{aligned}
\max_{\boldsymbol{\eta} \in \mathbb{R}^{d_z}, \boldsymbol{\upsilon} \in \mathbb{R}^{d_y}} \quad & \sum_{i=1}^{d_z} z_i \eta_i + \sum_{j=1}^{d_y} y_j \upsilon_j & \\
\text{s.t.} \quad & \eta_i \leq h_i, & \forall i \in [d_z], \\
& \upsilon_j \leq b_j, & \forall j \in [d_y], \\
& \eta_i + \upsilon_j \leq s_{i,j}, & \forall j \in \{i, i+1, \cdots, d_y\}, i \in [d_z].
\end{aligned} \tag{30}$$

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 the problem (28), while $\upsilon_j \in \mathbb{R}$ for each $j \in [d_y]$ represents the dual variable of the j-th

constraint in the second constraint set of the problem (28). According to the duality theorem, strong duality holds. Define that

$$\Psi^*_{\text{Inventory}}(f(\boldsymbol{x}), \boldsymbol{y}) := \max_{\boldsymbol{\eta}\in\mathbb{R}^{d_z}, \boldsymbol{\nu}\in\mathbb{R}^{d_y}} \left\{ \sum_{i=1}^{d_z} [f(\boldsymbol{x})]_i(\eta_i + c_i) + \sum_{j=1}^{d_y} y_j \nu_j \;\middle|\; \begin{array}{ll} \eta_i \le h_i, & \forall i \in [d_z], \\ \nu_j \le b_j, & \forall j \in [d_y], \\ \eta_i + \nu_j \le s_{i,j}, & \forall j \in \{i, i+1, \cdots, d_y\}, i \in [d_z] \end{array} \right\},$$

and then the problem (29) can be rewritten as

$$\inf_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\,\Re\,(\widehat{\mathbb{P}})} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\left[\Psi^*_{\text{Inventory}}(f(\boldsymbol{x}), \boldsymbol{y})\right] - \lambda \cdot R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p,$$

which is the same as (Soft-Causal-SDRO). Therefore, following the same reformulation processes as in Theorem 3, its dual formulation is given by

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

where

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

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

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

where

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

We compute the optimal value of $\Psi^*_{\text{Inventory}}(f(\boldsymbol{x}), \boldsymbol{y})$ and its associated gradient with respect to $f(\boldsymbol{x})$ using the commercial solver Gurobi (version 12.0.1).

F.2.2 Performance of the Proposed Method

We examine a scenario with $d_z = d_y = 3$ products, varying feature dimensions $d_x \in \{3, 5, 10, 20\}$ 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:

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

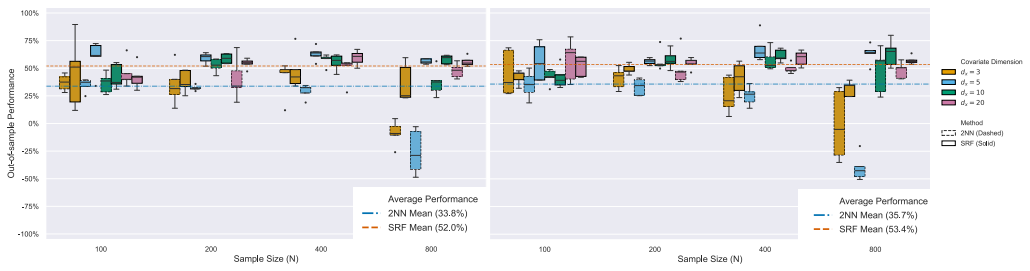
where $\boldsymbol{\beta} \in \mathbb{R}^{d_x}$ is sampled from $\mathcal{U}([-0.1, 0.1])$, and $\boldsymbol{x} \in \mathbb{R}^{d_x}$ is constructed by the procedure in Appendix F.1. We specify the cost parameters as

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

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

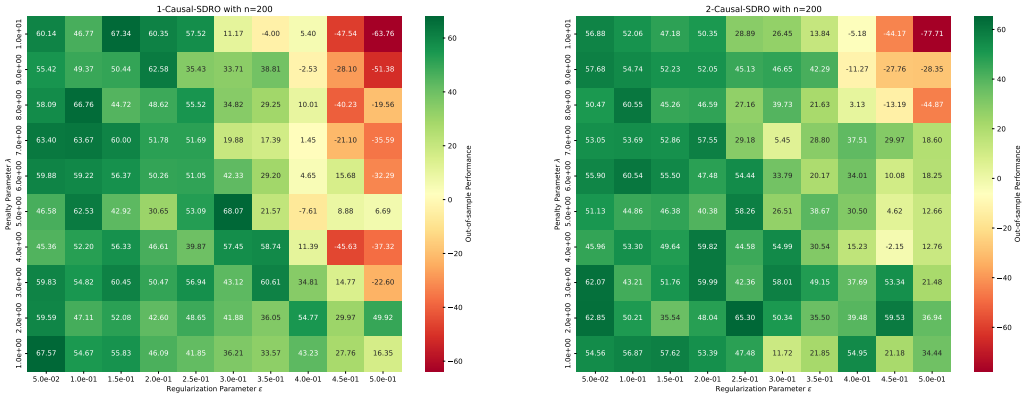
Figure 8 reports the out-of-sample performance of the proposed approach across varying p, N, and d_x. The SRF decision rule demonstrates superior efficacy, achieving average prescriptiveness scores of 48.5% (1-Causal-SDRO) and 49.4% (2-Causal-SDRO), while consistently maintaining positive out-of-sample performance across all instances. In comparison, the 2NN benchmark 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.

Figure 9 illustrates the out-of-sample performance of the proposed Causal-SDRO method across different combinations of regularization hyperparameters, meaning that Causal-SDRO models perform well across a relatively wide range of hyperparameter choices. Figure 10 compares Causal-SDRO against other DRO benchmarks. These numerical results are similar to those obtained in the experiments of the feature-based newsvendor problem.



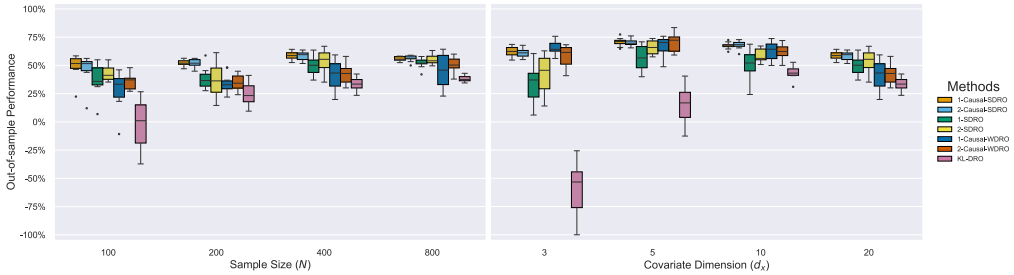
(a) 1-Causal-SDRO (b) 2-Causal-SDRO

Figure 8: Out-of-sample performance of the decision rules on the inventory substitution problem



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

Figure 9: Out-of-sample Performance of the inventory substitution problem with different parameters ($N = 200, d_x = 10$)



(a) $d_x = 20,\ N \in \{100, 200, 400, 800\}$ (b) $N = 400,\ d_x \in \{3, 5, 10, 20\}$

Figure 10: Comparison of different DRO models on the inventory substitute problem

F.3 Data-driven Portfolio Selection Problem

This section applies the proposed method to the data-driven portfolio selection problem. Appendix F.3.1 introduces the model and dataset for the portfolio selection problem; Appendix F.3.2 presents the performance of our approach on the portfolio problem; and Appendix F.3.3 demonstrates its intrinsic interpretability from an empirical perspective.

F.3.1 Introduction for Portfolio Selection Problem

Conditioned on a covariate $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ (for example, 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 returns, and risk can be characterized by variance, conditional value-at-risk, and related measures. We choose the variance of return as the measure of risk. In this example, this leads to the following model

$$\inf_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\mathfrak{R}(\widehat{\mathbb{P}})} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi_{\text{Portfolio}}(f(\boldsymbol{x}),\boldsymbol{y})\Big],$$

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

$$\Psi_{\text{Portfolio}}(\boldsymbol{z},\boldsymbol{y}) := \left\{ -\omega\cdot\sum_{i=1}^{d_y} y_i z_i + \left(\sum_{i=1}^{d_y} y_i z_i - z_0\right)^2 \;\middle|\; \begin{array}{ll} \sum_{i=1}^{d_y} z_i = 1, & \\ z_i \geq 0, & \forall i\in[d_y] \end{array} \right\},$$

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 $\boldsymbol{z} = \big(z_0, z_1, \cdots, z_{d_y}\big)^\top \in \mathcal{Z} \subseteq \mathbb{R}^{d_y+1}$, z_0 represents the expected portfolio return while z_i represents the portfolio on the asset i for each $i \in [d_y]$.

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 399 selected assets have been in the S&P 500 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 their historical values, thereby satisfying the causal structure of our model.

F.3.2 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 (that is, $\mathcal{H}^*$) when the information of the future 60 days is completely known:

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

(II) The equal-weighted (EW) model. This model assigns equal weights to each selected asset; that is, the investment amount for each asset 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 $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}}$:

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

(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:

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

(V) Our conditional p-Causal-SDRO mean-variance (p-Causal-SDRO) model shown in Appendix F.3.1.

Let $\tilde{r}_t$ 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 Mean$(\{\tilde{r}_t\}_{t\in[60]})$, and the standard deviation of return stdDev$(\{\tilde{r}_t\}_{t\in[60]})$. (2) The portfolio loss value $\Psi_{\text{Portfolio}}$. (3) The annualized Sharpe ratio $\sqrt{252} \times$ Mean$(\{\tilde{r}_t\}_{t\in[60]})$/stdDev$(\{\tilde{r}_t\}_{t\in[60]})$. (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 Eq. (13).

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-Causal-SDRO	0.950	1.352	0.067	0.985	2.004
	2-Causal-SDRO	**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-Causal-SDRO	0.929	1.387	**0.072**	1.044	2.109
	2-Causal-SDRO	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-Causal-SDRO	**0.800**	1.353	**0.074**	1.060	2.127
	2-Causal-SDRO	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-Causal-SDRO	**0.775**	1.192	0.062	1.085	2.222
	2-Causal-SDRO	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-Causal-SDRO	0.698	1.115	0.062	1.102	2.337
	2-Causal-SDRO	**0.618**	1.253	**0.074**	1.113	2.330

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

We employ the proposed SRF as the parametric decision rule for both the CMV and Causal-SDRO models. Table 1 presents experimental results for risk aversion levels $\omega \in \{1, 3, 5, 7, 9\}$, corresponding to different trade-offs 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 observe all covariates, our results demonstrate that our approach maintains robust performance even when the observed features are imperfect.

Figure 11 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

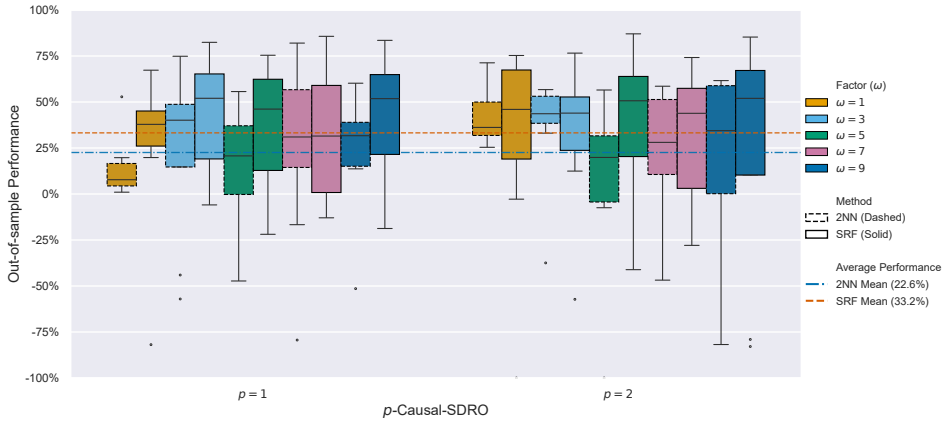
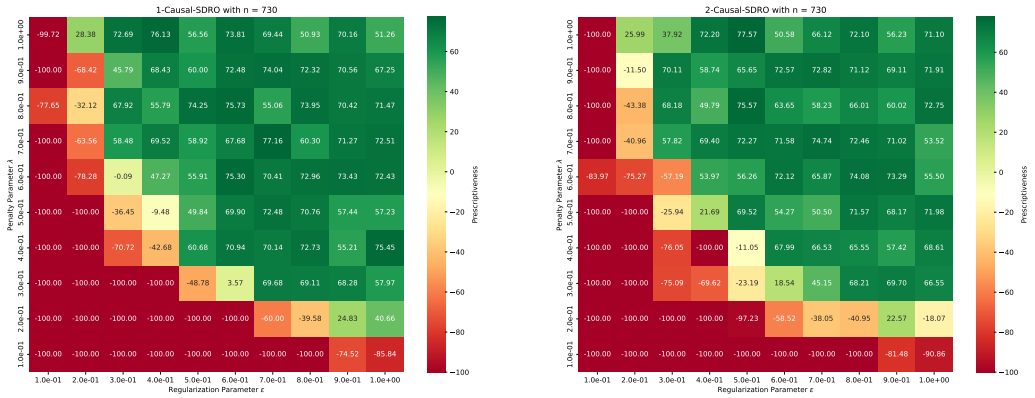


Figure 11: Out-of-sample performance of the decision rules on the portfolio problem

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.

The parameter sensitivity heatmaps for the data-driven portfolio selection problem are shown in Figure 12. As illustrated, Causal-SDRO models demonstrate robust performance across a relatively wide range of hyperparameter choices. In Figure 13, the Causal-SDRO methods achieve the best average out-of-sample performance compared with existing DRO baselines, as shown in Appendix C. These experimental results demonstrate the superior performance of our method on real data.



(a) 1-Causal-SDRO

(b) 2-Causal-SDRO

Figure 12: Out-of-sample Performance of the portfolio problem with different parameters ($\omega = 5$)

F.3.3 Interpretability of the Proposed SRF

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 14 illustrates the structure of the trained SRT, with normalized feature weights shown 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 $\boldsymbol{\pi}_8$ with weight probability (approximately) 1. The model maps $\boldsymbol{x}_2$ to decisions $\boldsymbol{\pi}_7$ and $\boldsymbol{\pi}_4$ with weight probabilities 0.996 and 0.004, respectively. These results confirm that the final decision is dominated by a few high-probability routes.

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 each feature's

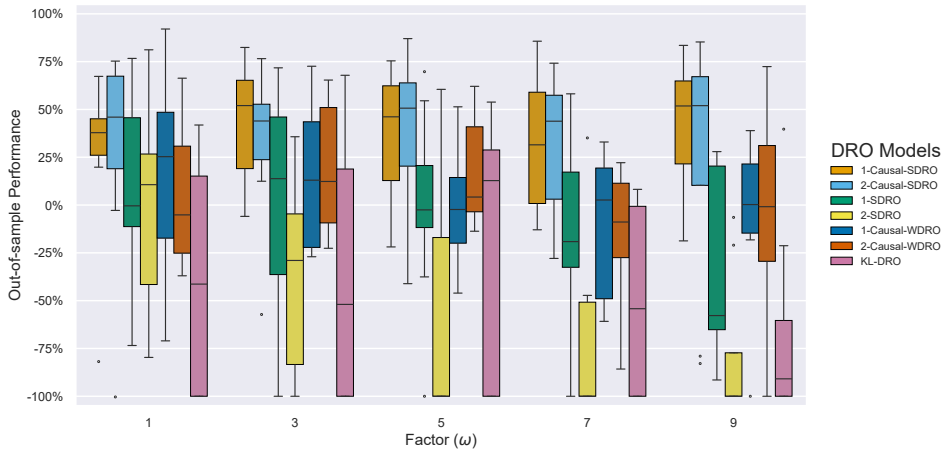


Figure 13: Comparison of different DRO models on the portfolio problem

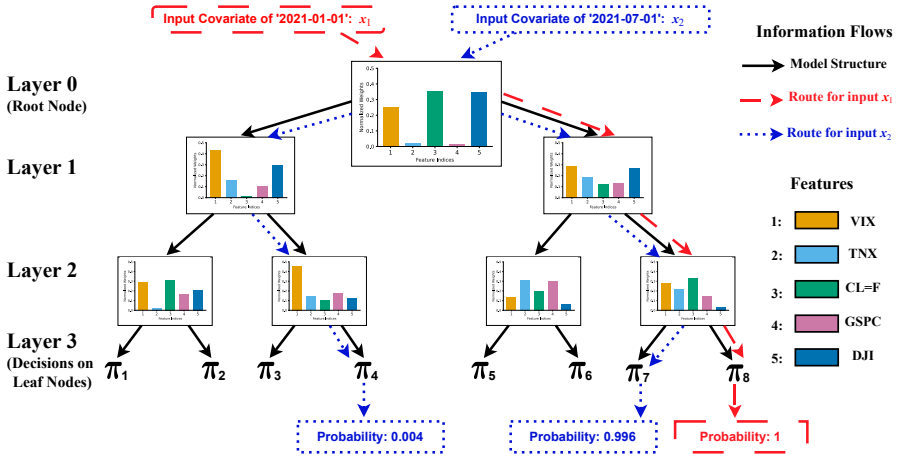


Figure 14: The structure of the trained SRT with three layers. Here, the solid black lines denote the model structure, while the red and blue dashed lines illustrate the decision routes and their corresponding selection probabilities for two input covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

contribution to the overall decision-making process, while the local interpretation shows how an individual decision is derived for a specific set of covariates (Dwivedi et al., 2023). In Appendix D.2, we introduce a global feature-importance measure and a local feature-attribution measure for SRF, both based on differentiability. In the following, we analyze the SRT using these interpretation measures and compare them with traditional post hoc explanation methods.

Globally, our feature importance measure is defined as the average marginal sensitivity of the decision to each feature over the training set. Figure 15 visualizes the relative feature importance of this SRT using the proposed measure in Appendix D.2 and a perturbation-based measure (Hastie et al., 2009). The difference between the two measures is that our measure shows feature importance intrinsically, based solely on the derivatives of the 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, our intrinsic interpretation measure, the empirical integrated gradient (EIG), decomposes the decision into the sum of feature contributions defined by integrated gradients (Sundararajan et al., 2017). Figure 16 visualizes the feature attributions derived from the proposed EIG measure and a traditional post hoc explanation (SHAP measure, Lundberg et al., 2020) for this SRT given a

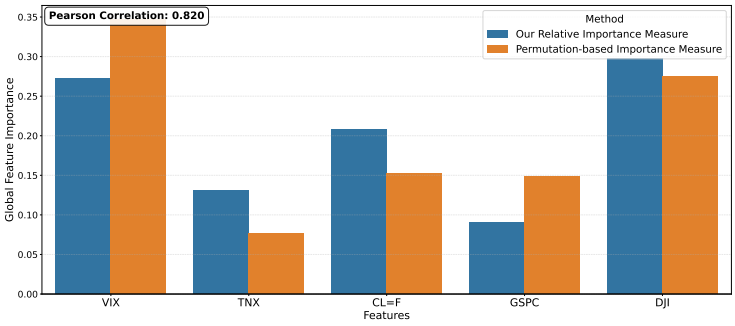
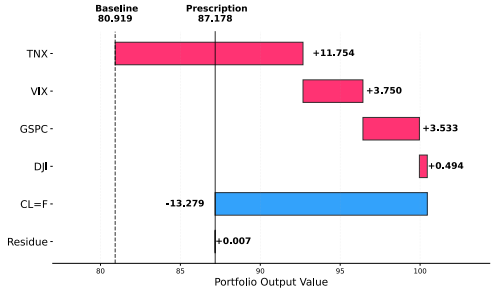
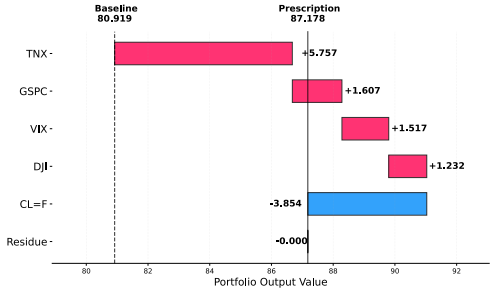


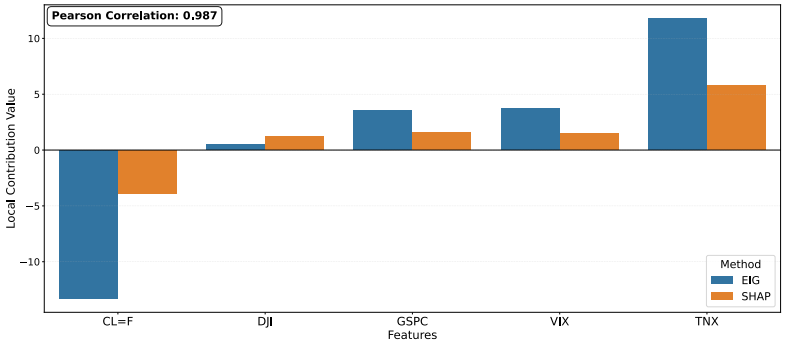
Figure 15: Global feature importance comparison for the trained SRT



(a) Intrinsic interpretation by EIG measure



(b) Post-hoc explanation by SHAP measure



(c) Comparison of the two measures

Figure 16: Local feature attribution comparison for the trained SRT

specific covariate. Figures 16(a) and 16(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 16(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.

G Omitted Analyses and Proof

G.1 Analysis for Condition 1 in Section 3.1

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

Proposition 20. *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 $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}, \boldsymbol{y})$, it holds that*

$$\Psi(f(\boldsymbol{x}), \boldsymbol{y}) \leq \alpha \cdot \lambda c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) + M(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), \tag{31}$$

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

Proof of Proposition 20. In Eq. (4), we set $\boldsymbol{\xi}_1 = \boldsymbol{x} - \widehat{\boldsymbol{x}}$ and $\boldsymbol{\xi}_2 = \boldsymbol{y} - \widehat{\boldsymbol{y}}$. When relation (31) holds, for $\widehat{\mathbb{P}} \otimes \nu_{\mathcal{X}}$-almost every $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x})$, we have

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

where the first equality follows from the change of variables $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{\xi}_2$ and the translation invariance of the Lebesgue measure, the first inequality is by (31), and the finiteness is by Assumptions 1(III) and (IV), and the fact that $e^{M(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})/\lambda\epsilon} < \infty$.

Similarly, Eq. (31) and Jensen's inequality imply that

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

for $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}$-almost every $\widehat{\boldsymbol{x}}$, where the finiteness follows from Assumptions 1(III)–(IV). □

G.2 Proof of Theorem 3 in Section 3.1

To prove the strong duality in Theorem 3, we first establish Lemma 21 and the weak duality result in Lemma 22.

Lemma 21. (Several Measurable Functions). *Assume that Assumption 1 holds, and 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*

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

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*

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

This function is jointly measurable with respect to $(\widehat{\boldsymbol{x}}, \boldsymbol{x})$ for any $\lambda \geq 0$.

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

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

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 21. 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
$$w_1(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}, 0) = \operatorname*{esssup}_{\boldsymbol{y} \sim \nu_{\mathcal{Y}}} \Psi(f(\boldsymbol{x}), \boldsymbol{y}).$$
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 & 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 (Proposition 5.2.1 Cohn, 2013), 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, that is, 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 $\lambda > 0$, Condition 1 ensures that the log-partition function defining $w_1(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}, \lambda)$ is positive and finite. Hence, $w_1(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}, \lambda)$ is a finite-valued measurable function.
Since $\widehat{\mathbb{P}}$ is an empirical distribution, $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}$ has finite support for every $\widehat{\boldsymbol{x}} \in \operatorname{supp}(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}})$. Therefore,
$$g(\widehat{\boldsymbol{x}}, \boldsymbol{x}, \lambda) = \mathbb{E}_{\widehat{\boldsymbol{Y}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}} \left[w_1(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{Y}}, \boldsymbol{x}, \lambda) \right],$$
is a finite weighted sum of measurable functions and is thus 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, the expectations of both w_1^+ and w_1^- with respect to $\widehat{\boldsymbol{y}}$ are well-defined measurable functions of $(\widehat{\boldsymbol{x}}, \boldsymbol{x})$. Since $\mathbb{E}_{\widehat{\boldsymbol{y}}\sim\widehat{\mathbb{P}}_{\widehat{Y}|\widehat{X}=\widehat{\boldsymbol{x}}}}\Big[w_1^-(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x}, 0)\Big]$ 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 22. (Weak Duality). *Under Assumption 1, it holds that* $v_{\mathrm{P}} \leq v_{\mathrm{D}}$.

Proof of Lemma 22. For the primal problem of (Causal-SDRO), let $\lambda \geq 0$ be the Lagrangian multiplier of the causal Sinkhorn ball constraint, and define $v_{\mathrm{P}}(\lambda)$ as the optimal value of the soft-constrained DRO problem with penalty parameter λ:

$$v_{\mathrm{P}}(\lambda) := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \Big[\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}[\Psi(f(\boldsymbol{x}), \boldsymbol{y})] - \lambda R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p\Big]. \tag{32}$$

Based on the Lagrangian weak duality, it holds that

$$\begin{aligned} v_{\mathrm{P}} &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \Big\{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}), \boldsymbol{y})\Big] : R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p \leq \rho^p\Big\} \\ &= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \inf_{\lambda\geq 0} \Big\{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}), \boldsymbol{y})\Big] - \lambda\Big(R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p - \rho^p\Big)\Big\} \\ &\leq \inf_{\lambda\geq 0} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \Big\{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}}\Big[\Psi(f(\boldsymbol{x}), \boldsymbol{y})\Big] - \lambda\Big(R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p - \rho^p\Big)\Big\} \\ &= \inf_{\lambda\geq 0} \Big\{\lambda\rho^p + v_{\mathrm{P}}(\lambda)\Big\}, \end{aligned} \tag{33}$$

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

Next, we reformulate the term $v_{\mathrm{P}}(\lambda)$. For the formulation of CSD

$$R_\epsilon(\widehat{\mathbb{P}}, \mathbb{P})^p = \inf_{\gamma\in\Gamma_c(\widehat{\mathbb{P}},\mathbb{P})} \mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma}\Big[c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) + \epsilon \cdot \mathbb{D}_{\mathrm{KL}}\Big(\gamma\|\widehat{\mathbb{P}} \otimes (\nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}})\Big)\Big], \tag{34}$$

where the second item can be reformulated as

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}\Big(\gamma\|\widehat{\mathbb{P}} \otimes (\nu_{\mathcal{X}} \otimes \nu_{\mathcal{Y}})\Big) \\ &\quad= \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) \\ &\quad= \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) \\ &\quad= \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) \\ &\quad= \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) \\ &\quad= \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{35}$$

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 (35)

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

$$\mathbb{E}_{\gamma}[\cdot] = \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\gamma_{\boldsymbol{Y}|\widehat{\boldsymbol{X}},\widehat{\boldsymbol{Y}},\boldsymbol{X}}}[\cdot \mid \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \boldsymbol{X}] \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\right] \mid \widehat{\boldsymbol{X}}\right]\right]. \tag{36}$$

From relations (35) and (36), the expectation term on the RHS of the Eq. (34) 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\mathbb{D}_{\mathrm{KL}}\left(\gamma\|\widehat{\mathbb{P}}\otimes(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}})\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}))\right.\right.\right.\right. \\
&\qquad\qquad\left.\left.\left.\left. + \epsilon\mathbb{D}_{\mathrm{KL}}\left(\gamma\|\widehat{\mathbb{P}}\otimes(\nu_{\mathcal{X}}\otimes\nu_{\mathcal{Y}})\right) \mid \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \boldsymbol{X}\right] \mid \widehat{\boldsymbol{X}}, \boldsymbol{X}\right] \mid \widehat{\boldsymbol{X}}\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}))\right.\right.\right.\right. \\
&+ \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}\Big] + \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}}\Big]\Big],
\end{aligned}$$

Therefore, we have

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

Similar to relation (36), 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). Since each conditional kernel induced by a feasible causal transport plan is feasible for the corresponding pointwise supremum, the conditional objective is bounded above by its pointwise supremum. Thus, taking expectations and then the supremum over all feasible transport plans yields

$$\begin{aligned}
v_{\mathrm{P}}(\lambda) &\le \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\sup_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\mathbb{E}_{\gamma_{\boldsymbol{X}|\widehat{\boldsymbol{X}}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\left[\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}}}\left[\Psi(f(\boldsymbol{x}),\boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\right.\right.\right.\right. \\
&\qquad\left.\left.\left.\left. - \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}\right] \mid \widehat{\boldsymbol{X}}\right] - \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}}\right]\right] \\
&= \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}}}}\left[\underbrace{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}}}\left[w_1(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\lambda) \mid \widehat{\boldsymbol{X}}\right]}_{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}}\right]}_{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 21. By the Fenchel duality, we have

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

and it follows that

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

Hence, according to Eq. (33), we have

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

This completes the proof of the weak duality. □

Next, we complete the proof of the strong duality theorem.

Proof of Theorem 3. To prove Theorem 3(I), we first rewrite the constraint of the primal problem (Causal-SDRO) as

$$\mathbb{E}_{((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) \sim \gamma} \Big[c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) + \epsilon \log \left(\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \right) \Big] \leq \rho^p. \tag{37}$$

Based on Assumption 1(II), the relation (37) can be reformulated as

$$\mathbb{E}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \sim \widehat{\mathbb{P}}} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \gamma_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})}} \Big[c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})) + \epsilon \log \left(\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \right) \Big] \leq \rho^p. \tag{38}$$

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 Eq. (4):

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

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

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

where $\mathbb{D}_{\mathrm{KL}} \Big(\gamma_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})} || \mathcal{K}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), \epsilon} \Big)$ 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 (38) as the following equivalent constraint in terms of the KL-divergence

$$\epsilon \cdot \mathbb{E}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \sim \widehat{\mathbb{P}}} \Big[\mathbb{D}_{\mathrm{KL}} \Big(\gamma_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})} || \mathcal{K}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), \epsilon} \Big) \Big] \leq \rho^p + \epsilon \cdot \mathbb{E}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \sim \widehat{\mathbb{P}}} \Big[\log Z(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \Big] = \bar{\rho}, \tag{39}$$

which implies that the constraint of the problem (Causal-SDRO) is equivalent to constraint (39),

It is straightforward to prove Theorems 3(I) using the reformulated constraint (39) and the non-negativity of the KL-divergence.

For Theorem 3(II), we first consider there exists $\lambda > 0$ satisfying Condition 1. According to Lemma 22, 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^*, we next show the optimal distribution of v_{P}, denoted by $\mathbb{P}^*$, exists. Suppose $\bar{\rho} \geq 0$ is bounded above such that the CSD constraint is binding, i.e., $R_\epsilon \left(\widehat{\mathbb{P}}, \mathbb{P}^* \right) = \rho$ and $\epsilon \cdot \mathbb{E}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \sim \widehat{\mathbb{P}}} \Big[\mathbb{D}_{\mathrm{KL}} \Big(\gamma_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})} || \mathcal{K}_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), \epsilon} \Big) \Big] = \bar{\rho}$. Then, by the theorem of complementary slackness, there exists $\lambda^* > 0$.

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

$$\rho^p + \epsilon \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \Big[\log \int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x}) \Big] = \frac{1}{\lambda^*} \mathbb{E}_{\widehat{\boldsymbol{x}} \sim \widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}} \Big[\frac{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} t(\widehat{\boldsymbol{x}}, \boldsymbol{x}) \cdot \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}}, \boldsymbol{x})} \mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})} \Big], \tag{40}$$

where

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

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

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

where the inequality relation holds because γ_* is a feasible solution in $\Gamma_c(\widehat{\mathbb{P}}, \mathbb{P}_*)$, and the fourth and fifth equalities are obtained by substituting the expression of γ_*, and the last equality is due to the first-order optimality condition (40). 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

$$
\begin{aligned}
v_{\mathrm{P}} &\geq \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{P}_*}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\Big] \\
&= \mathbb{E}_{((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))\sim\gamma_*}\Big[\Psi(f(\boldsymbol{x}),\boldsymbol{y})\Big] \\
&= \int_{(\mathcal{X}\times\mathcal{Y})\times(\mathcal{X}\times\mathcal{Y})} \Psi(f(\boldsymbol{x}),\boldsymbol{y})\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y})) \\
&\qquad\qquad \cdot \frac{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})}{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))} \cdot \frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} \\
&= \mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[\frac{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}t(\widehat{\boldsymbol{x}},\boldsymbol{x})\cdot\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}{\int_{\mathcal{X}} e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})}\Big] \\
&= \lambda^*\rho^p + \lambda^*\epsilon\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\left[\log\int_{\mathcal{X}}\exp\left(\frac{g(\widehat{\boldsymbol{x}},\boldsymbol{x},\lambda^*)}{\lambda^*\epsilon}\right)\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\right] \\
&= v_{\mathrm{D}}.
\end{aligned}
\tag{41}
$$

According to Lemma 22 and Eq. (41), 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 3(II) is completed.

If Condition 1 fails for every $\lambda > 0$, then the corresponding exponential moment is infinite, and thus $v_{\mathrm{D}} = \infty$. Moreover, when $\bar{\rho} > 0$, the equivalent conditional KL-divergence ambiguity set in Remark 8 has a positive radius, and the lack of exponential integrability implies that the worst-case expected loss is unbounded, i.e., $v_{\mathrm{P}} = \infty$. Therefore, we have $v_{\mathrm{P}} = v_{\mathrm{D}} = \infty$. □

G.3 Proof of Theorem 10 in Appendix B.4

Proof of Theorem 10. In the proof of Theorem 3, we have derived a worst-case distribution of v_{P}, that is,

$$
\frac{\mathrm{d}\gamma((\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}),(\boldsymbol{x},\boldsymbol{y}))}{\mathrm{d}\widehat{\mathbb{P}}(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\nu_{\mathcal{Y}}(\boldsymbol{y})} = \alpha_{\widehat{\boldsymbol{x}}}\cdot\beta_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x}}\cdot e^{r(\widehat{\boldsymbol{x}},\boldsymbol{x})+s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})},
\tag{42}
$$

where

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

and function $r(\widehat{\boldsymbol{x}},\boldsymbol{x})$ and $s(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}},\boldsymbol{x},\boldsymbol{y})$ are defined in (41). 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 in Eq. (14), and then we have

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

where

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

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

Here, we have

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

where

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

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

$$\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\widehat{\mathbb{P}}_{\widehat{\boldsymbol{X}}}}\Big[\int_{\mathcal{X}} u(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\cdot\int_{\mathcal{X}} e^{r(\lambda,\widehat{\boldsymbol{x}},\boldsymbol{x})}\mathrm{d}\nu_{\mathcal{X}}(\boldsymbol{x})\Big]\geq 0, \tag{43}$$

and it follows that

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

where the first inequality follows from relation (43), and the second follows from the Cauchy-Schwarz inequality.

Therefore, for any $\lambda>0$, we have $\nabla_\lambda^2 v(\lambda)\geq 0$, and equality holds if and only if the function $\Psi(f(\cdot),\cdot)$ is a constant function. Thus, the strict convexity of the dual problem (14a) holds for the dual objective, and it implies the uniqueness of λ^*. □

G.4 Proof of Theorem 6 in Section 5.2

We first prove the smoothness of the function $F(\boldsymbol{\theta})$ in (SCO).

Lemma 23 (Smoothness of the Objective)**.** *Let $\mathcal{U}$ be an open convex neighborhood of Θ. Under Assumption 2, function $F(\boldsymbol{\theta})$ in (SCO) is S_F-Lipschitz smooth on $\mathcal{U}$, where*

$$S_F := \lambda\epsilon\left(S_1C_2^2C_3^2 + C_1S_2C_3^2 + C_1C_2S_3\right).$$

Proof of Lemma 23. The uniform derivative bounds justify differentiation under the expectations, i.e., $J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi}_2\sim W_\epsilon}[J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta})]$. By the chain rule, we have

$$\nabla_{\boldsymbol{\theta}}\phi^{(1)}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta})^\top \nabla t_2(\phi^{(2)}(\boldsymbol{\theta}))\right],$$
$$\nabla F(\boldsymbol{\theta}) = \lambda\epsilon\, \mathbb{E}_{\widehat{\boldsymbol{x}}}\left[\nabla t_1(\phi^{(1)}\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}\phi^{(1)}(\boldsymbol{\theta})\right].$$

According to Proposition 18, under Assumption 2, we have

$$\|J_{\boldsymbol{\theta}}\phi^{(2)}\|_{\mathrm{op}} \le C_3, \qquad \|\nabla_{\boldsymbol{\theta}}\phi^{(1)}\|_2 \le C_2C_3.$$

Since $\mathcal{U}$ is convex, $\phi^{(2)}$ and $\phi^{(1)}$ are respectively C_3- and C_2C_3-Lipschitz continuous. We also have $\|J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}) - J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}')\|_{\mathrm{op}} \le S_3\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2$, and

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\theta}}\phi^{(1)}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\phi^{(1)}(\boldsymbol{\theta}')\right\|_2 \\
&= \Big\|\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\Big[\left(J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}) - J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}')\right)^\top \nabla t_2\left(\phi^{(2)}(\boldsymbol{\theta})\right) \\
&\qquad + J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}')^\top\left(\nabla t_2\left(\phi^{(2)}(\boldsymbol{\theta})\right) - \nabla t_2\left(\phi^{(2)}(\boldsymbol{\theta}')\right)\right)\Big]\Big\|_2 \\
&\le C_2\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left\|J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}) - J_{\boldsymbol{\theta}}\phi^{(2)}(\boldsymbol{\theta}')\right\|_{\mathrm{op}} + C_3S_2\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left\|\phi^{(2)}(\boldsymbol{\theta}) - \phi^{(2)}(\boldsymbol{\theta}')\right\|_2 \\
&\le \left(C_2S_3 + C_3S_2C_3\right)\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2.
\end{aligned}$$

Then, function $F(\boldsymbol{\theta})$ is S_F-Lipschitz smooth on $\mathcal{U}$:

$$\begin{aligned}
\|\nabla F(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta}')\|_2 &\le \lambda\epsilon\,\mathbb{E}_{\widehat{\boldsymbol{x}}}\Big[\left\|\nabla t_1\left(\phi^{(1)}(\boldsymbol{\theta})\right) - \nabla t_1\left(\phi^{(1)}(\boldsymbol{\theta}')\right)\right\|\left\|\nabla_{\boldsymbol{\theta}}\phi^{(1)}(\boldsymbol{\theta})\right\|_2 \\
&\qquad + \left\|\nabla t_1\left(\phi^{(1)}(\boldsymbol{\theta}')\right)\right\|\left\|\nabla_{\boldsymbol{\theta}}\phi^{(1)}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\phi^{(1)}(\boldsymbol{\theta}')\right\|_2\Big] \\
&\le \lambda\epsilon\left[S_1(C_2C_3)^2 + C_1(C_2S_3 + S_2C_3^2)\right]\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2 = S_F\cdot\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2.
\end{aligned}$$

where $S_F = \lambda\epsilon\left[S_1(C_2C_3)^2 + C_1(C_2S_3 + S_2C_3^2)\right]$. □

Proof of Theorem 6. For proving Theorem 6(I), we define the full version of Eq. (27):

$$\begin{aligned}
\phi^{(0)}(\boldsymbol{\theta};\widehat{\boldsymbol{x}}) &:= \mathbb{E}_{\widehat{\boldsymbol{x}}}\Big[t_1(\phi^{(1)}(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}})\Big], \\
\phi^{(1)}(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) &:= \mathbb{E}_{\boldsymbol{\xi}_1}\Big[t_2(\phi^{(2)}(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2);\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)\Big], \\
\phi^{(2)}(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) &:= \mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\Big].
\end{aligned} \tag{44}$$

All parameter iterates remain in the open convex region $\mathcal{U}$ of Lemma 23. Thus, the objective is S_F-smooth with respect to every parameter update. We use constant stepsizes α and β during the K iterations, with $0 < \beta \le 1/2$. At iteration k, the order in which the samples we draw are used for calculation is $\boldsymbol{\xi}_2^k$, $\boldsymbol{\xi}_1^k$, and $\widehat{\boldsymbol{x}}^k$. The three draws are mutually independent and independent of all previous draws. Define

$$\mathcal{F}_k := \sigma\left(\boldsymbol{\xi}_2^j, \boldsymbol{\xi}_1^j, \widehat{\boldsymbol{x}}^j : 0 \le j < k\right), \qquad \mathcal{F}_{k,2} := \mathcal{F}_k \vee \sigma(\boldsymbol{\xi}_2^k), \qquad \mathcal{F}_k^+ := \mathcal{F}_{k,2} \vee \sigma(\boldsymbol{\xi}_1^k).$$

Here $\sigma(\cdot)$ denotes the generated σ-algebra, and $\vee$ joins two σ-algebras. Thus, $\mathcal{F}_k$ contains the history before iteration k. The information $\mathcal{F}_{k,2}$ further contains $\boldsymbol{\xi}_2^k$, and the information $\mathcal{F}_k^+$ contains both $\boldsymbol{\xi}_2^k$ and $\boldsymbol{\xi}_1^k$ but not $\widehat{\boldsymbol{x}}^k$. In Algorithm 1, for every condition $(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1)$, we compute

$$\widetilde{\boldsymbol{y}}_2^{k+1}(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) = (1-\beta)\boldsymbol{y}_2^k(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) + t_3(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2^k) - (1-\beta)t_3(\boldsymbol{\theta}^{k-1};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2^k), \tag{45}$$

and apply the clipping rule to obtain the corresponding $\boldsymbol{y}_2^{k+1}$ with domain $[1,\exp(B/\lambda\epsilon)]^N$. Next, for every $\widehat{\boldsymbol{x}}$, compute

$$\widetilde{y}_1^{k+1}(\widehat{\boldsymbol{x}}) = (1-\beta)y_1^k(\widehat{\boldsymbol{x}}) + t_2\left(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1^k);\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1^k\right) - (1-\beta)t_2\left(\boldsymbol{y}_2^k(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1^k);\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1^k\right). \tag{46}$$

We can apply the same clipping rule to obtain y_1^{k+1}. Hence $\boldsymbol{y}_2^{k+1}$ is $\mathcal{F}_{k,2}$-measurable, and y_1^{k+1} is $\mathcal{F}_k^+$-measurable. We define the stochastic compositional gradient as

$$\boldsymbol{\nabla}^k = \lambda\epsilon\,\nabla t_1\left(y_1^{k+1}(\widehat{\boldsymbol{x}}^k);\widehat{\boldsymbol{x}}^k\right) J_{\boldsymbol{\theta}} t_3\left(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k,\boldsymbol{\xi}_2^k\right)^\top \nabla t_2\left(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k\right), \tag{47}$$

and then the parameters are updated by

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \alpha\boldsymbol{\nabla}^k. \tag{48}$$

Let $C_d := \lambda\epsilon C_1C_2C_3$. By Proposition 18 and Eq. (48), we obtain

$$\|\boldsymbol{\nabla}^k\|_2 \le C_d, \qquad \mathbb{E}\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}\|_2^2 \le C_d^2\alpha^2. \tag{49}$$

The second inequality also holds at $k=0$ by initialization.

We prove Theorem 6(I) in three steps, following Chen et al. (2021).

Step 1. Tracking variance. Define the variances of $y_2^k(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)$ given $\widehat{\boldsymbol{x}}$ and $\boldsymbol{\xi}_1$ and $y_1^k(\widehat{\boldsymbol{x}})$ given $\widehat{\boldsymbol{x}}$.

$$\begin{aligned} e_{2,k} &:= \mathbb{E}\left\|\boldsymbol{y}_2^k(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) - \phi^{(2)}(\boldsymbol{\theta}^{k-1};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)\right\|_2^2, \\ e_{1,k} &:= \mathbb{E}\left|y_1^k(\widehat{\boldsymbol{x}}) - \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\boldsymbol{y}_2^k(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1\right)\right]\right|^2, \\ E_k &:= e_{1,k} + e_{2,k}. \end{aligned} \tag{50}$$

In this step, we derive recursive bounds for the tracking errors $e_{1,k}$, $e_{2,k}$, and E_k.

For fixed $(\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1)$, define estimation error for the expectation of t_3:

$$\delta_3^k(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1) := t_3(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2^k) - \phi^{(2)}(\boldsymbol{\theta};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1).$$

Then, suppressing the fixed conditions for brevity, we have

$$\begin{aligned} \widetilde{\boldsymbol{y}}_2^{k+1} - \phi^{(2)}(\boldsymbol{\theta}^k) &= (1-\beta)\left[\boldsymbol{y}_2^k - \phi^{(2)}(\boldsymbol{\theta}^{k-1})\right] + \delta_3^k(\boldsymbol{\theta}^k) - (1-\beta)\delta_3^k(\boldsymbol{\theta}^{k-1}) \\ &= (1-\beta)\left[\boldsymbol{y}_2^k - \phi^{(2)}(\boldsymbol{\theta}^{k-1})\right] + (1-\beta)\left[\delta_3^k(\boldsymbol{\theta}^k) - \delta_3^k(\boldsymbol{\theta}^{k-1})\right] + \beta\delta_3^k(\boldsymbol{\theta}^k). \end{aligned} \tag{51}$$

Since $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}^{k-1}$ are $\mathcal{F}_k$-measurable and $\boldsymbol{\xi}_2^k$ is independent of $\mathcal{F}_k$, by the definition of $\phi^{(2)}$ we have

$$\mathbb{E}\left[\delta_3^k(\boldsymbol{\theta}^k) \mid \mathcal{F}_k\right] = \mathbb{E}\left[\delta_3^k(\boldsymbol{\theta}^{k-1}) \mid \mathcal{F}_k\right] = \mathbf{0}.$$

Hence, the last two terms in Eq. (51) have conditional mean zero, and their cross term with the previous tracking error vanishes. Therefore,

$$\begin{aligned} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{y}}_2^{k+1} - \phi^{(2)}(\boldsymbol{\theta}^k)\right\|_2^2 \mid \mathcal{F}_k\right] &= (1-\beta)^2\left\|\boldsymbol{y}_2^k - \phi^{(2)}(\boldsymbol{\theta}^{k-1})\right\|_2^2 \\ &\quad + \mathbb{E}\left[\left\|(1-\beta)\left[\delta_3^k(\boldsymbol{\theta}^k) - \delta_3^k(\boldsymbol{\theta}^{k-1})\right] + \beta\delta_3^k(\boldsymbol{\theta}^k)\right\|_2^2 \mid \mathcal{F}_k\right]. \end{aligned} \tag{52}$$

To bound the centered difference, define

$$\boldsymbol{D}_k := t_3(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2^k) - t_3(\boldsymbol{\theta}^{k-1};\widehat{\boldsymbol{x}},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2^k).$$

Then, by the definition of $\phi^{(2)}$, we have

$$\delta_3^k(\boldsymbol{\theta}^k) - \delta_3^k(\boldsymbol{\theta}^{k-1}) = \boldsymbol{D}_k - \mathbb{E}[\boldsymbol{D}_k \mid \mathcal{F}_k].$$

Thus, using the conditional variance identity and the L_3-Lipschitz continuity of t_3,

$$\begin{aligned} \mathbb{E}\left[\left\|\delta_3^k(\boldsymbol{\theta}^k) - \delta_3^k(\boldsymbol{\theta}^{k-1})\right\|_2^2 \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\|\boldsymbol{D}_k - \mathbb{E}[\boldsymbol{D}_k \mid \mathcal{F}_k]\|_2^2 \mid \mathcal{F}_k\right] \\ &\le \mathbb{E}\left[\|\boldsymbol{D}_k\|_2^2 \mid \mathcal{F}_k\right] \\ &\le L_3^2\left\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}\right\|_2^2. \end{aligned} \tag{53}$$

Moreover, each component of both $t_3(\boldsymbol{\theta}^k)$ and $\phi^{(2)}(\boldsymbol{\theta}^k)$ lies in $[1,\exp(B/(\lambda\epsilon))]$. Hence,

$$\left\|\delta_3^k(\boldsymbol{\theta}^k)\right\|_2^2 = \sum_{i=1}^N\left|[t_3(\boldsymbol{\theta}^k)]_i - [\phi^{(2)}(\boldsymbol{\theta}^k)]_i\right|^2 \le N\left(e^{B/(\lambda\epsilon)} - 1\right)^2. \tag{54}$$

Applying $\|\boldsymbol{u}+\boldsymbol{v}\|_2^2 \le 2\|\boldsymbol{u}\|_2^2 + 2\|\boldsymbol{v}\|_2^2$ to the second term in Eq. (52), and using Eqs. (53) and (54), yields

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{y}}_2^{k+1} - \phi^{(2)}(\boldsymbol{\theta}^k)\right\|_2^2 \mid \mathcal{F}_k\right] \le (1-\beta)^2\left\|\boldsymbol{y}_2^k - \phi^{(2)}(\boldsymbol{\theta}^{k-1})\right\|_2^2 + 2(1-\beta)^2 L_3^2\left\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}\right\|_2^2 + 2N\left(e^{B/(\lambda\epsilon)}-1\right)^2\beta^2.$$

Finally, since $\phi^{(2)}(\boldsymbol{\theta}^k) \in [1, \exp(B/(\lambda\epsilon))]^N$ and $\boldsymbol{y}_2^{k+1}$ is obtained by clipping $\widetilde{\boldsymbol{y}}_2^{k+1}$ onto this range, we have

$$\left\|\boldsymbol{y}_2^{k+1} - \phi^{(2)}(\boldsymbol{\theta}^k)\right\|_2^2 \le \left\|\widetilde{\boldsymbol{y}}_2^{k+1} - \phi^{(2)}(\boldsymbol{\theta}^k)\right\|_2^2.$$

Taking expectations over the algorithmic randomness, and using Eq. (49), we obtain

$$\begin{aligned} e_{2,k+1} &\le (1-\beta)^2 e_{2,k} + 2(1-\beta)^2 L_3^2\mathbb{E}\left[\left\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}\right\|_2^2\right] + 2N\left(e^{B/(\lambda\epsilon)}-1\right)^2\beta^2 \\ &\le (1-\beta)^2 e_{2,k} + 2(1-\beta)^2 L_3^2 C_d^2\alpha^2 + 2N\left(e^{B/(\lambda\epsilon)}-1\right)^2\beta^2. \end{aligned} \tag{55}$$

To analyze the update of $y_1^k(\widehat{\boldsymbol{x}})$, we condition on $\mathcal{F}_{k,2}$. Here both $\boldsymbol{y}_2^k$ and $\boldsymbol{y}_2^{k+1}$ are fixed. Then, for $j \in \{k, k+1\}$ and fixed $\widehat{\boldsymbol{x}}$, we have

$$\mathbb{E}\left[t_2\left(\boldsymbol{y}_2^j(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1^k); \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1^k\right) \mid \mathcal{F}_{k,2}\right] = \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\boldsymbol{y}_2^j(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1); \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1\right)\right].$$

Similar to the analysis for $e_{2,k}$, using the L_2-Lipschitz bound, we obtain

$$\begin{aligned} e_{1,k+1} \le (1-\beta)^2 e_{1,k} + 2\left(e^{B/(\lambda\epsilon)}-1\right)^2\beta^2 \\ + 2L_2^2\mathbb{E}\left\|\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1) - \boldsymbol{y}_2^k(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1)\right\|_2^2. \end{aligned} \tag{56}$$

For the third term in Eq. (56), by Eq. (45), we first have

$$\widetilde{\boldsymbol{y}}_2^{k+1} - \boldsymbol{y}_2^k = (1-\beta)\left[t_3(\boldsymbol{\theta}^k) - t_3(\boldsymbol{\theta}^{k-1})\right] + \beta\left[t_3(\boldsymbol{\theta}^k) - \boldsymbol{y}_2^k\right],$$

and it follows that

$$\mathbb{E}\left\|\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1) - \boldsymbol{y}_2^k(\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1)\right\|_2^2 \le 2(1-\beta)^2 L_3^2 C_d^2\alpha^2 + 2N\left(e^{B/(\lambda\epsilon)}-1\right)^2\beta^2, \tag{57}$$

where the first term is due to the L_3-Lipschitz continuity of t_3 and Eq. (49), and the second term is due to the range of t_3.

Combining Eqs. (55)–(57), we obtain

$$E_{k+1} \le (1-\beta)^2 E_k + c_\alpha\alpha^2 + c_\beta\beta^2, \tag{58}$$

where

$$c_\alpha := 2(1+2L_2^2)(1-\beta)^2 L_3^2 C_d^2, \qquad c_\beta := 2\left[(1+2L_2^2)N + 1\right]\left(e^{B/(\lambda\epsilon)}-1\right)^2.$$

Step 2. Gradient bias. In this step, we bound the difference between the actual direction $\boldsymbol{\nabla}^k$ and the ideal direction $\bar{\boldsymbol{\nabla}}^k$, and then bound the corresponding conditional bias. We first analyze the tracking errors evaluated at the reused samples. Given $\mathcal{F}_{k,2}$, the function $\boldsymbol{y}_2^{k+1}$ and the parameter $\boldsymbol{\theta}^k$ are fixed, while $\boldsymbol{\xi}_1^k$ and $\widehat{\boldsymbol{x}}^k$ are independent of $\mathcal{F}_{k,2}$. Therefore, according to the definition of $e_{2,k+1}$, we have

$$\mathbb{E}\left\|\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k, \boldsymbol{\xi}_1^k) - \phi^{(2)}(\boldsymbol{\theta}^k; \widehat{\boldsymbol{x}}^k, \boldsymbol{\xi}_1^k)\right\|_2^2 = e_{2,k+1}. \tag{59}$$

The tracking variable y_1^{k+1} is $\mathcal{F}_k^+$-measurable and $\widehat{\boldsymbol{x}}^k$ is independent of $\mathcal{F}_k^+$. Hence,

$$\begin{aligned} &\left(\mathbb{E}\left|y_1^{k+1}(\widehat{\boldsymbol{x}}^k) - \phi^{(1)}(\boldsymbol{\theta}^k; \widehat{\boldsymbol{x}}^k)\right|^2\right)^{1/2} \\ &\le \left(\mathbb{E}\left|y_1^{k+1}(\widehat{\boldsymbol{x}}^k) - \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k, \boldsymbol{\xi}_1); \widehat{\boldsymbol{x}}^k, \boldsymbol{\xi}_1\right)\right]\right|^2\right)^{1/2} \\ &\quad + \left(\mathbb{E}\left|\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\left(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k, \boldsymbol{\xi}_1); \widehat{\boldsymbol{x}}^k, \boldsymbol{\xi}_1\right)\right] - \phi^{(1)}(\boldsymbol{\theta}^k; \widehat{\boldsymbol{x}}^k)\right|^2\right)^{1/2} \end{aligned} \tag{60}$$

By the definition of $e_{1,k+1}$,

$$\left(\mathbb{E}\left|y_1^{k+1}(\widehat{\boldsymbol{x}}^k) - \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\big(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1\big)\right]\right|^2\right)^{1/2} = \sqrt{e_{1,k+1}}. \tag{61}$$

Moreover,

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\big(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1\big)\right] - \phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k)\right| \\
&= \left|\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\big(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1\big) - t_2\Big(\phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1\Big)\right]\right| \\
&\le \mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\left|t_2\big(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1\big) - t_2\Big(\phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1\Big)\right|\right] \\
&\le L_2\,\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[\left\|\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1) - \phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1)\right\|_2\right].
\end{aligned} \tag{62}$$

Thus, by Jensen's inequality,

$$\begin{aligned}
&\left(\mathbb{E}\left|\mathbb{E}_{\boldsymbol{\xi}_1\sim Q_\epsilon}\left[t_2\big(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1\big)\right] - \phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k)\right|^2\right)^{1/2} \\
&\le L_2\left(\mathbb{E}\left\|\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k) - \phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k)\right\|_2^2\right)^{1/2} \\
&= L_2\sqrt{e_{2,k+1}}.
\end{aligned} \tag{63}$$

Combining Eqs. (60)–(63), we have

$$\left(\mathbb{E}\left|y_1^{k+1}(\widehat{\boldsymbol{x}}^k) - \phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k)\right|^2\right)^{1/2} \le \sqrt{e_{1,k+1}} + L_2\sqrt{e_{2,k+1}}. \tag{64}$$

We next define an ideal direction by replacing the auxiliary variables in $\boldsymbol{\nabla}^k$ with their true expectations while keeping the same samples:

$$\bar{\boldsymbol{\nabla}}^k := \lambda\epsilon\,\nabla t_1\left(\phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k);\widehat{\boldsymbol{x}}^k\right) J_{\boldsymbol{\theta}} t_3\left(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k,\boldsymbol{\xi}_2^k\right)^\top \nabla t_2\left(\phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k);\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k\right). \tag{65}$$

Given $\mathcal{F}_k$, the parameter $\boldsymbol{\theta}^k$ is fixed, while $\boldsymbol{\xi}_2^k$, $\boldsymbol{\xi}_1^k$, and $\widehat{\boldsymbol{x}}^k$ are independent current samples. Taking expectations sequentially with respect to $\boldsymbol{\xi}_2^k$, $\boldsymbol{\xi}_1^k$, and $\widehat{\boldsymbol{x}}^k$, and using the chain rule, we obtain

$$\mathbb{E}[\bar{\boldsymbol{\nabla}}^k \mid \mathcal{F}_k] = \nabla F(\boldsymbol{\theta}^k). \tag{66}$$

We next bound the difference between the actual direction and the ideal direction. By adding and subtracting the derivative product evaluated at the true outer input and the auxiliary middle input, we have

$$\begin{aligned}
\boldsymbol{\nabla}^k - \bar{\boldsymbol{\nabla}}^k = {}& \lambda\epsilon\left[\nabla t_1\left(y_1^{k+1}(\widehat{\boldsymbol{x}}^k);\widehat{\boldsymbol{x}}^k\right) - \nabla t_1\left(\phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k);\widehat{\boldsymbol{x}}^k\right)\right] J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta}^k)^\top \nabla t_2\left(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k)\right) \\
&+ \lambda\epsilon\nabla t_1\left(\phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k);\widehat{\boldsymbol{x}}^k\right) J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta}^k)^\top\left[\nabla t_2\left(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k)\right) - \nabla t_2\left(\phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k)\right)\right].
\end{aligned}$$

By Proposition 18, we obtain

$$\begin{aligned}
\|\boldsymbol{\nabla}^k - \bar{\boldsymbol{\nabla}}^k\|_2 &\le \lambda\epsilon C_3C_2\left\|\nabla t_1\left(y_1^{k+1}(\widehat{\boldsymbol{x}}^k)\right) - \nabla t_1\left(\phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k)\right)\right\|_2 \\
&\quad + \lambda\epsilon C_3C_1\left\|\nabla t_2\left(\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k)\right) - \nabla t_2\left(\phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k)\right)\right\|_2 \\
&\le \lambda\epsilon C_3C_2S_1\left|y_1^{k+1}(\widehat{\boldsymbol{x}}^k) - \phi^{(1)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k)\right| \\
&\quad + \lambda\epsilon C_3C_1S_2\left\|\boldsymbol{y}_2^{k+1}(\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k) - \phi^{(2)}(\boldsymbol{\theta}^k;\widehat{\boldsymbol{x}}^k,\boldsymbol{\xi}_1^k)\right\|_2.
\end{aligned}$$

Using Eqs. (59), (64), and the triangle inequality in mean square, we obtain

$$\left(\mathbb{E}\|\boldsymbol{\nabla}^k - \bar{\boldsymbol{\nabla}}^k\|_2^2\right)^{1/2} \le A_1\sqrt{e_{1,k+1}} + A_2\sqrt{e_{2,k+1}},$$

where

$$A_1 := \lambda\epsilon C_3 C_2 S_1, \qquad A_2 := \lambda\epsilon C_3 (C_2 S_1 L_2 + C_1 S_2).$$

We take positive upper bounds of A_1 and A_2 such that $A_1^2 + A_2^2 > 0$. Define the conditional bias as

$$\boldsymbol{b}^k := \mathbb{E}[\boldsymbol{\nabla}^k \mid \mathcal{F}_k] - \nabla F(\boldsymbol{\theta}^k) = \mathbb{E}[\boldsymbol{\nabla}^k - \bar{\boldsymbol{\nabla}}^k \mid \mathcal{F}_k].$$

By Jensen's inequality and the Cauchy–Schwarz inequality, we have

$$\mathbb{E}\|\boldsymbol{b}^k\|_2^2 \le \mathbb{E}\|\boldsymbol{\nabla}^k - \bar{\boldsymbol{\nabla}}^k\|_2^2 \le \left(A_1\sqrt{e_{1,k+1}} + A_2\sqrt{e_{2,k+1}}\right)^2 \le (A_1^2 + A_2^2) E_{k+1}. \tag{67}$$

We next derive a descent bound for the objective function. By the S_F-Lipschitz smoothness of F and Eq. (49), we have

$$F(\boldsymbol{\theta}^{k+1}) \le F(\boldsymbol{\theta}^k) - \alpha\langle \nabla F(\boldsymbol{\theta}^k), \boldsymbol{\nabla}^k\rangle + \frac{S_F C_d^2}{2}\alpha^2.$$

Taking the conditional expectation given $\mathcal{F}_k$ and using $\mathbb{E}[\boldsymbol{\nabla}^k \mid \mathcal{F}_k] = \nabla F(\boldsymbol{\theta}^k) + \boldsymbol{b}^k$, we obtain

$$\mathbb{E}[F(\boldsymbol{\theta}^{k+1}) \mid \mathcal{F}_k] \le F(\boldsymbol{\theta}^k) - \alpha\|\nabla F(\boldsymbol{\theta}^k)\|_2^2 + \alpha\|\nabla F(\boldsymbol{\theta}^k)\|_2\|\boldsymbol{b}^k\|_2 + \frac{S_F C_d^2}{2}\alpha^2.$$

Using Young's inequality,

$$\alpha\|\nabla F(\boldsymbol{\theta}^k)\|_2\|\boldsymbol{b}^k\|_2 \le \frac{\alpha^2(A_1^2 + A_2^2)}{4\beta}\|\nabla F(\boldsymbol{\theta}^k)\|_2^2 + \frac{\beta}{A_1^2 + A_2^2}\|\boldsymbol{b}^k\|_2^2.$$

Taking the total expectation and applying Eq. (67), we finally obtain

$$\begin{aligned}\mathbb{E}F(\boldsymbol{\theta}^{k+1}) \le \mathbb{E}F(\boldsymbol{\theta}^k) - \alpha\left(1 - \frac{\alpha(A_1^2 + A_2^2)}{4\beta}\right)\mathbb{E}\|\nabla F(\boldsymbol{\theta}^k)\|_2^2 \\ + \beta E_{k+1} + \frac{S_F C_d^2}{2}\alpha^2.\end{aligned} \tag{68}$$

Step 3. Lyapunov descent and convergence rate. In this step, we establish a descent recursion for the Lyapunov function and derive the bound on $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^k)\|_2^2$. Let $F_{\inf} > -\infty$ be a fixed lower bound of F on $\mathcal{U}$. Define

$$\mathcal{V}_k := \mathbb{E}F(\boldsymbol{\theta}^k) - F_{\inf} + E_k \ge 0.$$

Adding E_{k+1} to Eq. (68) and using Eq. (58), we obtain

$$\begin{aligned}\mathcal{V}_{k+1} \le \mathcal{V}_k - \alpha\left(1 - \frac{\alpha(A_1^2 + A_2^2)}{4\beta}\right)\mathbb{E}\|\nabla F(\boldsymbol{\theta}^k)\|_2^2 \\ + \left[(1+\beta)(1-\beta)^2 - 1\right]E_k + (1+\beta)(c_\alpha\alpha^2 + c_\beta\beta^2) + \frac{S_F C_d^2}{2}\alpha^2.\end{aligned} \tag{69}$$

Since $0 < \beta \le 1/2$, we have $(1+\beta)(1-\beta)^2 \le 1$ and $1 + \beta \le 2$. Hence, the term containing E_k is nonpositive. Let

$$B_2 := 2c_\alpha + \frac{S_F C_d^2}{2}, \qquad B_3 := 2c_\beta,$$

and choose $\alpha = 2\beta/(A_1^2 + A_2^2)$. Then, we have

$$\mathcal{V}_{k+1} \le \mathcal{V}_k - \frac{\alpha}{2}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^k)\|_2^2 + B_2\alpha^2 + B_3\beta^2. \tag{70}$$

Summing over $k = 0, \dots, K-1$, the successive Lyapunov terms cancel. Since $\mathcal{V}_K \ge 0$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^k)\|_2^2 \le \frac{2\mathcal{V}_0}{K\alpha} + 2B_2\alpha + 2B_3\frac{\beta^2}{\alpha}.$$

For $K \ge \max\{1, (A_1^2 + A_2^2)^2\}$, we take $\alpha = \frac{1}{\sqrt{K}}$ and $\beta = \frac{A_1^2 + A_2^2}{2\sqrt{K}}$ that ensuring $0 < \beta \le 1/2$. Therefore,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^k)\|_2^2 \le \frac{2\mathcal{V}_0 + 2B_2 + \frac{B_3}{2}(A_1^2 + A_2^2)^2}{\sqrt{K}} = \frac{C_{\text{const}}}{\sqrt{K}},$$

where

$$C_{\text{const}} := 2\mathcal{V}_0 + 2B_2 + \frac{B_3}{2}(A_1^2 + A_2^2)^2.$$

Here, since all constants are independent of K, C_{const} is also independent of K. This completes the proof of Theorem 6(I).

For Theorem 6(II), to ensure that $\mathbb{E}\left[\|\nabla F(\widehat{\boldsymbol{\theta}})\|^2\right] \leq \varepsilon^2$, according to Chen et al. (2021), it follows that

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

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 $3 \cdot K$, implying that the sampling complexity of the SCSC method is also on 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 K gradient calculations, implying that their gradient complexities are of the same order, $\mathcal{O}(\varepsilon^{-4})$.

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

G.5 Analysis for Remark 9 in Appendix B.3

For regularization parameter $\epsilon > 0$, define $\tau := 1/\epsilon > 0$. We also define

$$H_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}}(\boldsymbol{x}, \boldsymbol{y}) := \Psi(f(\boldsymbol{x}), \boldsymbol{y}) - \lambda c_p((\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}), (\boldsymbol{x}, \boldsymbol{y})).$$

By the log-Laplace principle, for any fixed $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}, \boldsymbol{x})$, when $\epsilon \to 0$, we have

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

where the interchange of the limit and the conditional expectation in the third equality is because $\widehat{\mathbb{P}}_{\widehat{\boldsymbol{Y}}|\widehat{\boldsymbol{X}}=\widehat{\boldsymbol{x}}}$ has finite support, and the final equality is due to our lower semicontinuous assumption on $H_{\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}}}(\boldsymbol{x}, \boldsymbol{y})$. Then, for the Eq. (14a), similarly, applying the Laplace principle to the log-partition

function yields

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

When $\operatorname{supp}\nu_{\mathcal{X}}=\mathcal{X}$ and $\operatorname{supp}\nu_{\mathcal{Y}}=\mathcal{Y}$, the dual objective function of the Causal-SDRO converges to that of Causal-WDRO in Yang et al. (2026).

G.6 Proof of Proposition 15 in Appendix D.1

Proof of Proposition 15. 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:

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

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

$$\ln p_{l_t,t}(\boldsymbol{x})=\sum_{i\in\Lambda(l_t)}\ln\Omega_{i,t}(\boldsymbol{x}). \tag{71}$$

Eq. (71) is well-defined since $p_{l_t,t}(\boldsymbol{x})>0$ and $\Omega_{i,t}(\boldsymbol{x})>0$ strictly hold for any $\boldsymbol{x}\in\mathcal{X}$.

Taking the partial derivative of both sides of Eq. (71) with respect to feature x_j:

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

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_t in tree t, we define:

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

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:

$$\frac{\partial p_{l_t,t}(\boldsymbol{x})}{\partial x_j}=p_{l_t,t}(\boldsymbol{x})\cdot\sum_{i\in\Lambda(l_t)}\psi_{i,t}[\boldsymbol{w}_{i,t}]_j. \tag{72}$$

We next compute the second-order derivative $\frac{\partial^2 p_{l_t,t}(\boldsymbol{x})}{\partial x_j\partial x_k}$ by differentiating Eq. (72) with respect to x_k. Applying the product rule yields two terms:

$$\frac{\partial^2 p_{l_t,t}}{\partial x_j\partial x_k}=\underbrace{\frac{\partial p_{l_t,t}}{\partial x_k}\left(\sum_{i\in\Lambda(l_t)}\psi_{i,t}[\boldsymbol{w}_{i,t}]_j\right)}_{\mathbf{A}_1}+\underbrace{p_{l_t,t}\frac{\partial}{\partial x_k}\left(\sum_{i\in\Lambda(l_t)}\psi_{i,t}[\boldsymbol{w}_{i,t}]_j\right)}_{\mathbf{A}_2},$$

where $\mathbf{A}_1$ is equivalent to

$$\mathbf{A}_1 = p_{l_t,t}(\boldsymbol{x}) \left(\sum_{i \in \Lambda(l_t)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_k \right) \left(\sum_{i \in \Lambda(l_t)} \psi_{i,t} [\boldsymbol{w}_{i,t}]_j \right).$$

For $\mathbf{A}_2$, 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:

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

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

$$\mathbf{A}_2 = -p_{l_t,t}(\boldsymbol{x}) \sum_{i \in \Lambda(l_t)} s_{i,t}(1 - s_{i,t})[\boldsymbol{w}_{i,t}]_j [\boldsymbol{w}_{i,t}]_k.$$

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

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

□

G.7 Proof of Proposition 16 in Appendix D.1

Proof of Proposition 16. Since 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 $\|\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})\|$ is bounded by L^{SRF}. Similarly, to establish Lipschitz smoothness, it suffices to show that the Lipschitz constant of $\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})$ is bounded by S^{SRF}.

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

$$\begin{aligned} \|\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})\| &= \left\| \frac{1}{T} \sum_{t=1}^{T} \sum_{l_t=1}^{2^{D(t)}} \nabla_{\boldsymbol{x}} p_{l_t,t}(\boldsymbol{x}) \cdot \boldsymbol{\pi}_{l_t,t}^{\top} \right\| \leq \frac{1}{T} \sum_{t=1}^{T} \sum_{l_t=1}^{2^{D(t)}} \left\| \nabla_{\boldsymbol{x}} p_{l_t,t}(\boldsymbol{x}) \cdot \boldsymbol{\pi}_{l_t,t}^{\top} \right\| \\ &= \frac{1}{T} \sum_{t=1}^{T} \sum_{l_t=1}^{2^{D(t)}} \left\| \nabla_{\boldsymbol{x}} p_{l_t,t}(\boldsymbol{x}) \right\| \cdot \left\| \boldsymbol{\pi}_{l_t,t} \right\|, \end{aligned} \tag{73}$$

where the inequality follows from the triangle inequality, and the second equality follows from the spectral-norm property for outer products.

We next analyze the gradient of the route probability $p_{l_t,t}(\boldsymbol{x})$. Let $W_{\max} := \max_{i,t} \|\boldsymbol{w}_{i,t}\|$ be the maximum norm of gating weights, $\Pi_{\max} := \max_{l_t,t} \|\boldsymbol{\pi}_{l_t,t}\|$ be the maximum norm of leaf vectors, and $D_{\max}$ be the maximum tree depth. According to Proposition 15, we have

$$\begin{aligned} \left\| \nabla_{\boldsymbol{x}} p_{l_t,t}(\boldsymbol{x}) \right\| &= \left\| p_{l_t,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l_t)} \psi_{i,t} \boldsymbol{w}_{i,t} \right\| \leq p_{l_t,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l_t)} \left\| \psi_{i,t} \boldsymbol{w}_{i,t} \right\| \leq p_{l_t,t}(\boldsymbol{x}) \cdot \sum_{i \in \Lambda(l_t)} \left\| \boldsymbol{w}_{i,t} \right\| \\ &\leq p_{l_t,t}(\boldsymbol{x}) \cdot D_{\max} \cdot W_{\max}. \end{aligned} \tag{74}$$

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

Based on Eq. (74), Eq. (73) can be bounded by

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

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

For Lipschitz smoothness, we bound the spectral norm of the Hessian of the route probability:

$$\begin{aligned}\left\|\nabla_{\boldsymbol{x}}^2 p_{l_t,t}(\boldsymbol{x})\right\| &= \Bigg\| p_{l_t,t}(\boldsymbol{x}) \cdot \Bigg[\Big(\sum_{i\in\Lambda(l_t)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big) \Big(\sum_{i\in\Lambda(l_t)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big)^\top - \\ &\qquad \sum_{i\in\Lambda(l_t)} \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big(1 - \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big)\boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}^\top \Bigg] \Bigg\| \\ &\leq p_{l_t,t}(\boldsymbol{x}) \cdot \Bigg[\Big\| \Big(\sum_{i\in\Lambda(l_t)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big) \Big(\sum_{i\in\Lambda(l_t)} \psi_{i,t}\boldsymbol{w}_{i,t} \Big)^\top \Big\| + \\ &\qquad \Big\| \sum_{i\in\Lambda(l_t)} \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big(1 - \mathrm{S}\Big((\boldsymbol{w}_{i,t})^\top \boldsymbol{x} + b_{i,t}\Big)\Big)\boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}^\top \Big\| \Bigg] \\ &\leq p_{l_t,t}(\boldsymbol{x}) \cdot \Bigg[\Big\| \Big(\sum_{i\in\Lambda(l_t)} \boldsymbol{w}_{i,t} \Big) \Big(\sum_{i\in\Lambda(l_t)} \boldsymbol{w}_{i,t} \Big)^\top \Big\| + \frac{1}{4} \Big\| \sum_{i\in\Lambda(l_t)} \boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}^\top \Big\| \Bigg] \\ &\leq p_{l_t,t}(\boldsymbol{x}) \cdot \Bigg[\sum_{i\in\Lambda(l_t)} \Big\|\boldsymbol{w}_{i,t}\Big\| \cdot \sum_{i\in\Lambda(l_t)} \Big\|\boldsymbol{w}_{i,t}\Big\| + \frac{1}{4} \sum_{i\in\Lambda(l_t)} \Big\|\boldsymbol{w}_{i,t}\Big\| \Big\|\boldsymbol{w}_{i,t}\Big\| \Bigg] \\ &\leq p_{l_t,t}(\boldsymbol{x}) \cdot \Bigg[D_{\max}^2 \cdot W_{\max}^2 + \frac{1}{4} D_{\max} \cdot W_{\max}^2 \Bigg],\end{aligned} \tag{75}$$

where the first inequality is due to the triangle inequality, the second inequality is due to the range of $\psi_{i,t}$ and the 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 Eq. (73), the upper bound of the gradient of $\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})$ is given by

$$\begin{aligned}\|\nabla_{\boldsymbol{x}}^2 f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})\| &= \Big\| \frac{1}{T}\sum_{t=1}^{T}\sum_{l_t=1}^{2^{D(t)}} \nabla_{\boldsymbol{x}}^2 p_{l_t,t}(\boldsymbol{x}) \cdot \boldsymbol{\pi}_{l_t,t}^\top \Big\| \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{l_t=1}^{2^{D(t)}} \Big\|\nabla_{\boldsymbol{x}}^2 p_{l_t,t}(\boldsymbol{x})\Big\| \cdot \Big\|\boldsymbol{\pi}_{l_t,t}\Big\| \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{l_t=1}^{2^{D(t)}} p_{l_t,t}(\boldsymbol{x}) \cdot \Bigg[D_{\max}^2 \cdot W_{\max}^2 + \frac{1}{4} D_{\max} \cdot W_{\max}^2 \Bigg] \cdot \Pi_{\max} \\ &= D_{\max}(D_{\max} + \frac{1}{4}) \cdot W_{\max}^2 \cdot \Pi_{\max} \\ &= S^{\text{SRF}},\end{aligned}$$

where the second inequality is due to the definition of $\Pi_{\max}$ and Eq. (75). Therefore, $f_{\boldsymbol{\theta}}^{\text{SRF}}(\boldsymbol{x})$ is S^{SRF}-Lipschitz smoothness. □

G.8 Proof of Proposition 18 in Appendix E.1

Proof of Proposition 18. 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 $\boldsymbol{u}(\boldsymbol{\theta}; \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) : \Theta \to [0, B]^N$

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

For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and samples $\widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2$, and $\{\boldsymbol{y}\}_{i=1}^N$, we write $\boldsymbol{u}(\boldsymbol{\theta}) := \boldsymbol{u}(\boldsymbol{\theta}; \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ and $\boldsymbol{u}(\boldsymbol{\theta}') := \boldsymbol{u}(\boldsymbol{\theta}'; \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ for brevity. Then, under Assumption 2(II), we have

$$\|\boldsymbol{u}(\boldsymbol{\theta}) - \boldsymbol{u}(\boldsymbol{\theta}')\|^2 = \sum_{i\in[N]} |u_i(\boldsymbol{\theta}) - u_i(\boldsymbol{\theta}')|^2 \le N L_{\boldsymbol{\theta}}^2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \tag{76}$$

For each $i \in [N]$, we have

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

which implies that

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

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

$$\begin{aligned}
\|t_3(\boldsymbol{u}(\boldsymbol{\theta})) - t_3(\boldsymbol{u}(\boldsymbol{\theta}'))\|^2 &= \sum_{i\in[N]} \Big|\Big[t_3(\boldsymbol{u}(\boldsymbol{\theta}))\Big]_i - \Big[t_3(\boldsymbol{u}(\boldsymbol{\theta}'))\Big]_i\Big|^2 \\
&\le (L_3')^2 \sum_{i\in[N]} |u_i(\boldsymbol{\theta}) - u_i(\boldsymbol{\theta}')|^2 = (L_3')^2 \|\boldsymbol{u}(\boldsymbol{\theta}) - \boldsymbol{u}(\boldsymbol{\theta}')\|^2 \\
&\le L_3^2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2,
\end{aligned} \tag{77}$$

where $L_3 = \sqrt{N} L_{\boldsymbol{\theta}} L_3' = \frac{\sqrt{N} L_{\boldsymbol{\theta}}}{\lambda\epsilon}\exp(B/\lambda\epsilon)$, and the final inequality is due to Eq. (76). This result shows that the function t_3 is L_3-Lipschitz continuous.

According to the second-order derivation of the function t_3, its Jacobian matrix is given by

$$J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta}) = T(\boldsymbol{\theta}) J_{\boldsymbol{\theta}} \boldsymbol{u}(\boldsymbol{\theta}),$$

where $T(\boldsymbol{\theta}) := J_{\boldsymbol{u}} t_3(\boldsymbol{u}(\boldsymbol{\theta})) = \operatorname{diag}\Big(\frac{1}{\lambda\epsilon}\exp(u_1(\boldsymbol{\theta})/\lambda\epsilon), \cdot, \frac{1}{\lambda\epsilon}\exp(u_N(\boldsymbol{\theta})/\lambda\epsilon)\Big)$ is an N-dimensional diagonal matrix. According to Eq. (77), we have $\|T(\boldsymbol{\theta})\|_{\text{op}} \le L_3'$ and

$$\begin{aligned}
\|T(\boldsymbol{\theta}) - T(\boldsymbol{\theta}')\|_{\text{op}} &= \max_{j\in[N]} \Big|\frac{\exp(u_j(\boldsymbol{\theta})/\lambda\epsilon)}{\lambda\epsilon} - \frac{\exp(u_j(\boldsymbol{\theta}')/\lambda\epsilon)}{\lambda\epsilon}\Big| \\
&\le S_3' \cdot \max_{j\in[N]} \Big|u_j(\boldsymbol{\theta}) - u_j(\boldsymbol{\theta}')\Big| \\
&\le S_3' \cdot L_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|,
\end{aligned}$$

where the first inequality is because the exponential function is Lipschitz continuous on the interval $[0, B/\lambda\epsilon]$ and $S_3' = \frac{1}{(\lambda\epsilon)^2}\exp(B/\lambda\epsilon)$, and the second inequality follows from Eq. (76), i.e., $|u_i(\boldsymbol{\theta}) - u_i(\boldsymbol{\theta}')| \le L_{\boldsymbol{\theta}} \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ for any $i \in [N]$.

In Assumption 2(II), for fixed $\boldsymbol{x}$ and $\boldsymbol{y}$, the Lipschitz smoothness implies that

$$\|\nabla_{\boldsymbol{\theta}} u_i(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} u_i(\boldsymbol{\theta}')\|_2 \le S_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \quad \forall i \in [N].$$

Then, for $J_{\boldsymbol{\theta}} \boldsymbol{u}(\boldsymbol{\theta})$, we have

$$\|J_{\boldsymbol{\theta}} \boldsymbol{u}(\boldsymbol{\theta})\|_{\text{F}}^2 = \sum_{i\in[N]} \|\nabla_\theta u_i(\boldsymbol{\theta})\|_2^2 \le N \cdot L_{\boldsymbol{\theta}}^2,$$

$$\|J_\theta \boldsymbol{u}(\boldsymbol{\theta}) - J_\theta \boldsymbol{u}\left(\boldsymbol{\theta}'\right)\|_{\mathrm{F}}^2 = \sum_{j=1} \|\nabla_\theta u_j(\boldsymbol{\theta}) - \nabla_\theta u_j\left(\boldsymbol{\theta}'\right)\|_2^2 \leq N S_{\boldsymbol{\theta}}^2 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 .$$

Therefore,

$$\begin{aligned}
\|J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta}) - J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta}')\|_{\mathrm{F}} &= \|T(\boldsymbol{\theta})\Big(J_\theta \boldsymbol{u}(\boldsymbol{\theta}) - J_\theta \boldsymbol{u}\left(\boldsymbol{\theta}'\right)\Big) + \Big(T(\boldsymbol{\theta}) - T(\boldsymbol{\theta}')\Big) J_{\boldsymbol{\theta}} \boldsymbol{u}(\boldsymbol{\theta}')\|_{\mathrm{F}} \\
&\leq \Big\|T(\boldsymbol{\theta})\Big(J_\theta \boldsymbol{u}(\boldsymbol{\theta}) - J_\theta \boldsymbol{u}\left(\boldsymbol{\theta}'\right)\Big)\Big\|_{\mathrm{F}} + \Big\|\Big(T(\boldsymbol{\theta}) - T(\boldsymbol{\theta}')\Big) J_{\boldsymbol{\theta}} \boldsymbol{u}(\boldsymbol{\theta}')\Big\|_{\mathrm{F}} \\
&\leq \|T(\boldsymbol{\theta})\|_{\mathrm{op}} \|J_\theta \boldsymbol{u}(\boldsymbol{\theta}) - J_\theta \boldsymbol{u}\left(\boldsymbol{\theta}'\right)\|_{\mathrm{F}} + \|T(\boldsymbol{\theta}) - T(\boldsymbol{\theta}')\|_{\mathrm{op}} \|J_{\boldsymbol{\theta}} \boldsymbol{u}(\boldsymbol{\theta})\|_{\mathrm{F}} \\
&\leq L_3' \cdot \sqrt{N} S_{\boldsymbol{\theta}} \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 + \sqrt{N} \cdot L_{\boldsymbol{\theta}} S_3' \cdot L_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \\
&= S_3 \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2,
\end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality is due to the property of matrix norm, i.e., $\|\boldsymbol{AB}\|_{\mathrm{F}} \leq \|\boldsymbol{A}\|_{\mathrm{op}} \|\boldsymbol{B}\|_{\mathrm{F}}$, and

$$S_3 := \sqrt{N}\Big(L_3' S_{\boldsymbol{\theta}} + L_{\boldsymbol{\theta}} S_3'\Big) = \sqrt{N} e^{B/\lambda\epsilon}\Big(\frac{S_{\boldsymbol{\theta}}}{\lambda\epsilon} + \frac{L_{\boldsymbol{\theta}}^2}{\lambda^2\epsilon^2}\Big).$$

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

$$\|J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta})\|_{\mathrm{F}} \leq \|T(\boldsymbol{\theta})\|_{\mathrm{op}} \|J_\theta \boldsymbol{u}(\boldsymbol{\theta})\|_{\mathrm{F}} \leq \sqrt{N} L_{\boldsymbol{\theta}} L_3',$$

and thus

$$\mathbb{E}\Big[\|J_{\boldsymbol{\theta}} t_3(\boldsymbol{\theta})\|_{\mathrm{F}}^2\Big] \leq N L_{\boldsymbol{\theta}}^2 (L_3')^2 = C_3^2,$$

where $C_3 = L_3$, which implies that the Jacobian of t_3 with respect to $\boldsymbol{\theta}$ has a uniformly bounded second moment.

According to a fundamental result in Casella & Berger (2024), the variance of a bounded random variable (or vector) is finite. Thus, to prove that the function t_3 has finite variance, it suffices to show that its norm is bounded. According to Assumption 2(III), given $\widehat{\boldsymbol{x}}$, we have

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

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

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

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

Therefore, we obtain

$$\begin{aligned}
\|\nabla t_2(\boldsymbol{v})\|^2 &= \sum_{i=1}^{N} \Big(\frac{\partial t_2(\boldsymbol{v})}{\partial v_i}\Big)^2 \\
&= \sum_{i=1}^{N} \Big(t_2(\boldsymbol{v}) \cdot \frac{p_i}{v_i}\Big)^2 \\
&\leq \exp\Big(2B/\lambda\epsilon\Big).
\end{aligned} \tag{78}$$

where the inequality arises from the domain and range of the function t_2 and $\sum_{j=1}^N p_j^2 \leq 1$. Let $L_2 = C_2 = \exp(B/\lambda\epsilon)$. From Eq. (78), the function t_2 is $L2$-Lipschitz continuous and has bounded stochastic gradients in expectation, that is, $\mathbb{E}\Big[\|\nabla t_2(\boldsymbol{v}; \widehat{\boldsymbol{x}}, \boldsymbol{\xi}_1)\|^2\Big] \leq C_2^2$, and its variance is also finite.

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

$$\begin{aligned} \Big| t_2(\boldsymbol{v}_1) - t_2(\boldsymbol{v}_2) - \nabla t_2(\boldsymbol{v}_2)^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2) \Big| &= \Big| \frac{1}{2} \cdot (\boldsymbol{v}_1 - \boldsymbol{v}_2)^\top \nabla^2 t_2(\boldsymbol{c})(\boldsymbol{v}_1 - \boldsymbol{v}_2) \Big| \\ &\leq \frac{1}{2} \|\nabla^2 t_2(\boldsymbol{c})\|_2 \cdot \|(\boldsymbol{v}_1 - \boldsymbol{v}_2)\|^2, \end{aligned} \tag{79}$$

where the inequality is due to the Cauchy-Schwarz inequality. According to the definition of Lipschitz smoothness, if $\|\nabla^2 t_2(\boldsymbol{c})\|_2$ is bounded by a constant S_2, that is, $\|\nabla^2 t_2(\boldsymbol{c})\|_2 \leq S_2$, then the function t_2 is S_2-Lipschitz smooth. Denote the j-th element in vector $\boldsymbol{c}$ as c_j for any $j \in [N]$. For any $j, k \in [N]$, 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

$$\begin{aligned} \|\nabla^2 t_2(\boldsymbol{c})\|_2^2 &\leq \|\nabla^2 t_2(\boldsymbol{c})\|_{\mathrm{F}}^2 \\ &= \sum_{j=1}^{N} \Big(t_2(\boldsymbol{c}) \left(\frac{p_j^2}{c_j^2} - \frac{p_j}{c_j^2} \right) \Big)^2 + \sum_{j \neq k} \Big(t_2(\boldsymbol{c}) \frac{p_j p_k}{c_j c_k} \Big)^2 \\ &\leq \sum_{j=1}^{N} \Big(e^{\frac{B}{\lambda \epsilon}} \left(\frac{p_j - p_j^2}{1} \right) \Big)^2 + \sum_{j \neq k} \Big(e^{\frac{B}{\lambda \epsilon}} \frac{p_j p_k}{1} \Big)^2 \\ &\leq e^{\frac{2B}{\lambda \epsilon}} \Big[\sum_{j=1}^{N} p_j^2 + \Big(\sum_{j=1}^{N} p_j^2 \Big)^2 \Big] \\ &\leq 2 \cdot e^{\frac{2B}{\lambda \epsilon}}, \end{aligned}$$

where the first inequality is because the Frobenius norm $\|\cdot\|_{\mathrm{F}}$ is an upper bound of the spectral 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 $j \in [N]$ and the Cauchy-Schwarz inequality, and the last inequality is due to $\sum_{j=1}^{N} p_j^2 \leq 1$. Let $S_2 = \sqrt{2}\exp(B/\lambda\epsilon)$, based on relation (79), we have

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

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

- **For function** t_1, its range is $[0, B/\lambda\epsilon]$ because its domain is the same as the range of function t_2, that is, $[1, \exp(B/\lambda\epsilon)]$. 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 $z \in \big[1, \exp(B/\lambda\epsilon)\big]$, we can bound the absolute values of these derivatives:

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

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

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

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

Since the value of the function t_1 is bounded, its variance is finite, that is,

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

This completes the proof. □

G.9 Proof of Theorem 19 in Appendix E.2

As an essential part of sample complexity analysis, we first introduce the following lemmas. Based on Cramér's large deviations theorem, we introduce the following Lemma 24.

Lemma 24. (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*

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

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 $\delta\in(0,\delta_1)$, $I(\delta)\geq \frac{\delta^2}{(2+\kappa)\sigma^2}$.

Hu et al. (2020) extend Cramér's Inequality from random variables to random vectors, as shown in Lemma 25.

Lemma 25. (Concentration Inequality, 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 $\mathbb{E}\|\boldsymbol{X}\|^2=\sigma^2<\infty$. Then for any $\kappa>0$, there exists $\delta_1>0$ such that for any $\delta\in(0,\delta_1)$, it holds that*

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

Using Lemma 25, we present the following Lemma 26.

Lemma 26. *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*

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

Proof of Lemma 26. For $\upsilon\in(0,1)$, the set $\{\boldsymbol{x}_l\}_{l=1}^Q$ is said to be a υ-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 $\|\boldsymbol{x}-\boldsymbol{x}_{l(\boldsymbol{x})}\|\leq\upsilon$. We construct a υ-net to eliminate the supremum over $\boldsymbol{\theta}$ and apply a concentration inequality to bound the probability. First, we pick a υ-net $\{\boldsymbol{\theta}_l\}_{l=1}^Q$ on the decision set $\Theta\subseteq\mathbb{R}^{d_\theta}$, such that $L_1L_2L_3\upsilon=\delta/4$. Under Assumption 2(I), Θ has a finite diameter D_Θ, for any $\upsilon\in(0,1)$, there exists a υ-net of Θ, and the size of the υ-net is bounded, $Q\leq\mathcal{O}((D_\times/\upsilon)^{d_\theta})$ (Shapiro et al., 2021). By definition of υ-net, we have $\forall\boldsymbol{\theta}\in\Theta,\exists l(\boldsymbol{\theta})\in\{1,2,\ldots,Q\}$, s.t.

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

Based on Proposition 18, we have

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

and

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

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

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

It follows that

$$\begin{aligned}
&\Pr\left(\sup_{\boldsymbol{\theta}\in\Theta}\left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}\right)-F\left(\boldsymbol{\theta}\right)\right|>\delta\right) \\
\leq&\Pr\left(\sum_{l=1}^{Q}\left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_l\right)-F\left(\boldsymbol{\theta}_l\right)\right|>\frac{\delta}{2}\right) \leq \sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_l\right)-F\left(\boldsymbol{\theta}_l\right)\right|>\frac{\delta}{2}\right) \\
\leq&\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_l\right)-\widehat{F}_{N_1,N_2}\left(\boldsymbol{\theta}_l\right)\right|+\left|\widehat{F}_{N_1,N_2}\left(\boldsymbol{\theta}_l\right)-\widehat{F}_{N_1}\left(\boldsymbol{\theta}_l\right)\right|+\left|\widehat{F}_{N_1}\left(\boldsymbol{\theta}_l\right)-F\left(\boldsymbol{\theta}_l\right)\right|>\frac{\delta}{2}\right) \\
\leq&\underbrace{\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_l\right)-\widehat{F}_{N_1,N_2}\left(\boldsymbol{\theta}_l\right)\right|>\frac{\delta}{6}\right)}_{\Delta_1}+\underbrace{\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2}\left(\boldsymbol{\theta}_l\right)-\widehat{F}_{N_1}\left(\boldsymbol{\theta}_l\right)\right|>\frac{\delta}{6}\right)}_{\Delta_2} \\
&+\underbrace{\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1}\left(\boldsymbol{\theta}_l\right)-F\left(\boldsymbol{\theta}_l\right)\right|>\frac{\delta}{6}\right)}_{\Delta_3}.
\end{aligned} \tag{80}$$

For the term Δ_1 in Eq. (80), we have

$$\begin{aligned}
\Delta_1=&\sum_{l=1}^{Q}\Pr\left(\left|\widehat{F}_{N_1,N_2,N_3}\left(\boldsymbol{\theta}_l\right)-\widehat{F}_{N_1,N_2}\left(\boldsymbol{\theta}_l\right)\right|>\frac{\delta}{6}\right) \\
=&\sum_{l=1}^{Q}\Pr\Big(\Big|\frac{\lambda\epsilon}{N_1}\sum_{i=1}^{N_1}t_1\Big(\frac{1}{N_2}\sum_{j=1}^{N_2}t_2\Big(\frac{1}{N_3}\sum_{k=1}^{N_3}t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\boldsymbol{\xi}_2^k\Big);\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j\Big)\Big) \\
&-\frac{\lambda\epsilon}{N_1}\sum_{i=1}^{N_1}t_1\Big(\frac{1}{N_2}\sum_{j=1}^{N_2}t_2\Big(\mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\boldsymbol{\xi}_2^k\Big)\Big];\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j\Big)\Big)\left(\boldsymbol{\theta}_l\right)\Big|>\frac{\delta}{6}\Big).
\end{aligned} \tag{81}$$

By Proposition 18,

$$\begin{aligned}
\Delta_1 \leq&\sum_{l=1}^{Q}\Pr\left(\max_{i=1,\cdots,N_1;j=1,\cdots,N_2}\left|L_1L_2\|\frac{1}{N_3}\sum_{k=1}^{N_3}t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\boldsymbol{\xi}_2^k\Big)-\mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\boldsymbol{\xi}_2^k\Big)\Big]\|\right|>\frac{\delta}{6\lambda\epsilon}\right) \\
\leq&\sum_{l=1}^{Q}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\Pr\left(\|\frac{1}{N_3}\sum_{k=1}^{N_3}t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\boldsymbol{\xi}_2^k\Big)-\mathbb{E}_{\boldsymbol{\xi}_2}\Big[t_3\Big(\boldsymbol{\theta};\widehat{\boldsymbol{x}}^i,\boldsymbol{\xi}_1^j,\boldsymbol{\xi}_2^k\Big)\Big]\|>\frac{\delta}{6\lambda\epsilon L_1L_2}\right) \\
\leq&QN_1N_2\cdot 2N_1\exp\Big(-\frac{N_3\delta^2}{36(2+\kappa)\lambda^2\epsilon^2L_1^2L_2^2\sigma_3^2}\Big),
\end{aligned} \tag{82}$$

where the first inequality follows from Lipschitz continuity, and the last inequality follows from Lemma 25. Similarly, we obtain

$$\Delta_2 \leq QN_1 \cdot 2\exp\Big(-\frac{N_2\delta^2}{36(2+\kappa)\lambda^2\epsilon^2L_1^2\sigma_2^2}\Big), \tag{83}$$

and

$$\Delta_3 \leq Q \cdot 2\exp\Big(-\frac{N_1\delta^2}{36(2+\kappa)\lambda^2\epsilon^2\sigma_1^2}\Big). \tag{84}$$

Combining with Eqs. (80)- (84), and the fact that $Q \leq \mathcal{O}(1)\left(4L_1L_2L_3D_\Theta/\delta\right)^{d_\theta}$, we can obtain the desired result of Lemma 26. □

In the following, we prove the results in Theorem 19.

Proof of Theorem 19. (I) For $\Pr\left(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) - F(\boldsymbol{\theta}^*) \leq \delta\right)$ in Theorem 19(I), we have

$$\begin{aligned}
&\Pr\left(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) - F(\boldsymbol{\theta}^*) > \delta\right)\\
=&\Pr\Big(\Big[F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) - \widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)\Big]\\
&\quad + \Big[\widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) - \widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*)\Big] + \Big[\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*) - F(\boldsymbol{\theta}^*)\Big] > \delta\Big)\\
\leq&\Pr\left(F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) - \widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) > \frac{\delta}{2}\right) + \Pr\left(\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*) - F(\boldsymbol{\theta}^*) > \frac{\delta}{2}\right)\\
\leq&\Pr\left(\Big|F\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) - \widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right)\Big| > \frac{\delta}{2}\right) + \Pr\left(\Big|\widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*) - F(\boldsymbol{\theta}^*)\Big| > \frac{\delta}{2}\right)
\end{aligned} \tag{85}$$

where the first inequality is due to $\widehat{F}_{N_1,N_2,N_3}\left(\widehat{\boldsymbol{\theta}}_{N_1,N_2,N_3}\right) - \widehat{F}_{N_1,N_2,N_3}(\boldsymbol{\theta}^*) \leq 0$. Using the result of Lemma 26, we can obtain the desired result of Theorem 19(I) using Eq. (85).

(II) For Theorem 19(II), to analyze

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

it suffices to study

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

Let each of the three terms on the right-hand side (RHS) of the inequality in Theorem 19(I) be no more than $\alpha/3$. This leads to

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

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

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

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 the problem (SAA) for achieving a δ-optimal solution is $T = N_1 + N_2 + N_3 = \mathcal{O}\Big(d_\theta/\delta^2\Big)$.

This completes the proof. □