Robust Decision-Focused Learning via Worst-Case Regret Minimization

Shoki Yamao[1] **Ken Kobayashi**[1] **Ryo Matsui**[2] **Shota Nagai**[1] **Naoki Nishimura**[2] **Kazuhide Nakata**[1]

[1]Institute of Science Tokyo, Tokyo, Japan
[2]Recruit Co., Ltd., Chiyoda, Tokyo, Japan,

Abstract

In optimization-based decision-making, when the objective coefficient vector is unknown, a common approach is to predict it from covariates using a predictive mode and solve the downstream optimization problem. However, improving predictive accuracy does not necessarily lead to better decisions. This has motivated Decision-Focused Learning (DFL), which trains predictive models to minimize decision loss measured by regret. Despite recent progress, existing DFL methods do not sufficiently address two sources of uncertainty: (i) observation errors in the measured coefficients, and (ii) distribution shift in the coefficient vector at deployment due to environmental changes. These uncertainties can degrade solution quality. In this paper, we propose two robust regret losses to address these uncertainties. For uncertainty (i), we introduce an uncertainty set around the observed coefficient vector that captures measurement errors and define the loss as the worst-case regret over this set. For uncertainty (ii), we construct a Wasserstein ambiguity set around the conditional empirical distribution and define the loss using the worst-case distribution within the set. We enable efficient training via Danskin-based subgradients. Experiments demonstrate that our method reduces regret and yields more stable solutions than existing robust DFL approaches.

1 INTRODUCTION

Many real-world decision-making tasks can be formulated as mathematical optimization problems, including shortest-path and resource-allocation problems [Hillier and Lieberman, 2015, Bertsimas and Tsitsiklis, 1997]. In such problems, parameters such as travel times and demands appear as coefficients in the objective function. We denote this coefficient vector by $\boldsymbol{c} \in \mathbb{R}^n$. However, these coefficients are often unknown at the time of decision-making, so the optimization problem cannot be solved directly. A widely studied approach is the two-stage predict-then-optimize pipeline: given covariates $\boldsymbol{x} \in \mathcal{X}$ (e.g., weather, seasonality, and historical demand), a predictive model $f_{\boldsymbol{\theta}}$ estimates the unknown coefficients as $\hat{\boldsymbol{c}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$, and we then solve the downstream optimization problem

$$z(\hat{\boldsymbol{c}}) := \min_{\boldsymbol{w} \in \mathcal{W}} \hat{\boldsymbol{c}}^\top \boldsymbol{w}, \tag{1}$$

where $\mathcal{W}$ is a nonempty compact feasible set, $\boldsymbol{w}$ denotes the decision variables, and $\boldsymbol{\theta} \in \Theta$ denotes the parameters of the predictive model [Bertsimas and Kallus, 2020, Huber et al., 2019, Rios et al., 2015, El Balghiti et al., 2023, Elmachtoub and Grigas, 2022, Sadana et al., 2025]. As a concrete example, let us consider energy scheduling, where the coefficients $\boldsymbol{c}$ are the hourly electricity prices. A model produces predicted prices $\hat{\boldsymbol{c}}$ from covariates $\boldsymbol{x}$, and these prices are used to determine an hourly consumption schedule $\boldsymbol{w}$ subject to operational constraints. Since the realized prices are revealed only after the schedule is committed, the decision must be made before $\boldsymbol{c}$ is known.

Prior work typically trained the predictive model $f_{\boldsymbol{\theta}}$ by minimizing standard prediction losses such as mean squared error. Under this paradigm, the goal is to make the predicted coefficient vector $\hat{\boldsymbol{c}}$ match the ground-truth coefficients $\boldsymbol{c}$ as closely as possible, with the expectation that this would improve the quality of the downstream optimal solution [Bertsimas and Kallus, 2020, Huber et al., 2019, Rios et al., 2015]. However, it has been recognized that higher predictive accuracy does not necessarily lead to better downstream solutions [Elmachtoub and Grigas, 2022, Sadana et al., 2025, Donti et al., 2017, Wilder et al., 2019].

Motivated by this observation, recent work has moved away from focusing solely on matching the predicted coefficients to the ground-truth coefficients. Instead, these approaches directly target the quality of the solution obtained when solving the downstream optimization problem with those

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7701–7736.

predictions [Elmachtoub and Grigas, 2022, Sadana et al., 2025, Donti et al., 2017, Wilder et al., 2019, Kotary et al., 2021, Mandi et al., 2024, Bengio et al., 2021]. This paradigm is known as Decision-Focused Learning (DFL). DFL aims to couple prediction and optimization by training model parameters to directly minimize a decision loss (e.g., regret), thereby improving downstream solution quality under the predicted coefficient vector.

Despite this progress, existing DFL methods often do not adequately account for the following practical uncertainties in real-world deployments:

- **Uncertainty (i):** Observation errors caused by sensor aging in training data.
- **Uncertainty (ii):** Distribution shift between training and deployment, where the conditional distribution of the cost coefficients given the covariates changes due to environmental variations.

Because conventional approaches train predictive models without explicitly accounting for (i) and (ii), the resulting decisions can be brittle, and the downstream solution quality can degrade substantially. This motivates a new DFL framework that incorporates these uncertainties to enable robust decision-making.

Prior work has incorporated ideas from robust optimization into DFL to handle Uncertainty (i) [Schutte et al., 2024, Wang et al., 2026]. Schutte et al. [2024] study robustness to uncertainty arising from limited data and noisy observations and propose three robust loss formulations. Wang et al. [2026] use a generative model to learn the distribution of coefficient vectors, aiming to learn a model that is robust to the predicted distribution.

However, important limitations remain. First, existing robust DFL methods [Schutte et al., 2024, Wang et al., 2026] do not explicitly control the worst-case decision loss induced by observation errors, leaving them vulnerable to noisy measurements. Second, a robust DFL for Uncertainty (ii) has not been established where the underlying data-generating process changes between training and deployment. Most approaches minimize empirical risk under the training distribution, offering no guarantees when the testing distribution deviates. We discuss these limitations in more detail in Section 2.

To address these issues, we propose a unified framework that evaluates the worst-case regret under uncertainty. Based on this framework, we introduce two robust loss functions targeting each source of uncertainty:

- **For Uncertainty (i) (Observation Errors):** Robust Regret Loss $\mathcal{L}_{\mathrm{rob}}$, which minimizes the worst-case regret over a parameter-based uncertainty set.
- **For Uncertainty (ii) (Distribution Shift):** Distributionally Robust Regret Loss $\mathcal{L}_{\mathrm{dro}}$, which minimizes the worst-case expected regret over a Wasserstein ambiguity set.

Concretely, we leverage robust optimization techniques [Ben-Tal et al., 2009, Bertsimas and Sim, 2004] to model observation errors (Uncertainty (i)) and distributionally robust optimization (DRO) techniques [Delage and Ye, 2010, Mohajerin Esfahani and Kuhn, 2018] to model distribution shifts (Uncertainty (ii)), yielding uncertainty sets for the coefficient vector. We then formulate a learning objective that minimizes the worst-case decision loss over these sets. Our goal is to obtain models that reliably produce high-quality solutions even in the presence of observation errors or distribution shifts in the coefficient vector.

However, directly minimizing these worst-case measures presents computational challenges because the optimization involves infinite-dimensional constraints or nested maximization problems, making direct gradient computation difficult. To enable efficient training, we reformulate these optimization problems into tractable forms solvable by standard techniques and derive subgradients for the proposed losses.

The main contributions of this paper are as follows:

1. To address observation errors, we develop a robust DFL loss that models such errors via a parameter-space uncertainty set and minimizes the worst-case decision loss.
2. To handle distribution shifts, we propose a distributionally robust DFL loss that minimizes the expected decision loss under the worst-case conditional distribution within a Wasserstein ambiguity set.
3. We reduce the infinite-dimensional distribution optimization problem to a computationally tractable regularized loss function.
4. We derive subgradients for the proposed worst-case loss functions to establish an efficient gradient-based learning algorithm.
5. Numerical experiments demonstrate that our method improves downstream solution quality under uncertainty in terms of average regret and risk metrics compared with existing DFL baselines.

2 PRELIMINARIES

2.1 DECISION-FOCUSED LEARNING

In approaches that decouple learning a prediction model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^n$ from solving the downstream optimization problem, the training loss typically targets prediction accuracy. However, more accurate predictions of the coefficient vector $\hat{\boldsymbol{c}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ do not necessarily yield better decisions, i.e., better solutions $\boldsymbol{w}^*(\hat{\boldsymbol{c}}) \in \arg\min_{\boldsymbol{w}\in\mathcal{W}} \hat{\boldsymbol{c}}^\top \boldsymbol{w}$. Decision-focused learning (DFL), in contrast, trains $f_{\boldsymbol{\theta}}$ end-to-end to improve downstream decision quality by incorporating the optimization problem into the learning objective.

The downstream decision quality is commonly quantified by the regret [Sadana et al., 2025]:

$$\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c}) := \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}). \tag{2}$$

This quantity measures how much the decision optimized for the prediction, $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$, degrades the true objective value relative to the true optimum under $\boldsymbol{c}$.

In standard DFL, given training data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{c}_i)\}_{i=1}^N$, we learn the model parameters $\boldsymbol{\theta}$ by minimizing the empirical average regret:

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathrm{Reg}\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{c}_i\big). \tag{3}$$

Because $\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c})$ is generally nonconvex and discontinuous in the prediction $\hat{\boldsymbol{c}}$, practitioners typically optimize a continuous convex surrogate, such as the SPO+ loss [Elmachtoub and Grigas, 2022].

2.2 RELATED WORK

Due to space limitations, we provide a brief overview of related work here. A more detailed discussion is available in Appendix H.

2.2.1 Existing research on Uncertainty (i)

Several approaches have been proposed to address Uncertainty (i) in DFL. Schutte et al. [2024] proposed three robust loss functions to handle Uncertainty (i). The first method, the k-NN loss, replaces the coefficient vector $\boldsymbol{c}$ in (2) with a local average $\bar{\boldsymbol{c}}$ computed from neighbors in the covariate space. However, this approach can be sensitive to large observation errors and does not guarantee worst-case regret robustness. The second method is the RO loss, which uses a robust solution $\boldsymbol{w}^{\mathrm{RO}}(\boldsymbol{c})$ defined with respect to an uncertainty set $\mathcal{C}$:

$$\mathcal{L}_{\mathrm{RO}}(\hat{\boldsymbol{c}}, \boldsymbol{c}) = \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{c}^\top \boldsymbol{w}^{\mathrm{RO}}(\boldsymbol{c}). \tag{4}$$

Here, $\boldsymbol{w}^{\mathrm{RO}}(\boldsymbol{c}) \in \arg\min_{\boldsymbol{w}\in\mathcal{W}} \max_{\boldsymbol{\delta}\in\mathcal{C}} (\boldsymbol{c}+\boldsymbol{\delta})^\top \boldsymbol{w}$. Crucially, the second term $\boldsymbol{c}^\top \boldsymbol{w}^{\mathrm{RO}}(\boldsymbol{c})$ in (4) does not depend on the prediction $\hat{\boldsymbol{c}}$. Consequently, minimizing the RO loss is equivalent to minimizing the standard regret up to a constant shift. The third method also minimizes the standard regret up to a constant shift, failing to explicitly control the worst-case regret induced by Uncertainty (i).

Gen-DFL [Wang et al., 2026] is another approach addressing Uncertainty (i). It trains a generative model q to capture conditional distributions and minimizes the CVaR of regret. However, if the generative model q is trained using observed coefficient vectors $\boldsymbol{c}_{\mathrm{obs}}$ that contain errors, it may learn the observation-noise structure itself. If the learned distribution deviates from the true underlying distribution, optimizing against q does not guarantee robustness under the true distribution.

2.2.2 Existing research on Uncertainty (ii)

While DRO is well-studied, to the best of our knowledge, there is no existing work that applies it to the DFL setting to address Uncertainty (ii). As noted in Uncertainty (ii), standard DFL objectives do not account for shifts in the conditional distribution. Without explicit distributional robustness, the optimizer $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ can be fragile to out-of-distribution data.

3 PROPOSED METHOD

In this section, we present a framework for robust decision-focused learning based on worst-case regret minimization. We study worst-case regret minimization under two uncertainty models: a set-based model for observation errors and a Wasserstein distributional model for deployment shifts. Although these uncertainties arise from different sources, we address them through a common principle: minimizing the worst-case decision loss over a defined uncertainty set. Sections 3.1 and 3.2 detail the specific formulations for observation errors and distribution shifts, respectively.

3.1 DECISION-FOCUSED LEARNING ROBUST TO OBSERVATION ERRORS

In this subsection, we propose a robust DFL loss that accounts for observation errors in the coefficient vector. We first define a robust regret loss function based on worst-case regret minimization over an uncertainty set. Next, we derive its subgradient for model training. However, as the computational cost of the subgradient is high, we also introduce an approximate robust regret loss to improve learning efficiency.

3.1.1 Robust Regret Loss Function

To address observation errors, we regard the observed coefficient vector $\boldsymbol{c}_{\mathrm{obs}}$ as a noisy observation of the true, unobserved coefficient vector $\boldsymbol{c}$, and model the range of plausible perturbations by an uncertainty set $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$ centered at $\boldsymbol{c}_{\mathrm{obs}}$. Our goal is to learn parameters $\boldsymbol{\theta}$ such that the induced decision $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$, with $\hat{\boldsymbol{c}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$, remains reliable against any coefficient vector $\boldsymbol{c}$ within this uncertainty set. Accordingly, we evaluate robustness by the worst-case regret over $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$ and define the following robust regret loss:

$$\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) := \max_{\boldsymbol{c}\in\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \Big\{ \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}) \Big\}, \tag{5}$$

where the uncertainty set $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$ is nonempty and compact.

This loss corresponds to the regret incurred under the worst-case coefficient vector $\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$. Training to minimize $\mathcal{L}_{\mathrm{rob}}$, we obtain a model whose decisions remain robust to observation errors, i.e., it reduces worst-case regret and limits the degradation of solution quality. Typical choices for $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$ include ellipsoidal and budgeted uncertainty

sets [Ben-Tal et al., 2009, Bertsimas and Sim, 2004]; their definitions and detailed formulations are provided in Appendix B.1.

3.1.2 Gradient Calculation of Robust Regret Loss

To train a predictive model (e.g., a neural network) with the robust regret loss (5), we need to compute gradients of the loss with respect to the model output $\hat{\boldsymbol{c}}$ and update the parameters $\boldsymbol{\theta}$. However, the robust regret loss $\mathcal{L}_{\mathrm{rob}}$ involves a maximization over the uncertainty set, and its objective further contains the optimal value function $z(\boldsymbol{c})$. This nested structure makes gradient computation nontrivial. In this section, we derive a subgradient of $\mathcal{L}_{\mathrm{rob}}$ via Danskin's theorem. We then obtain a practical training algorithm by combining the resulting subgradient with the surrogate gradient.

Using Danskin's theorem [Bertsekas, 1999], we derive a subgradient of the robust regret loss $\mathcal{L}_{\mathrm{rob}}$ with respect to the decision $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ obtained from the predicted coefficient vector $\hat{\boldsymbol{c}}$. The subdifferential is characterized by the following theorem. (See Appendix D.1 for the proof.)

Theorem 3.1. *Let $\mathcal{W}$ and $\mathcal{C}$ be nonempty compact sets. Then the subdifferential of $\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}})$ with respect to $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ is given by the convex hull of the worst-case coefficient vectors, i.e.,*

$$\partial_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) = \mathrm{conv}\Big\{\operatorname*{arg\,max}_{\boldsymbol{c}\in\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})}\big(\boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c})\big)\Big\}.$$

Here conv *denotes the convex hull of a set.*

Next, we explain how to update the model parameters $\boldsymbol{\theta}$ using a subgradient with respect to $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$. If $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ is differentiable with respect to $\hat{\boldsymbol{c}}$, the chain rule yields

$$\nabla_{\hat{\boldsymbol{c}}}\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) = J_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}^\top \boldsymbol{c}^*,\ \boldsymbol{c}^* \in \partial_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}),$$

where $J_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$ denotes the Jacobian of $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ with respect to $\hat{\boldsymbol{c}}$. However, differentiating through the argmin map $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ is often intractable. Therefore, we employ existing techniques to compute the surrogate of $\nabla_{\hat{\boldsymbol{c}}}\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}})$ [Elmachtoub and Grigas, 2022, Blondel et al., 2020, Huang and Gupta, 2024]. See Appendix E.2.1 for details of the training algorithm.

3.1.3 Computationally Efficient Training via Approximate Robust Loss

To compute the gradient $\nabla_{\hat{\boldsymbol{c}}}\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}})$, we use the worst-case coefficient vector $\boldsymbol{c}^* \in \arg\max_{c\in\mathcal{C}} \mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c})$ as an element of the subgradient in practice. However, calculating the worst-case coefficient vector is computationally intensive and incurs significant processing time due to solving a bilevel or nonconvex optimization problem at every training iteration. In this section, we propose a computationally efficient surrogate loss that avoids this costly maximization. Specifically, given the current optimal solution $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$, we first select a coefficient vector $\boldsymbol{c}^\dagger$ that maximizes the objective value $\boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}})$ over the uncertainty set $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$:

$$\boldsymbol{c}^\dagger(\hat{\boldsymbol{c}}) \in \operatorname*{arg\,max}_{\boldsymbol{c}\in\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})}\ \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}). \tag{6}$$

Using $\boldsymbol{c}^\dagger(\hat{\boldsymbol{c}})$, we define the following approximate robust regret loss:

$$\widetilde{\mathcal{L}}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) := \boldsymbol{c}^\dagger(\hat{\boldsymbol{c}})^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z\big(\boldsymbol{c}^\dagger(\hat{\boldsymbol{c}})\big). \tag{7}$$

As a result, because the optimal value function $z(\boldsymbol{c})$ no longer appears inside the maximization, the inner problem at each iteration reduces to maximizing a linear function of $\boldsymbol{c}$ for a fixed vector $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$. Moreover, for both ellipsoidal and budgeted uncertainty sets [Ben-Tal et al., 2009, Bertsimas and Sim, 2004], the maximizer $\boldsymbol{c}^\dagger$ admits a closed-form solution; see Appendix B.1.3 for the detailed derivations.

This approximation substantially reduces the computational cost. At each iteration, we only need to solve the optimization problem that produces the decision $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ and avoid solving the additional nonconvex problem required by $\mathcal{L}_{\mathrm{rob}}$. Therefore, using (7) as the training loss enables robust learning against perturbations that degrade solution quality, while keeping the overall training time manageable. Moreover, gradients can be computed in the same manner as in the previous section, using Danskin's theorem together with existing techniques [Elmachtoub and Grigas, 2022, Blondel et al., 2020, Huang and Gupta, 2024]. Intuitively, $\widetilde{\mathcal{L}}_{\mathrm{rob}}$ (`ropt`) pushes the model toward perturbation-robust decisions because it trains against the coefficient vector in the uncertainty set that most degrades the current decision $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$. Although it lacks the formal worst-case guarantees of $\mathcal{L}_{\mathrm{rob}}$, it is a computationally efficient alternative.

3.2 DECISION-FOCUSED LEARNING ROBUST TO DISTRIBUTION SHIFTS

In this section, rather than treating the empirical distribution induced by the observed data as the true distribution, we consider an uncertainty set of candidate true distributions around the empirical distribution. We then adopt a DRO framework that minimizes the expected regret under the worst-case distribution in this set. Specifically, we construct a Wasserstein ambiguity set defined using the Wasserstein distance between distributions and derive a DFL loss that is robust to distribution shifts. We first define the conditional empirical distribution based on the dataset $\mathcal{D} := \{(\boldsymbol{x}_i, \boldsymbol{c}_i)\}_{i=1}^N$. We then formulate the loss using the Wasserstein distance.

Here, we denote by $\delta_{\boldsymbol{c}}$ the Dirac measure concentrated at $\boldsymbol{c}$, i.e., the distribution that assigns probability one to $\boldsymbol{c}$. Given a covariate value $\boldsymbol{x}_i$ in the dataset, define the index set of samples sharing the same covariate as $\mathcal{I}(\boldsymbol{x}_i) := \{j \in$

$\{1, \ldots, N\} \mid \boldsymbol{x}_j = \boldsymbol{x}_i\}$. The conditional empirical distribution of the coefficient vector given $\boldsymbol{x}_i$ is then

$$P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i} := \frac{1}{|\mathcal{I}(\boldsymbol{x}_i)|} \sum_{j \in \mathcal{I}(\boldsymbol{x}_i)} \delta_{\boldsymbol{c}_j}, \quad (|\mathcal{I}(\boldsymbol{x}_i)| \geq 1).$$

In many applications, the same covariate value may not appear multiple times in the training data; hence $\mathcal{I}(\boldsymbol{x}_i) = \{i\}$ in most cases, and $P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}$ reduces to the point mass at $\boldsymbol{c}_i$. To learn a model that is robust to unknown distribution shifts, we consider an ambiguity set that spreads this point-mass distribution and evaluate the worst-case regret over that set.

3.2.1 Robust Loss against Distribution Shift

First, under the conditional empirical distribution of the coefficient vector, the empirical risk can be written as

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}} \big[\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c})\big]. \tag{8}$$

This treats $\mathrm{Reg}(\hat{\boldsymbol{c}}, \tilde{\boldsymbol{c}})$ as the per-sample loss, takes the expectation over $\tilde{\boldsymbol{c}} \sim P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}$ for each fixed covariate value $\boldsymbol{x}_i$, and then averages over $i = 1, \ldots, N$. However, at test time, the conditional distribution of the coefficient vector given $\boldsymbol{x}$ may differ from that observed during training. Minimizing (8) can then overfit to the conditional empirical distribution induced by the training data, potentially increasing regret under the true conditional distribution. To address this issue, we introduce, for each $\boldsymbol{x}_i$, an uncertainty set of conditional distributions and aim to control the worst-case expected regret over that set.

Specifically, for each covariate value $\boldsymbol{x}_i$, we measure the discrepancy between conditional distributions of coefficient vectors using the 1-Wasserstein distance:

$$W_1\big(Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}, P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}\big) := \inf_{\pi \in \Pi\left(Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}, P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}\right)} \int \|\tilde{\boldsymbol{c}} - \tilde{\boldsymbol{c}}'\| \mathrm{d}\pi(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{c}}').$$

Here, $\Pi\left(Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}, P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}\right)$ denotes the set of couplings π whose marginals are $Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}$, $P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}$, and $\|\cdot\|$ denotes a norm. The integral $\int \|\tilde{\boldsymbol{c}} - \tilde{\boldsymbol{c}}'\| \, \mathrm{d}\pi(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{c}}')$ is the expected transportation cost under π.

Using this distance, we model uncertainty in the conditional distribution of the coefficient vector via a 1-Wasserstein ambiguity set:

$$\mathbb{B}_\varepsilon(P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}) := \Big\{ Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i} \;\Big|\; W_1\big(Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}, P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}_i}\big) \leq \varepsilon \Big\}.$$

We then aim to control the worst-case expected regret over this set. For a fixed covariate value $\boldsymbol{x} = \boldsymbol{x}_i$, we define the distributionally robust regret loss as

$$\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}) := \sup_{Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}} \in \mathbb{B}_\varepsilon(P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}})} \mathbb{E}_{\tilde{\boldsymbol{c}} \sim Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}}} \big[\mathrm{Reg}(\hat{\boldsymbol{c}}, \tilde{\boldsymbol{c}})\big].$$

Finally, the empirical regret loss minimized is

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}_i).$$

In what follows, since the empirical regret loss is an average over samples, we fix an arbitrary $\boldsymbol{x}_i$ and denote it by $\boldsymbol{x}$ for simplicity. Because $\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x})$ is an infinite-dimensional optimization problem over probability distributions, directly computing its optimizer is generally intractable. We therefore reduce it into a finite-dimensional optimization problem that is tractable for commercial solvers. Define the regret evaluation function for a fixed prediction $\hat{\boldsymbol{c}}$ as:

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}) := \mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c}).$$

Using the strong duality result of Blanchet and Murthy [2019], we show that the distributionally robust regret loss can be written as the sum of the empirical loss and a regularization term, as stated in the following theorem. (See Appendix D.3 for the proof.)

Theorem 3.2. *Let $\mathcal{W}$ be a nonempty compact set. Then the distributionally robust regret loss $\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x})$ over the 1-Wasserstein ambiguity set $\mathbb{B}_\varepsilon(P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}})$ defined with a norm $\|\cdot\|$ admits the following representation:*

$$\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}) = \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}}} \big[\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}})\big] + \varepsilon \, \mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big). \tag{9}$$

Here, $\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big)$ denotes the Lipschitz constant of $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ and is given by $\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big) := \sup_{\boldsymbol{w} \in \mathcal{W}} \big\|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\big\|_*$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

This theorem provides an intuitive interpretation of the proposed distributionally robust regret. Minimizing the worst-case regret over the Wasserstein ambiguity set is equivalent to minimizing a weighted sum of the expected regret under the empirical distribution induced by the available data and a regularization term that encourages a smaller Lipschitz constant $\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$. Here, $\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$ can be interpreted as a sensitivity measure that quantifies how much the regret may change in response to small perturbations of the coefficient vector $\boldsymbol{c}$. If $\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$ is large, even a mild distribution shift can lead to a rapid increase in regret. Conversely, keeping this value small promotes stability of regret under distributional changes. Therefore, the proposed method can be viewed as learning a balance between decision quality on the observed data and robustness to distribution shifts.

During training, we evaluate the distributionally robust loss for each sample $\boldsymbol{x}_i$ and minimize $\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}_i)$. Using Theorem 3.2, the resulting empirical loss can be written as

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ \mathrm{Reg}(\hat{\boldsymbol{c}}(\boldsymbol{x}_i), \boldsymbol{c}_i) + \varepsilon \, \mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}}(\boldsymbol{x}_i))\big) \right\}.$$

In other words, the empirical loss decomposes into the standard empirical regret term plus the expected Lipschitz regularizer. As a result, we can incorporate distributional robustness without explicitly solving for a worst-case distribution.

3.2.2 Gradient Calculation of Distributionally Robust Loss

To compute gradients for the proposed distributionally robust regret loss $\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x})$, we use the decomposition from the previous section

$$\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}) = \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\hat{\boldsymbol{c}}|\boldsymbol{x}}}\big[\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}})\big] + \varepsilon\,\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big),$$

and differentiate the two terms separately. The expectation term coincides with the standard DFL objective, so existing surrogate gradients such as SPO+ can be applied directly.

For the second term, $\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big)$ can be written as

$$\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big) = \max_{(\boldsymbol{w},\boldsymbol{g})\in\mathcal{W}\times\mathcal{G}} \boldsymbol{g}^\top\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\big),$$

where $\mathcal{G}$ denotes the unit ball associated with the dual norm. Hence, its (sub)gradient can be obtained by applying Danskin's theorem. See Appendix C.2 for the detailed derivation and the resulting learning algorithm.

4 THEORETICAL GUARANTEES

We provide generalization guarantees for the proposed robust regret loss and the distributionally robust regret loss. Let $\mathcal{P}$ be the joint distribution over $\mathcal{X}\times\mathbb{R}^n$. In particular, we show that these losses are uniformly bounded and derive a uniform generalization error bound via Rademacher complexity. (see Appendix B.2.)

Theorem 4.1 (Uniform Generalization Error Bound)**.** *Let $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta}\in\Theta\}$ be a class of predictive models. Define the loss class $\mathcal{G}_{\mathrm{rob}}$ associated with the robust regret loss as $\mathcal{G}_{\mathrm{rob}} := \{g_{\boldsymbol{\theta}} : (\boldsymbol{x},\boldsymbol{c}_{\mathrm{obs}}) \mapsto \mathcal{L}_{\mathrm{rob}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{c}_{\mathrm{obs}}) \mid f_{\boldsymbol{\theta}}\in\mathcal{F}\}$. Under Assumption B.3, for any $\delta\in(0,1)$, with probability at least $1-\delta$ on a dataset $\mathcal{D}$, the following generalization bound holds uniformly for all $\boldsymbol{\theta}\in\Theta$:*

$$R(\boldsymbol{\theta}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{\theta}) \le 2\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}_{\mathrm{rob}}) + 3B\sqrt{\frac{\log(2/\delta)}{2N}}, \quad (10)$$

where $R(\boldsymbol{\theta})$ and $\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})$ are the expected risk (with respect to $\mathcal{P}$) and empirical risk defined with respect to the loss $\mathcal{L}_{\mathrm{rob}}$, $\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}_{\mathrm{rob}})$ is the empirical Rademacher complexity of $\mathcal{G}_{\mathrm{rob}}$, and B is a uniform upper bound on the loss.

An analogous bound holds for the distributionally robust regret loss $\mathcal{L}_{\mathrm{dro}}$ by defining the corresponding loss class $\mathcal{G}_{\mathrm{dro}}$. (see Appendix C.1.)

5 NUMERICAL EXPERIMENTS

In this section, we evaluate the proposed method through experiments on synthetic and real-world datasets. All experiments are fully reproducible, and our implementation is available at `https://github.com/isct-nakatalab/UAI2026_Robust_Decision-Focused_Learning_via_Worst-Case_Regret_Minimization`.

5.1 EXPERIMENTAL SETTINGS

Optimization Problems and Datasets In our experiments, we considered three optimization problems: Shortest Path (SP), Traveling Salesman Problem (TSP) and Knapsack Problem (KP). For the synthetic-data experiments, we used PyEPO [Tang and Khalil, 2024], a DFL library. Specifically, we generated coefficient vectors $\boldsymbol{c}_i$ via a nonlinear transformation of covariates $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, I_p)$. See Appendix F.2 and E.1 for the detailed problem formulations and data-generation procedures. For the real-world experiments, we used the energy scheduling problem (EN) constructed from public datasets [Meinecke et al., 2020, Panadero, 2024, Lamprecht, 2026], following the experimental setup in PyEPO. See Appendix E.1.4 for the detailed formulation of the energy scheduling problem.

Predictive Models and Learning Algorithms For the predictive model $f_{\boldsymbol{\theta}}$, we used a linear model in the synthetic-data experiments. For the real-world experiments, we used a multilayer perceptron with three fully connected layers (hidden dimensions: 64 and 32) and ReLU activations. We employ the technique from Elmachtoub and Grigas [2022] to compute a surrogate for the gradient of the proposed losses. Detailed hyperparameter settings and the data split are provided in Appendix F.1.

Evaluation Metrics At test time, for each instance i, we compute the optimal solution under the true coefficient vector $\boldsymbol{w}^*(\boldsymbol{c}_i)$, and the optimal solution induced by the prediction $\boldsymbol{w}^*(\hat{\boldsymbol{c}}_i)$, where $\hat{\boldsymbol{c}}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$. Following Section 3.1, we define the regret as $r_i := \mathrm{Reg}(\hat{\boldsymbol{c}}_i, \boldsymbol{c}_i)$. To account for scale differences, we report the relative regret $\tilde{r}_i := \frac{r_i}{|z(\boldsymbol{c}_i)|}$ and use the mean relative regret over the test set as the primary evaluation metric. To assess robustness, we additionally report the Conditional Value-at-Risk (CVaR_α) at confidence level $\alpha = 0.9$.

Comparison Methods The methods compared in this experiment are summarized as follows. For clarity, we denote the method names corresponding to our proposed losses below:

- `reg`: The robust regret loss $\mathcal{L}_{\mathrm{rob}}$, minimizing worst-case regret over an uncertainty set.
- `ropt`: The approximate robust regret loss $\widetilde{\mathcal{L}}_{\mathrm{rob}}$, a computationally efficient surrogate.
- `droreg`: The distributionally robust regret loss $\mathcal{L}_{\mathrm{dro}}$, minimizing worst-case expected regret over a Wasserstein ambiguity set.

We append `_budget` or `_ellip` to indicate the use of budgeted or ellipsoidal uncertainty sets, respectively. For `droreg`, we specify the ground norm ℓ_p (e.g., `droreg-l2`).

Baselines include:

- `twostage`: A predictive model trained by minimizing mean squared error (MSE).

- `emp`: A predictive model trained with the standard regret loss (3).
- `knn`: A predictive model trained with the local average regret loss based on k-nearest neighbors [Schutte et al., 2024]. We set $k = 5$.
- `gendfl`: A predictive model trained with the regret loss augmented by a generative model [Wang et al., 2026]. Following the original paper, we train the generative model using conditional normalizing flows [Winkler et al., 2019] and set $\alpha = 0.9$ for CVaR.

5.1.1 Uncertainty and Perturbation Settings

In this experiment, we use the following perturbation settings to mimic two types of uncertainty: observation errors and distribution shifts.

Observation Error Settings To artificially introduce observation errors into the training data, for each training instance i we generate a perturbed observed coefficient vector $\boldsymbol{c}_{\mathrm{obs},i}$ from the true coefficient vector $\boldsymbol{c}_i$ and use it for training. We consider the following two perturbation types.

- **Ellipsoidal Perturbation (ell)**: For each instance i, we set the baseline standard deviation as $\sigma_i^{(\mathrm{base})} := \|\boldsymbol{c}_i\|_2/\sqrt{n}$. We then draw Gaussian noise $\boldsymbol{g}_i \sim \mathcal{N}(\mathbf{0}, (\sigma_i^{(\mathrm{base})})^2 I_n)$ and define the observed coefficient vector as $\boldsymbol{c}_{\mathrm{obs},i} = \boldsymbol{c}_i + \boldsymbol{g}_i$.
- **Sparse Perturbation (sparse)**: Sparse perturbations selectively introduce large shifts in a subset of coordinates. Specifically, we add a fixed shift of width $\delta \geq 0$ to the randomly selected $\lfloor \gamma_{\mathrm{frac}} n \rfloor$ coordinates, where $\gamma_{\mathrm{frac}} \in (0, 1]$ controls the fraction of perturbed dimensions. See Appendix F.1 for details. For each instance i, let $J_i \subset \{1, \ldots, n\}$ be the indices of the selected coordinates and $s \in \{+1, -1\}$ denote the problem direction, where $s = +1$ for the maximization problems and $s = -1$ for the minimization problems. Then we set $c_{\mathrm{obs},ij} = c_{ij} + s\,\delta\,\mathbf{1}\{j \in J_i\}$.

Distribution Shift Settings To evaluate robustness to distribution shifts, we artificially alter the conditional distribution of coefficient vectors in the training data. Concretely, we use a mixture perturbation model that combines two types of shifts: covariate-dependent bias, where a polynomial map of the covariates is added to the coefficients to simulate systematic shifts and global noise inflation, where covariate-independent Gaussian noise is injected to simulate increased environmental variance. Detailed mathematical formulations and parameter values are provided in Appendix F.1.

5.2 EXPERIMENTAL RESULTS AND DISCUSSION

5.2.1 Robustness against Observation Errors

Figure 1 (a) reports the mean relative regret for each optimization problem and perturbation setting on the test set.

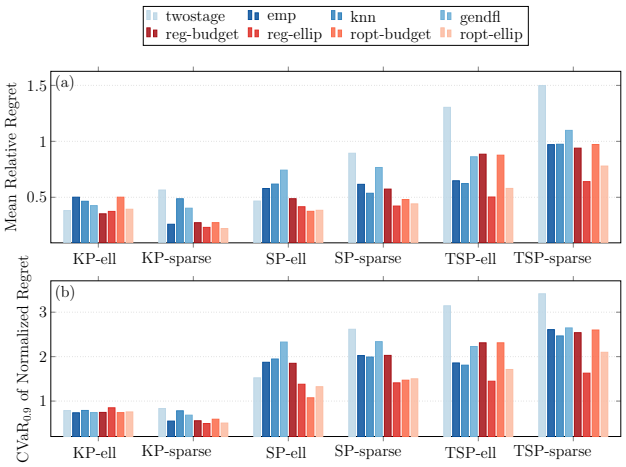

Figure 1: Mean relative regret (a) and $\mathrm{CVaR}_{0.9}$ (b) in the presence of observation errors.

Since `twostage` directly minimizes MSE, it does not show consistent improvements over `emp`, which is trained with a regret-based loss; in particular, it performs substantially worse on TSP. We also observe that `gendfl` yields a relatively large mean relative regret overall. In the presence of observation errors in the training data, the generative model learns a distorted conditional distribution, causing tail-focused objectives (e.g., CVaR) to target the wrong tail and leading to evaluation under a misspecified distribution. In contrast, the proposed methods (`reg` and `ropt`) achieve lower regret than `emp` and `knn` in many settings. Notably, `reg-ellip` provides a marked improvement on TSP, highlighting the benefit of optimizing a loss that accounts for worst-case outcomes. Moreover, even under mismatched uncertainty settings, where the ellipsoidal-set models (`reg-ellip`/`ropt-ellip`) are evaluated under structurally different sparse perturbations, they still tend to outperform `emp` and `knn`. This suggests that a model trained with the ellipsoidal set can remain robust even when the perturbations encountered at evaluation differ from those assumed during training.

Figure 1 (b) shows the resulting $\mathrm{CVaR}_{0.9}$ values. Beyond improving the average regret, the proposed methods reduce $\mathrm{CVaR}_{0.9}$ by suppressing a small number of extreme regret instances. This reduction in tail risk also contributes to the lower mean relative regret. Conversely, methods without explicit worst-case robustness, such as `knn`, exhibit higher $\mathrm{CVaR}_{0.9}$, which is consistent with their increased mean relative regret.

5.2.2 Robustness against Distribution Shifts

Figure 2 compares the mean relative regret and $\mathrm{CVaR}_{0.9}$ for each optimization problem under distribution shift. Overall, `droreg` exhibits more stable performance than the baselines in terms of both mean relative regret and $\mathrm{CVaR}_{0.9}$. This is consistent with the fact that `twostage`, which minimizes only prediction error, is sensitive to distributional

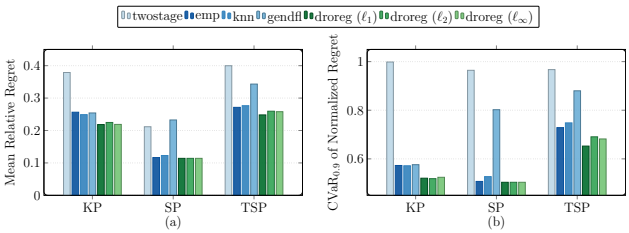

Figure 2: Comparison of mean relative regret (a) and $\text{CVaR}_{0.9}$ (b) for each optimization problem under distribution shift.

changes, while `emp` can degrade out-of-distribution due to overfitting to the training distribution. Similarly, `gendfl` deteriorates when the generative model fails to capture the shifted distribution. In contrast, the proposed distributionally robust approach yields substantial gains on KP and TSP; notably, using the ℓ_1 leads to particularly strong robustness on TSP. We note that the synthetic shifts were generated by a specific mechanism that was not assumed during training (Appendix F.1). Even so, `droreg` still outperformed the baselines, indicating that it remains robust even when the deployment shifts are not the ones assumed during training.

5.2.3 Sensitivity Analysis

We show the results of sensitivity analyses on the uncertainty-set hyperparameters, using `reg-ellip` on TSP under sparse perturbation as a representative case (Figure 3). Both the mean and $\text{CVaR}_{0.9}$ of the relative regret are lowest at an intermediate κ and larger when κ is too small or too large. Across most of the range, the proposed method outperforms the `emp` and `knn` baselines in both metrics, matching them only at the smallest κ. We also show the results for `droreg` by varying the parameter ε for the Wasserstein ambiguity set with several ground norms (ℓ_1, ℓ_2, ℓ_∞) on KP (Figure 4). Both the mean and $\text{CVaR}_{0.9}$ of the relative regret are lowest at an intermediate ε and larger when ε is too small or too large, especially under the ℓ_2 and ℓ_∞ norms, whereas the ℓ_1 norm is the most stable across ε. With an appropriately chosen ε, `droreg` outperforms the baselines, and the best ground norm depends on the problem structure. Comprehensive results of the sensitivity analysis are shown in Appendix G (Figure 9).

5.2.4 Comparison of Computation Time

We investigate how much faster `ropt` is compared to other methods to verify its computational efficiency. Specifically, we compare the computation times of the existing methods (`emp`, `knn`) and the proposed methods (`ropt`, `reg`) for robust regret learning against observation errors. Figure 5 shows the distribution of per-epoch training time (the vertical axis is on a logarithmic scale). Both `emp` and `knn` have low computational cost because they only require solving the downstream optimization problem for the predicted coefficient vector. In contrast, `reg` computes the worst-case

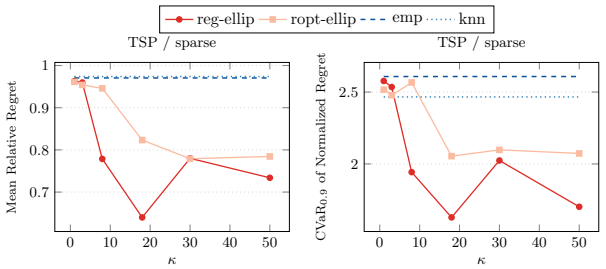

Figure 3: Sensitivity of mean relative regret (left) and $\text{CVaR}_{0.9}$ (right) for `reg-ellip` (`ropt-ellip` for reference) to the parameter κ on TSP under sparse perturbation, with the `emp` (dashed) and `knn` (dotted) baselines.

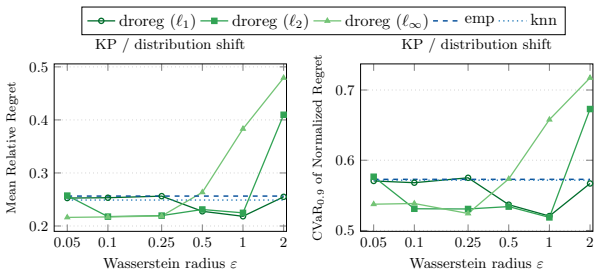

Figure 4: Sensitivity of mean relative regret (left) and $\text{CVaR}_{0.9}$ (right) for `droreg` with $\ell_1, \ell_2, \ell_\infty$ ground norms to the parameter ε on KP under distribution shift, with the `emp` (dashed) and `knn` (dotted) baselines.

regret exactly, incurring the longest training time; `ropt`, in turn, takes more time than `emp`/`knn` but far less than `reg`, while still improving robustness. Overall, `ropt` provides a practical trade-off, enhancing robustness within feasible computation time.

5.2.5 Verification using Real-World Data

The results on the real-world energy scheduling problem (EN) are shown in Figure 6 (observation errors) and Figure 7 (distribution shift).

Under Observation Error In the presence of observation errors, the proposed methods consistently outperform the baselines in both mean relative regret and $\text{CVaR}_{0.9}$. This indicates that the benefits of our loss functions observed on synthetic data also carry over to a more realistic setting based on public datasets. Notably, the improvement in $\text{CVaR}_{0.9}$ suggests that the proposed methods not only reduce regret but also suppress rare yet severe regret events.

Under Distribution Shift Even under distribution shift, the proposed methods tend to keep both mean relative regret and $\text{CVaR}_{0.9}$ low. In practice, electricity prices and demand can shift due to seasonality, system and policy changes, or evolving market conditions, making risk reduction under distribution shift particularly important. Overall, real-world experiments demonstrate that the proposed methods improve solution robustness in terms of both regret and CVaR

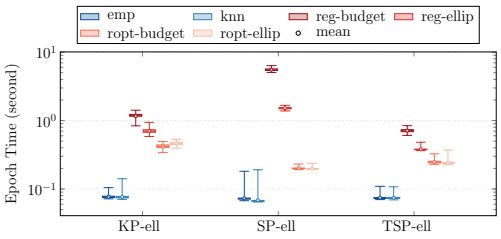

Figure 5: Comparison of computation time using the proposed robust regret.

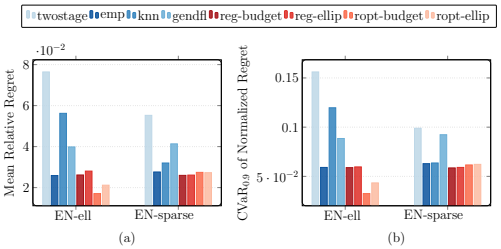

Figure 6: Comparison of mean relative regret (a) and $\text{CVaR}_{0.9}$ (b) under observation errors in real-world data.

against two distinct sources of uncertainty.

6 CONCLUSION

We proposed robust DFL methods that address (i) observation errors in training coefficient vectors and (ii) distribution shifts at deployment. For Uncertainty (i), we define an uncertainty set for observation errors and minimize the worst-case regret over the set. For Uncertainty (ii), we adopt a Wasserstein ambiguity set and optimize a distributionally robust regret. Experiments show that our method improves regret and risk measures over existing baselines. Future work includes automating the selection of uncertainty-set hyperparameters.

Limitations and Future Work Several limitations remain. First, using the approximate loss $\widetilde{\mathcal{L}}_{\text{rob}}$ (`ropt`) in place of the exact loss $\mathcal{L}_{\text{rob}}$ is a heuristic without formal worst-case guarantees. Although it performed comparably to the exact loss at much lower computation cost, establishing such guarantees remains an open problem. Second, our real-world validation relies on a single energy-scheduling dataset, and our synthetic shifts may not fully capture the shifts arising in practice, so broader evaluation is needed. Third, automating the data-driven selection of the parameters of the uncertainty and ambiguity sets is an important direction for future work.

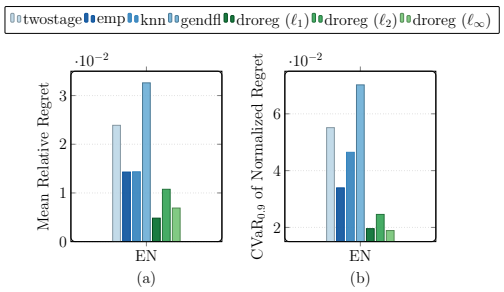

Figure 7: Comparison of mean relative regret (a) and $\text{CVaR}_{0.9}$ (b) under distribution shifts in real-world data.

References

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 136–145, 2017.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009. doi: 10.1515/9781400831050.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, 2021. doi: 10.1016/j.ejor.2020.07.063.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2 edition, 1999. ISBN 1886529000.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3): 1025–1044, 2020. doi: 10.1287/mnsc.2018.3253.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004. doi: 10.1287/opre.1030.0065.

Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. doi: 10.1137/080734510.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. doi: 10.1287/moor.2018.0936.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. URL https://jmlr.org/papers/v21/19-021.html.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3): 595–612, 2010. doi: 10.1287/opre.1090.0741.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5490–5500, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Othman El Balghiti, Adam N. Elmachtoub, Paul Grigas, and Ambuj Tewari. Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research*, 48(4):2043–2065, 2023. doi: 10.1287/moor.2022.1330. URL https://doi.org/10.1287/moor.2022.1330.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022. doi: 10.1287/mnsc.2020.3922. URL https://doi.org/10.1287/mnsc.2020.3922.

Aaron M. Ferber, Bryan Wilder, Bistra Dilkina, and Milind Tambe. MIPaaL: Mixed integer program as a layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1504–1511, 2020. doi: 10.1609/aaai.v34i02.5509.

Gurobi Optimization, LLC. Gurobi optimizer reference manual. Software documentation, 2025. URL https://docs.gurobi.com/projects/optimizer/en/current/. (2026-01-21).

Frederick S. Hillier and Gerald J. Lieberman. *Introduction to Operations Research*. McGraw-Hill Education, 10 edition, 2015.

Michael Huang and Vishal Gupta. Decision-focused learning with directional gradients. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jakob Huber, Sebastian Müller, Moritz Fleischmann, and Heiner Stuckenschmidt. A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, 278(3):904–915, 2019. doi: 10.1016/j.ejor.2019.04.043.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. End-to-end constrained optimization learning: A survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 4475–4482, 2021. doi: 10.24963/ijcai.2021/610.

Christian Sebastian Lamprecht. meteostat: Access and analyze historical weather and climate data with python. Software (Python package / PyPI), 2026. URL https://pypi.org/project/meteostat/. (2026-01-21).

Xutao Ma, Chao Ning, and Wenli Du. Differentiable distributionally robust optimization layers. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 2024. doi: 10.1613/jair.1.15320.

Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188, 1989.

Steffen Meinecke, Džanan Sarajlić, Simon Ruben Drauz, Annika Klettke, Lars-Peter Lauven, Christian Rehtanz, Albert Moser, and Martin Braun. Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. *Energies*, 13(12):3290, 2020. doi: 10.3390/en13123290.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3462–3471, 2018.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018. doi: 10.1007/s10107-017-1172-1.

Eugenio Panadero. aiopvpc: Simple aio library to download spanish electricity hourly prices. GitHub repository, 2024. URL https://github.com/azogue/aiopvpc. (2026-01-21).

Ignacio Rios, Roger J.-B. Wets, and David L. Woodruff. Multi-period forecasting and scenario generation with limited data. *Computational Management Science*, 12(2): 267–295, 2015. doi: 10.1007/s10287-015-0230-5.

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. doi: 10.1016/j.ejor.2024.03.020.

Noah J. Schutte, Krzysztof S. Postek, and Neil Yorke-Smith. Robust losses for decision-focused learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 4868–4875, 2024. doi: 10.24963/ijcai.2024/538.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-focused learning without decision-making: Learning locally optimized decision losses. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://dblp.org/rec/conf/nips/Shah0WPT22.

Bo Tang and Elias B. Khalil. PyEPO: a PyTorch-based end-to-end predict-then-optimize library for linear and integer programming. *Mathematical Programming Computation*, July 2024. ISSN 1867-2957. doi: 10.1007/s12532-024-00255-x.

Marin Vlastelica, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolínek. Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*, 2020.

Prince Zizhuang Wang, Shuyi Chen, Jinhao Liang, Ferdinando Fioretto, and Shixiang Zhu. Gen-DFL: Decision-focused generative learning for robust decision making. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GU2197a3Lm.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014. doi: 10.1287/opre.2014.1314.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint*, 2019. URL https://arxiv.org/abs/1912.00042.

Robust Decision-Focused Learning via Worst-Case Regret Minimization (Supplementary Material)

Shoki Yamao[1] **Ken Kobayashi**[1] **Ryo Matsui**[2] **Shota Nagai**[1] **Naoki Nishimura**[2] **Kazuhide Nakata**[1]

[1]Institute of Science Tokyo, Tokyo, Japan
[2]Recruit Co., Ltd., Chiyoda, Tokyo, Japan,

This Supplementary Material should be submitted together with the main paper.

A MATHEMATICAL PRELIMINARIES

In this section, we summarize external results used in this paper to clarify the theoretical background.

Proposition A.1 (Danskin's Theorem; [Bertsekas, 1999, Proposition B.25])**.** *Let Y be a non-empty compact set, and let the function $\psi : \mathbb{R}^m \times Y \to \mathbb{R}$ be continuous, and for each $y \in Y$, $\psi(\cdot, y)$ be convex and of class C^1. Define the optimal value function $f(x) := \max_{y \in Y} \psi(x, y)$ and the set of optimal solutions $Y(x) := \arg\max_{y \in Y} \psi(x, y)$. In this case, if $\nabla_x \psi(x, \cdot)$ is continuous on Y, the subgradient of $f(x)$ is given by:*

$$\partial f(x) = \mathrm{conv}\{\nabla_x \psi(x, y) \mid y \in Y(x)\}.$$

Here, conv *denotes the convex hull. In particular, if the optimal solution is unique, the gradient is $\nabla f(x) = \nabla_x \psi(x, y^*)$.*

Theorem A.2 (Blanchet–Murthy's Strong Duality [Blanchet and Murthy, 2019, Theorem 1])**.** *Consider a distance $d(\boldsymbol{u}, \boldsymbol{v}) := \|\boldsymbol{u} - \boldsymbol{v}\|$ induced by a norm $\|\cdot\|$.*

Let μ be a probability measure on $\mathbb{R}^n$, and write $\boldsymbol{C} \sim \mu$. Assume that the evaluation function $f : \mathbb{R}^n \to \mathbb{R}$ is

- *absolutely integrable with respect to μ,*
- *upper semicontinuous on $\mathbb{R}^n$.*

In this case, for any $\varepsilon > 0$, the worst-case expected value on the uncertainty set with respect to W_1

$$\mathbb{B}_\varepsilon(\mu) := \left\{ Q \;\middle|\; W_1(Q, \mu) \leq \varepsilon \right\}$$

is expressed as the following univariate dual problem:

$$\sup_{Q \in \mathbb{B}_\varepsilon(\mu)} \mathbb{E}_Q[f(\boldsymbol{C})] = \inf_{\lambda \geq 0} \Big\{ \lambda\varepsilon + \mathbb{E}_\mu \big[\sup_{\boldsymbol{y} \in \mathbb{R}^n} \{ f(\boldsymbol{y}) - \lambda \|\boldsymbol{y} - \boldsymbol{C}\| \} \big] \Big\}. \tag{11}$$

The minimum value on the right-hand side is finite and equal to the maximum value on the left-hand side.

B DETAILS OF ROBUST REGRET LOSS

B.1 DETAILED FORMULATION OF ROBUST REGRET LOSS

In this section, we describe the detailed formulation of the robust regret loss $\mathcal{L}_{\mathrm{rob}}$.

As concrete shapes of the uncertainty set $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$, we adopt the following two sets used in robust optimization.

1. Ellipsoidal Uncertainty Set [Ben-Tal et al., 2009]: It is a set defined based on the Mahalanobis distance, considering the correlation between coefficient vectors.

$$\mathcal{C}_{\mathrm{ell}}(\boldsymbol{c}_{\mathrm{obs}}) = \{\, \boldsymbol{c} \in \mathbb{R}^n \mid (\boldsymbol{c} - \boldsymbol{c}_{\mathrm{obs}})^\top \Sigma^{-1} (\boldsymbol{c} - \boldsymbol{c}_{\mathrm{obs}}) \leq \kappa^2 \,\}.$$

   Here, Σ is the covariance matrix of the error, and κ is a parameter that controls the magnitude of uncertainty.

2. Budgeted Uncertainty Set [Bertsimas and Sim, 2004]: It is an uncertainty set where each component can fluctuate independently, but the total amount (budget) of components having large errors at the same time is limited. Specifically,

$$\mathcal{C}_{\Gamma}(\boldsymbol{c}_{\mathrm{obs}}) = \{\boldsymbol{c} \in \mathbb{R}^n \mid \boldsymbol{c} = \boldsymbol{c}_{\mathrm{obs}} + \mathrm{Diag}(\boldsymbol{\rho})\boldsymbol{\zeta},\ \|\boldsymbol{\zeta}\|_\infty \leq 1,\ \|\boldsymbol{\zeta}\|_1 \leq \Gamma\}\,.$$

   Here, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)^\top$ is a vector in which the fluctuation range of each component is arranged, and $\mathrm{Diag}(\boldsymbol{\rho}) \in \mathbb{R}^{n \times n}$ represents a diagonal matrix with diagonal elements $\rho_1, \ldots, \rho_n$ and 0 otherwise. Also, Γ represents the budget of uncertainty.

First, defining the support function on the uncertainty set as $h_{\mathcal{C}}(\boldsymbol{d}) := \sup_{\boldsymbol{c} \in \mathcal{C}} \boldsymbol{c}^\top \boldsymbol{d}$, the robust regret loss can be transformed as follows:

$$\begin{aligned}
\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) &= \max_{\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \left\{ \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \min_{\boldsymbol{w} \in \mathcal{W}} \boldsymbol{c}^\top \boldsymbol{w} \right\} && (12)\\
&= \max_{\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \max_{\boldsymbol{w} \in \mathcal{W}} \boldsymbol{c}^\top (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) && (13)\\
&= \max_{\boldsymbol{w} \in \mathcal{W}} \max_{\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \boldsymbol{c}^\top (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) && (14)\\
&= \max_{\boldsymbol{w} \in \mathcal{W}} h_{\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}). && (15)
\end{aligned}$$

In the following, we show that the solution of the inner optimization by the uncertainty set can be analytically calculated when using the ellipsoidal uncertainty set and the budgeted uncertainty set for Equation (15), and formulate it as an optimization problem.

B.1.1 Formulation with Ellipsoidal Uncertainty Set

We describe the details when using the ellipsoidal uncertainty set $\mathcal{C}_{\mathrm{ell}}(\boldsymbol{c}_{\mathrm{obs}})$. The support function on this set (the value maximizing the linear function $\boldsymbol{c}^\top \boldsymbol{d}$ on the set) can be analytically obtained as follows [Ben-Tal et al., 2009].

$$h_{\mathcal{C}_{\mathrm{ell}}(\boldsymbol{c}_{\mathrm{obs}})}(\boldsymbol{d}) = \sup_{\boldsymbol{c} \in \mathcal{C}_{\mathrm{ell}}(\boldsymbol{c}_{\mathrm{obs}})} \boldsymbol{c}^\top \boldsymbol{d} = \boldsymbol{c}_{\mathrm{obs}}^\top \boldsymbol{d} + \kappa \, \|\Sigma^{1/2} \boldsymbol{d}\|_2.$$

By substituting this into Equation (15), the calculation of the worst regret becomes the following maximization problem:

$$\mathcal{L}_{\mathrm{rob}} = \max_{\boldsymbol{w} \in \mathcal{W}} \left\{ \boldsymbol{c}_{\mathrm{obs}}^\top \big(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\big) + \kappa \left\| \Sigma^{1/2} \big(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\big) \right\|_2 \right\}. \tag{16}$$

Furthermore, by introducing an auxiliary variable $t \geq 0$, it can be transformed into the following formally equivalent problem:

$$\begin{aligned}
\max_{\boldsymbol{w} \in \mathcal{W},\ t \geq 0} \quad & \boldsymbol{c}_{\mathrm{obs}}^\top (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) + \kappa \, t \\
\text{s.t.} \quad & t^2 \leq (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w})^\top \Sigma \, (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}).
\end{aligned} \tag{17}$$

Although this problem is a non-convex Quadratically Constrained Quadratic Programming (QCQP), it is possible to find a globally optimal solution using mathematical optimization solvers such as Gurobi Optimizer [Gurobi Optimization, LLC, 2025].

B.1.2 Formulation with Budgeted Uncertainty Set

Next, we describe the details when using the budgeted uncertainty set $\mathcal{C}_{\Gamma}(\boldsymbol{c}_{\mathrm{obs}})$. The support function in this case can be formulated as follows.

$$h_{\mathcal{C}_{\Gamma}(\boldsymbol{c}_{\mathrm{obs}})}(\boldsymbol{d}) = \sup_{\boldsymbol{c} \in \mathcal{C}_{\Gamma}(\boldsymbol{c}_{\mathrm{obs}})} \boldsymbol{c}^\top \boldsymbol{d} = \boldsymbol{c}_{\mathrm{obs}}^\top \boldsymbol{d} + \max_{\substack{|\zeta_j| \leq 1 \\ \sum_{j=1}^n |\zeta_j| \leq \Gamma}} \sum_{j=1}^n \rho_j \, \zeta_j \, d_j.$$

Here, if we transform the variables as $\zeta_j d_j = \mathrm{sgn}(d_j) s_j$ (where $s_j \in [0,1]$), the maximization problem of the second term becomes

$$\max_{\substack{0 \le s_j \le 1 \\ \sum_{j=1}^n s_j \le \Gamma}} \sum_{j=1}^n \rho_j \, |d_j| \, s_j \tag{18}$$

and the optimal solution is obtained by selecting s_j in descending order of weight $\rho_j |d_j|$. Specifically, the support function is expressed as

$$h_{\mathcal{C}_\Gamma(\boldsymbol{c}_{\mathrm{obs}})}(\boldsymbol{d}) = \boldsymbol{c}_{\mathrm{obs}}^\top \boldsymbol{d} + \sum_{k=1}^{\lfloor \Gamma \rfloor} \rho_{(k)} \, |d_{(k)}| + (\Gamma - \lfloor \Gamma \rfloor) \, \rho_{(\lfloor \Gamma \rfloor + 1)} \, |d_{(\lfloor \Gamma \rfloor + 1)}|.$$

Here, $\rho_{(k)}|d_{(k)}|$ represents the k-th value when $\{\rho_j |d_j|\}_{j=1}^n$ is sorted in descending order. In particular, when Γ is an integer, it simplifies to

$$h_{\mathcal{C}_\Gamma(\boldsymbol{c}_{\mathrm{obs}})}(\boldsymbol{d}) = \boldsymbol{c}_{\mathrm{obs}}^\top \boldsymbol{d} + \sum_{k=1}^{\Gamma} \rho_{(k)} \, |d_{(k)}|.$$

By substituting $\boldsymbol{d} = \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}$, the robust loss function can be written as

$$\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) = \max_{\boldsymbol{w} \in \mathcal{W}} \left\{ \boldsymbol{c}_{\mathrm{obs}}^\top (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) + \sum_{k=1}^{\Gamma} \rho_{(k)} \, |w^*(\hat{\boldsymbol{c}})_k - w_k| \right\}$$

.

Furthermore, when $\mathcal{W}$ is a polyhedral set and Γ is an integer, this maximization problem can be formulated as a Mixed Integer Linear Programming (MILP). We introduce a variable $u_j \ge |w^*(\hat{\boldsymbol{c}})_j - w_j|$ representing the absolute value of each component, and a $0-1$ variable $y_j \in \{0,1\}$ representing whether it is included in the top Γ. Furthermore, by using a variable t_j to linearize the product term, it can be reduced to the following MILP as a whole:

$$u_j \ge w^*(\hat{\boldsymbol{c}})_j - w_j, \qquad u_j \ge -(w^*(\hat{\boldsymbol{c}})_j - w_j), \qquad u_j \ge 0 \qquad (j = 1, \dots, n).$$

We introduce a $0-1$ variable $y_j \in \{0,1\}$ to represent the top Γ selection, and set $\sum_{j=1}^n y_j \le \Gamma$. To linearize the product $\rho_j \, u_j \, y_j$ appearing in the objective function, we introduce a new variable $t_j = \rho_j \, u_j \, y_j$. Summarizing the above, the outer maximization reduces to the following MILP:

$$\max_{\substack{\boldsymbol{w} \in \mathcal{W},\, \boldsymbol{u} \ge 0, \\ \boldsymbol{y} \in \{0,1\}^n,\, \boldsymbol{t} \ge 0}} \quad \boldsymbol{c}_{\mathrm{obs}}^\top (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) + \sum_{j=1}^n t_j \tag{19}$$

$$\text{s.t.} \quad u_j \ge w^*(\hat{\boldsymbol{c}})_j - w_j, \qquad u_j \ge -(w^*(\hat{\boldsymbol{c}})_j - w_j) \qquad (j = 1, \dots, n), \tag{20}$$

$$0 \le u_j \le M_j \qquad (j = 1, \dots, n), \tag{21}$$

$$\sum_{j=1}^n y_j \le \Gamma, \qquad y_j \in \{0,1\} \qquad (j = 1, \dots, n), \tag{22}$$

$$0 \le t_j \le \rho_j \, u_j, \quad t_j \le \rho_j \, M_j \, y_j, \quad t_j \ge \rho_j \, u_j - \rho_j \, M_j \, (1 - y_j) \quad (j = 1, \dots, n). \tag{23}$$

The objective function is $\boldsymbol{c}_{\mathrm{obs}}^\top (\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) + \sum_{j=1}^n t_j$, and since t_j matches $\rho_j u_j$ for the selected Γ components, it is equivalent to the sum of the top Γ of the support function. Note that M_j is a sufficiently large constant satisfying $M_j \ge \max_{\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}} |w_j - w'_j|$.

B.1.3 Analytical Solution for Approximate Robust Regret Loss

We describe the solution method for the maximization problem over the uncertainty set of the coefficient vector, which is important in the calculation of the approximate robust regret loss $\widetilde{\mathcal{L}}_{\mathrm{rob}}$:

$$\boldsymbol{c}^\dagger(\hat{\boldsymbol{c}}) \in \operatorname*{arg\,max}_{\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}).$$

This $\boldsymbol{c}^\dagger$ is used to approximate the inner minimization in the regret calculation.

Also, for any set $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$, let $\max_{\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) = h_{\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$.

Case of Budget Uncertainty Set The support function of the budget uncertainty set $\mathcal{C}_\Gamma(\boldsymbol{c}_{\mathrm{obs}})$ can be written as follows, by a similar argument to Equation (18):

$$h_{\mathcal{C}_\Gamma(\boldsymbol{c}_{\mathrm{obs}})}(\boldsymbol{w}) \;=\; \boldsymbol{c}_{\mathrm{obs}}^\top \boldsymbol{w} \;+\; \sum_{k=1}^{\lfloor \Gamma \rfloor} \rho_{(k)} \, |w_{(k)}| \;+\; \left(\Gamma - \lfloor \Gamma \rfloor\right) \rho_{(\lfloor \Gamma \rfloor + 1)} \, |w_{(\lfloor \Gamma \rfloor + 1)}|. \tag{24}$$

Therefore, $\boldsymbol{c}^\dagger$ is obtained by determining the signs such that $\zeta_j = \mathrm{sgn}(w_j)$ for components with large absolute values $|w_j|$ of $\boldsymbol{w} = \boldsymbol{w}^*(\hat{\boldsymbol{c}})$, and assigning weights ρ_j within the range where the total amount does not exceed Γ. Thus, the approximate robust regret loss $\widetilde{\mathcal{L}}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) = h_{\mathcal{C}_\Gamma(\boldsymbol{c}_{\mathrm{obs}})}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big) - z(\boldsymbol{c}^\dagger)$ can be calculated analytically.

Case of Ellipsoidal Uncertainty Set In the case of the ellipsoidal uncertainty set $\mathcal{C}_{\mathrm{ell}}(\boldsymbol{c}_{\mathrm{obs}})$, the support function is $h_{\mathcal{C}_{\mathrm{ell}}}(\boldsymbol{w}) = \boldsymbol{c}_{\mathrm{obs}}^\top \boldsymbol{w} + \kappa \, \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{w}\|_2$, and its maximizing solution $\boldsymbol{c}^\dagger$ can be analytically calculated as

$$\boldsymbol{c}^\dagger \;=\; \boldsymbol{c}_{\mathrm{obs}} \;+\; \kappa \cdot \frac{\boldsymbol{\Sigma} \, \boldsymbol{w}^*(\hat{\boldsymbol{c}})}{\left\| \boldsymbol{\Sigma}^{1/2} \boldsymbol{w}^*(\hat{\boldsymbol{c}}) \right\|_2} \quad \left(\, \|\boldsymbol{\Sigma}^{1/2} \boldsymbol{w}^*(\hat{\boldsymbol{c}})\|_2 > 0 \, \right). \tag{25}$$

Therefore, $\widetilde{\mathcal{L}}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) = \boldsymbol{c}^{\dagger\top} \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}^\dagger)$ can be easily calculated.

B.2 UNIFORM GENERALIZATION OF ROBUST REGRET LOSS

In this section, we show that the robust regret loss defined for observation errors (5) is uniformly bounded from above by the diameter of the feasible set $\mathcal{W}$ and the radius of the uncertainty set $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$, and derive an upper bound on the generalization error.

B.2.1 Uniform Boundedness of Robust Regret Loss

Preliminaries: Definitions of Diameter and Radius Let $\|\cdot\|$ be a norm and $\|\cdot\|_*$ be its dual norm. We define the diameter of the feasible set $\mathcal{W} \subset \mathbb{R}^n$ with respect to the dual norm as

$$\mathrm{diam}_*(\mathcal{W}) := \sup_{\boldsymbol{u}, \boldsymbol{v} \in \mathcal{W}} \|\boldsymbol{u} - \boldsymbol{v}\|_* \in [0, \infty). \tag{26}$$

Also, for each $\boldsymbol{c}_{\mathrm{obs}}$, we define the radius of the uncertainty set $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$ as

$$R_{\mathcal{C}}(\boldsymbol{c}_{\mathrm{obs}}) := \sup_{\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \|\boldsymbol{c}\| \in [0, \infty). \tag{27}$$

Hereinafter, we assume the following.

Assumption B.1 (Compactness)**.** *$\mathcal{W} \subset \mathbb{R}^n$ is non-empty and compact. For each $\boldsymbol{c}_{\mathrm{obs}}$, $\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}}) \subset \mathbb{R}^n$ is non-empty and compact.*

From Assumption B.1, $\mathrm{diam}_*(\mathcal{W}) < \infty$ and $R_{\mathcal{C}}(\boldsymbol{c}_{\mathrm{obs}}) < \infty$ hold.

Upper Bound of Robust Regret Loss

Lemma B.2 (Upper Bound of Robust Regret Loss)**.** *Under Assumption B.1, for any $\hat{\boldsymbol{c}} \in \mathbb{R}^n$ and any $\boldsymbol{c}_{\mathrm{obs}}$,*

$$0 \le \mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) \le R_{\mathcal{C}}(\boldsymbol{c}_{\mathrm{obs}}) \cdot \mathrm{diam}_*(\mathcal{W}) \tag{28}$$

holds.

(Proof in Section D.6)

Uniform Boundedness in Population In the proof of generalization, it is necessary that the loss is almost surely bounded through the random variable $\tilde{\boldsymbol{c}}_{\mathrm{obs}}$, so we assume the following.

Assumption B.3 (Uniform Boundedness)**.** *There exists a constant $R < \infty$ such that $R_{\mathcal{C}}(\tilde{\boldsymbol{c}}_{\mathrm{obs}}) \leq R$ holds almost surely.*

Under Assumption B.3, from Lemma B.2, $0 \leq \mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \tilde{\boldsymbol{c}}_{\mathrm{obs}}) \leq B$ holds almost surely. Here $B := R \cdot \mathrm{diam}_*(\mathcal{W})$.

B.2.2 Generalization Based on Uniform Boundedness

Hereinafter, let $\tilde{\boldsymbol{s}} := (\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{c}}_{\mathrm{obs}}) \sim \mathcal{D}$, and let $S = (\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_N)$ be i.i.d. samples from $\mathcal{D}$. Here $\mathcal{D}$ represents the joint distribution of the covariate $\tilde{\boldsymbol{x}}$ and the observed coefficient vector $\tilde{\boldsymbol{c}}_{\mathrm{obs}}$. Let the model class be $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, and the loss class be defined as

$$\mathcal{G} := \Big\{ g_{\boldsymbol{\theta}} : \ (\boldsymbol{x}, \boldsymbol{c}_{\mathrm{obs}}) \mapsto \mathcal{L}_{\mathrm{rob}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{c}_{\mathrm{obs}}) \ \Big| \ \boldsymbol{\theta} \in \Theta \Big\}. \tag{29}$$

We define the expected loss on the population and the empirical loss as

$$R(\boldsymbol{\theta}) := \mathbb{E}[g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}})], \qquad \widehat{R}_S(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i). \tag{30}$$

Rademacher Complexity We introduce Rademacher complexity as a measure of the expressive power of the loss class $\mathcal{G}$ (how well it can fit random signs on finite samples) [Bartlett and Mendelson, 2002]. First, we define a sequence of independent and identically distributed random variables $\sigma_1, \ldots, \sigma_N$ by

$$\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = \frac{1}{2}, \qquad i = 1, \ldots, N. \tag{31}$$

These σ_i are called Rademacher variables.

The empirical Rademacher complexity for sample $S = (\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_N)$ is defined as

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) := \mathbb{E}_\sigma \Big[\sup_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i g(\tilde{\boldsymbol{s}}_i) \Big]. \tag{32}$$

Here $\mathbb{E}_\sigma[\cdot]$ denotes the expectation with respect to the Rademacher variables $\sigma_1, \ldots, \sigma_N$.

Furthermore, we define the average Rademacher complexity, which also averages the randomness of the sample itself, as

$$\mathfrak{R}_N(\mathcal{G}) := \mathbb{E}_S[\widehat{\mathfrak{R}}_S(\mathcal{G})]. \tag{33}$$

Uniform Generalization

Theorem B.4 (Uniform Convergence for Uniformly Bounded Loss)**.** *Assume Assumption B.3 holds, and thus $0 \leq g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}) \leq B$ holds almost surely. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sup_{\boldsymbol{\theta} \in \Theta} \big(R(\boldsymbol{\theta}) - \widehat{R}_S(\boldsymbol{\theta}) \big) \leq 2\, \widehat{\mathfrak{R}}_S(\mathcal{G}) + 3B \sqrt{\frac{\log(2/\delta)}{2N}} \tag{34}$$

holds. Furthermore, taking the expectation,

$$\mathbb{E}_S \Big[\sup_{\boldsymbol{\theta} \in \Theta} \big(R(\boldsymbol{\theta}) - \widehat{R}_S(\boldsymbol{\theta}) \big) \Big] \leq 2\, \mathfrak{R}_N(\mathcal{G}) \tag{35}$$

holds.

(Proof in Section D.7)

Corollary B.5 (Generalization Guarantee for Optimal Solution of Robust Regret Loss)**.** *Let $\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \widehat{R}_S(\boldsymbol{\theta})$. From Theorem B.4,*

$$R(\hat{\boldsymbol{\theta}}) \leq \widehat{R}_S(\hat{\boldsymbol{\theta}}) + 2\, \widehat{\mathfrak{R}}_S(\mathcal{G}) + 3B \sqrt{\frac{\log(2/\delta)}{2N}}$$

holds.

(Proof in Section D.8)

C DETAILS OF DISTRIBUTIONALLY ROBUST REGRET LOSS

C.1 UNIFORM GENERALIZATION OF DISTRIBUTIONALLY ROBUST REGRET LOSS

C.1.1 Setup

Let $\tilde{\boldsymbol{s}} := (\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{c}}) \sim \mathcal{D}$, and let the sample $S = (\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_N)$ be i.i.d. samples from $\mathcal{D}$. Here, $\mathcal{D}$ represents the joint distribution of the covariate $\tilde{\boldsymbol{x}}$ and the coefficient vector $\tilde{\boldsymbol{c}}$.

Let the feasible set $\mathcal{W} \subset \mathbb{R}^n$ be non-empty and compact. For the norm $\|\cdot\|$ and the dual norm $\|\cdot\|_*$, we define $\mathrm{diam}_*(\mathcal{W})$ by (26).

We define the optimal value function and the optimal solution as

$$z(\boldsymbol{c}) := \min_{\boldsymbol{w} \in \mathcal{W}} \boldsymbol{c}^\top \boldsymbol{w}, \qquad \boldsymbol{w}^*(\boldsymbol{c}) \in \arg\min_{\boldsymbol{w} \in \mathcal{W}} \boldsymbol{c}^\top \boldsymbol{w}.$$

Let a predictive model be $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^n$, and the model class be $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.

We define the regret as

$$\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c}) := \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}) \tag{36}$$

.

C.1.2 Definition of Distributionally Robust Regret Loss

Similar to the distributionally robust formulation in the main text, we consider the conditional distribution $P_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}}$ of the coefficient vector for each covariate vector $\boldsymbol{x}$. Here, $\tilde{\boldsymbol{c}}$ represents the random variable of the coefficient vector conditioned on $\tilde{\boldsymbol{x}}$.

For each $\boldsymbol{x}$, we introduce the 1-Wasserstein ball of the conditional distribution $P_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}}$

$$\mathbb{B}_\varepsilon(P_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}}) := \left\{ Q_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}} \;\middle|\; W_1\big(Q_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}},\, P_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}}\big) \le \varepsilon \right\}$$

as the uncertainty set. In this case, we define the distributionally robust regret loss against the shift of the conditional distribution as

$$\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}; \boldsymbol{\theta}) := \sup_{Q_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}} \in \mathbb{B}_\varepsilon(P_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}})} \mathbb{E}_{\tilde{\boldsymbol{c}} \sim Q_{\tilde{\boldsymbol{c}}|\tilde{\boldsymbol{x}}=\boldsymbol{x}}} \big[\mathrm{Reg}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \tilde{\boldsymbol{c}})\big] \tag{37}$$

.

Let the population loss be

$$R_{\mathrm{dro}}(\boldsymbol{\theta}) := \mathbb{E}\big[\mathcal{L}_{\mathrm{dro}}(\tilde{\boldsymbol{x}}; \boldsymbol{\theta})\big] \tag{38}$$

.

By the strong duality argument regarding 1-Wasserstein used in the main text, the right-hand side of (37) can be written in the following form:

$$\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}\big[\mathrm{Reg}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \tilde{\boldsymbol{c}}) \mid \tilde{\boldsymbol{x}} = \boldsymbol{x}\big] + \varepsilon \,\mathrm{Lip}\left(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x}))\right). \tag{39}$$

Here, $\mathrm{Lip}(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x})))$ is a regularization term depending only on $\boldsymbol{x}$.

We define the function $g_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{c})$ for a sample $(\boldsymbol{x}, \boldsymbol{c})$ as follows:

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{c}) := \mathrm{Reg}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{c}) + \varepsilon \,\mathrm{Lip}\left(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x}))\right). \tag{40}$$

Using the law of iterated expectations, the population loss $R_{\mathrm{dro}}(\boldsymbol{\theta})$ can be expressed as the expectation of $g_{\boldsymbol{\theta}}$:

$$R_{\mathrm{dro}}(\boldsymbol{\theta}) = \mathbb{E}\big[\mathcal{L}_{\mathrm{dro}}(\tilde{\boldsymbol{x}}; \boldsymbol{\theta})\big] = \mathbb{E}\Big[\mathbb{E}\big[g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{c}}) \mid \tilde{\boldsymbol{x}}\big]\Big] = \mathbb{E}[g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{c}})]. \tag{41}$$

.

Similarly, as an estimator for the expected value of the population loss $R_{\mathrm{dro}}(\boldsymbol{\theta}) = \mathbb{E}[g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}})]$, we define the quantity obtained by replacing the expectation with the empirical distribution (sample mean) as

$$\widehat{R}_{\mathrm{dro},S}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i) \tag{42}$$

.

C.1.3 Uniform Boundedness

Assumption C.1 (Upper Bound on Norm of Coefficient Vector)**.** *There exists a constant $\overline{R}_{\mathcal{C}} < \infty$ such that*

$$\|\tilde{\boldsymbol{c}}\| \le \overline{R}_{\mathcal{C}}$$

holds almost surely.

Lemma C.2 (Uniform Boundedness of $g_{\boldsymbol{\theta}}$)**.** *Under Assumption C.1, for any $\boldsymbol{\theta} \in \Theta$ and any $(\boldsymbol{x}, \boldsymbol{c})$,*

$$0 \le g_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{c}) \le B_{\mathrm{dro}}$$

holds. Here,

$$B_{\mathrm{dro}} := \big(\overline{R}_{\mathcal{C}} + \varepsilon\big)\,\mathrm{diam}_*(\mathcal{W}). \tag{43}$$

(Proof in Section D.9)

C.1.4 Uniform Generalization of Population Loss

For each $\boldsymbol{\theta} \in \Theta$, the per-sample loss $g_{\boldsymbol{\theta}} \colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ is given by (40). Let the loss class be

$$\mathcal{G}_{\mathrm{dro}} := \big\{\, g_{\boldsymbol{\theta}} \;\big|\; \boldsymbol{\theta} \in \Theta \,\big\},$$

and define $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\mathrm{dro}})$ and $\mathfrak{R}_N(\mathcal{G}_{\mathrm{dro}})$ similarly to (32)–(33).

Theorem C.3 (Uniform Generalization of Population Loss)**.** *Suppose Assumption C.1 holds, and thus $0 \le g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}) \le B_{\mathrm{dro}}$ holds almost surely by Lemma C.2. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{\boldsymbol{\theta}\in\Theta}\Big(R_{\mathrm{dro}}(\boldsymbol{\theta}) - \widehat{R}_{\mathrm{dro},S}(\boldsymbol{\theta})\Big) \le 2\,\widehat{\mathfrak{R}}_S(\mathcal{G}_{\mathrm{dro}}) + 3B_{\mathrm{dro}}\sqrt{\frac{\log(2/\delta)}{2N}}$$

holds. Furthermore, taking the expectation,

$$\mathbb{E}_S\Big[\sup_{\boldsymbol{\theta}\in\Theta}\big(R_{\mathrm{dro}}(\boldsymbol{\theta}) - \widehat{R}_{\mathrm{dro},S}(\boldsymbol{\theta})\big)\Big] \le 2\,\mathfrak{R}_N(\mathcal{G}_{\mathrm{dro}})$$

holds.

(Proof in Section D.10)

Corollary C.4 (Generalization Guarantee for Population Loss)**.** *Let $\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{R}_{\mathrm{dro},S}(\boldsymbol{\theta})$. From Theorem C.3,*

$$R_{\mathrm{dro}}(\hat{\boldsymbol{\theta}}) \le \widehat{R}_{\mathrm{dro},S}(\hat{\boldsymbol{\theta}}) + 2\,\widehat{\mathfrak{R}}_S(\mathcal{G}_{\mathrm{dro}}) + 3B_{\mathrm{dro}}\sqrt{\frac{\log(2/\delta)}{2N}}$$

holds.

(Proof in Section D.11)

C.1.5 Implications of This Section

The properties of the upper bound obtained from the proof in this section can be interpreted as follows.

- The upper bound of $g_{\boldsymbol{\theta}}$ is given by
$$B_{\mathrm{dro}} = (\overline{R}_{\mathcal{C}} + \varepsilon)\,\mathrm{diam}_*(\mathcal{W})$$
(Lemma C.2). Therefore, the larger the possible magnitude $\overline{R}_{\mathcal{C}}$ of the coefficient vector, the robust radius ε, and the diameter $\mathrm{diam}_*(\mathcal{W})$ of the feasible set, the larger the upper bound B_{dro} becomes, and the larger the second term of the upper bound becomes.
- The gap
$$\sup_{\boldsymbol{\theta}\in\Theta}\Big(R_{\mathrm{dro}}(\boldsymbol{\theta}) - \widehat{R}_{\mathrm{dro},S}(\boldsymbol{\theta})\Big)$$
is controlled by $2\widehat{\mathfrak{R}}_S(\mathcal{G}_{\mathrm{dro}})$ (Theorem C.3). This is a quantity representing how well the loss class can follow random signs on the sample, meaning that if the expressive power is excessively large, the upper bound becomes loose.
- Increasing ε makes the objective function more conservative in preparation for a wider distribution shift, but since B_{dro} of the upper bound also increases, the gap tends to become large with finite samples.
- Corollary C.4 provides a non-asymptotic guarantee that the learning solution $\hat{\boldsymbol{\theta}}$ that minimizes the DRO objective $\widehat{R}_{\mathrm{dro},S}$ based on the empirical distribution also has a small R_{dro} for the unknown population distribution with high probability.

From the above, the upper bound in this section clarifies how the magnitude of uncertainty (radius), the geometry of optimization (diameter), and the model expressive power affect generalization.

C.2 GRADIENT CALCULATION OF DISTRIBUTIONALLY ROBUST LOSS

We describe the details of the gradient calculation of the distributionally robust loss. Regarding $\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$, we need to calculate the gradient through the subgradient of

$$\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) = \sup_{\boldsymbol{w}\in\mathcal{W}} \|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\|_*$$

introduced in the previous section.

To do this, using the dual norm

$$\|\boldsymbol{u}\|_* = \max_{\boldsymbol{g}\in\mathcal{G}} \boldsymbol{g}^\top \boldsymbol{u}, \qquad \mathcal{G} := \{\boldsymbol{g}\in\mathbb{R}^n : \|\boldsymbol{g}\| \le 1\},$$

we rewrite it as

$$\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) = \max_{(\boldsymbol{w},\boldsymbol{g})\in\mathcal{W}\times\mathcal{G}} \boldsymbol{g}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}).$$

Here, $\mathcal{V} := \mathcal{W}\times\mathcal{G}$ is a compact set, and letting $\phi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}); \boldsymbol{w}, \boldsymbol{g}) := \boldsymbol{g}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w})$, ϕ is linear and continuous with respect to $(\boldsymbol{w}^*(\hat{\boldsymbol{c}}), \boldsymbol{w}, \boldsymbol{g})$, and for each $(\boldsymbol{w}, \boldsymbol{g})$, $\boldsymbol{w}^*(\hat{\boldsymbol{c}}) \mapsto \phi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}); \boldsymbol{w}, \boldsymbol{g})$ is C^1 and convex, and

$$\nabla_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}\phi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}); \boldsymbol{w}, \boldsymbol{g}) = \boldsymbol{g}$$

holds. Furthermore, $(\boldsymbol{w}, \boldsymbol{g}) \mapsto \nabla_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}\phi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}); \boldsymbol{w}, \boldsymbol{g}) = \boldsymbol{g}$ is continuous on $\mathcal{V}$. Therefore, the assumptions of Danskin's theorem are satisfied.

Thus, letting

$$\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) = \max_{v\in\mathcal{V}} \phi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}); v),$$

we have

$$\partial\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) = \mathrm{conv}\{\nabla_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}\phi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}); v) \mid v \in V(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\}.$$

The gradient of the parameters $\boldsymbol{\theta}$ of the predictive model with respect to the distributionally robust loss $\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x})$ can be written by the chain rule as

$$\frac{\partial}{\partial\hat{\boldsymbol{c}}}\,\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big) = J^\top_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}\,\boldsymbol{g}, \qquad \boldsymbol{g} \in \partial\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big).$$

Here, $J_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$ is the Jacobian matrix of $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$, and $\boldsymbol{g}$ is the subgradient obtained from Danskin's theorem above. In implementation, we perform the calculation using methods such as the one proposed in [Huang and Gupta, 2024] as a method to calculate the approximate gradient of $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ without using the gradient of SPO+.

D PROOFS

D.1 PROOF OF THEOREM 3.1

First, as a preparation for gradient calculation, we show a lemma about the continuity of the optimal value function $z(\boldsymbol{c})$.

Lemma D.1. *Let $\mathcal{W} \subseteq \mathbb{R}^n$ be a compact set, and define $z : \mathbb{R}^n \to \mathbb{R}$ by $z(\boldsymbol{c}) := \min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{c}^\top \boldsymbol{w}$. Then z is Lipschitz continuous.*

The proof of this lemma is given in the next section (Section D.2).

Proof of Theorem 3.1. We apply Danskin's theorem [Bertsekas, 1999, Proposition B.25] (Proposition A.1). For the proposed robust regret loss, regarding $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ as a variable and $\boldsymbol{c} \in \mathcal{C}$ as a coefficient vector, let

$$\psi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}), \boldsymbol{c}) := \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}).$$

From the non-empty compactness of $\mathcal{W}, \mathcal{C}$ and the continuity of $z(\boldsymbol{c})$ (Lemma D.1), ψ is continuous in $(\boldsymbol{w}^*(\hat{\boldsymbol{c}}), \boldsymbol{c})$, and since it is affine with respect to $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ for each $\boldsymbol{c}$, it is convex and C^1, and

$$\nabla_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})} \psi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}), \boldsymbol{c}) = \boldsymbol{c}.$$

Also, since $\mathcal{C}$ is compact, it satisfies the assumption of Proposition A.1. Therefore, the subgradient set of the robust regret loss with respect to $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$ is derived as the convex hull of the worst-case coefficient vectors $\boldsymbol{c}^* \in \arg\max_{\boldsymbol{c}\in\mathcal{C}} \psi(\boldsymbol{w}^*(\hat{\boldsymbol{c}}), \boldsymbol{c})$:

$$\partial_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})} \mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) = \mathrm{conv}\Big\{ \operatorname*{arg\,max}_{\boldsymbol{c}\in\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \big(\boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c})\big) \Big\}. \tag{44}$$

□

D.2 PROOF OF LEMMA D.1

Proof. From the compactness of $\mathcal{W}$,

$$M := \sup_{w\in\mathcal{W}} \|w\| < \infty$$

holds. Take arbitrary $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^n$, and let $\boldsymbol{w}' \in \mathcal{W}$ be a point satisfying $z(\boldsymbol{c}') = \min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{c}'^\top \boldsymbol{w} = \boldsymbol{c}'^\top \boldsymbol{w}'$. Note that since $\mathcal{W}$ is compact and the objective function is continuous, the minimum is always achieved at some point in $\mathcal{W}$. At this time,

$$\begin{aligned} z(\boldsymbol{c}) - z(\boldsymbol{c}') &= z(\boldsymbol{c}) - \boldsymbol{c}'^\top \boldsymbol{w}' \\ &\le \boldsymbol{c}^\top \boldsymbol{w}' - \boldsymbol{c}'^\top \boldsymbol{w}' \\ &= (\boldsymbol{c} - \boldsymbol{c}')^\top \boldsymbol{w}' \\ &\le \|\boldsymbol{c} - \boldsymbol{c}'\| \, \|\boldsymbol{w}'\| \\ &\le M \|\boldsymbol{c} - \boldsymbol{c}'\| \end{aligned}$$

holds. Similarly, by swapping $\boldsymbol{c}$ and $\boldsymbol{c}'$,

$$z(\boldsymbol{c}') - z(\boldsymbol{c}) \le M \|\boldsymbol{c} - \boldsymbol{c}'\|$$

also follows, so

$$|z(\boldsymbol{c}) - z(\boldsymbol{c}')| \le M \|\boldsymbol{c} - \boldsymbol{c}'\|$$

holds for arbitrary $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^n$. Therefore, z is Lipschitz continuous. □

D.3 PROOF OF THEOREM 3.2

Proof. To derive an equivalent representation of the distributionally robust loss $\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x})$, we apply the Blanchet–Murthy strong duality theorem [Blanchet and Murthy, 2019]. As a preliminary step, we first state a proposition about the Lipschitz continuity of the objective function $\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$.

Proposition D.2 (Lipschitz continuity of $\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$). *Let the constant for arbitrary $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^n$ be*

$$\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) := \sup_{\boldsymbol{w} \in \mathcal{W}} \|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\|_*,$$

then the following inequality holds:

$$\left|\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}) - \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}')\right| \;\le\; \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\, \|\boldsymbol{c} - \boldsymbol{c}'\|.$$

That is, $\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$ is Lipschitz continuous with respect to the norm $\|\cdot\|$, and $\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$ is the Lipschitz constant.

The proof of this proposition is given in the next section (Section D.4).

From Proposition D.2, $\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$ is a Lipschitz continuous function and is integrable with respect to $P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}}$.

Letting the distance be $d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|$, Assumptions (A1), (A2) of the Blanchet–Murthy strong duality [Blanchet and Murthy, 2019] are satisfied by Proposition D.2. Therefore,

$$\sup_{Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}} \in \mathbb{B}_\varepsilon(P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}})} \mathbb{E}_{Q_{\tilde{\boldsymbol{c}}|\boldsymbol{x}}}[\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}})] = \inf_{\lambda \ge 0} \Big\{ \lambda\varepsilon + \mathbb{E}_{P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}}}\big[\sup_{\boldsymbol{y}} \{\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|\}\big] \Big\}$$

holds. Although maximization by $\boldsymbol{y}$ for each $\tilde{\boldsymbol{c}}$ remains on the right side, by the following Theorem D.3, this term is divided into a finite value or infinity depending on the value of λ.

Theorem D.3. *For any $\lambda \ge 0,\ \tilde{\boldsymbol{c}} \in \mathbb{R}^n$,*

$$\sup_{\boldsymbol{y} \in \mathbb{R}^n} \{\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|\} = \begin{cases} \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}}), & \lambda \ge \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})), \\ +\infty, & \lambda < \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})), \end{cases}$$

holds. In particular, the optimal value in

$$\inf_{\lambda \ge 0} \Big\{ \lambda\varepsilon + \mathbb{E}_{\mu}\big[\sup_{\boldsymbol{y}} \{\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|\}\big] \Big\}$$

appearing in the Blanchet–Murthy strong duality is achieved at $\lambda^* = \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$.

The proof of this theorem is given in Section D.5.

Using Theorem D.3, the optimal value of the dual problem is achieved at $\lambda = \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$, and

$$\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}) = \mathbb{E}_{\tilde{\boldsymbol{c}} \sim P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}}}[\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}})] + \varepsilon\, \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$$

is obtained. □

D.4 PROOF OF PROPOSITION D.2

Proof. $\mathcal{W}$ is compact, and since the mapping $\boldsymbol{w} \mapsto \boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w})$ is continuous, the maximum value is always achieved at some point in $\mathcal{W}$. For each $\boldsymbol{c} \in \mathbb{R}^n$, choose one

$$\boldsymbol{w}_{\boldsymbol{c}} \in \arg\max_{\boldsymbol{w} \in \mathcal{W}} \ \boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}).$$

Then

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}) = \ \boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}_{\boldsymbol{c}}) \ \le \ \boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}_{\boldsymbol{c}'})$$

holds, so

$$\begin{aligned}\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}) - \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}') \leq\ & \boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}_{\boldsymbol{c}'}) - \ \boldsymbol{c}'^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}_{\boldsymbol{c}'}) \\ =\ & (\boldsymbol{c} - \boldsymbol{c}')^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}_{\boldsymbol{c}'}).\end{aligned}$$

From the definition of the dual norm

$$(\boldsymbol{c} - \boldsymbol{c}')^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}_{\boldsymbol{c}'}) \ \leq\ \|\boldsymbol{c} - \boldsymbol{c}'\|\,\|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}_{\boldsymbol{c}'}\|_* \ \leq\ \|\boldsymbol{c} - \boldsymbol{c}'\| \sup_{\boldsymbol{w}\in\mathcal{W}} \|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\|_* = \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\,\|\boldsymbol{c} - \boldsymbol{c}'\|.$$

Therefore

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}) - \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}') \ \leq\ \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\,\|\boldsymbol{c} - \boldsymbol{c}'\|.$$

Similarly, swapping the roles of $\boldsymbol{c}$ and $\boldsymbol{c}'$,

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}') - \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}) \ \leq\ \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\,\|\boldsymbol{c} - \boldsymbol{c}'\|$$

also follows, so combining the two,

$$\left|\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}) - \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{c}')\right| \ \leq\ \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\,\|\boldsymbol{c} - \boldsymbol{c}'\|$$

holds for arbitrary $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^n$. This shows the claim. □

D.5 PROOF OF THEOREM D.3

Proof. First, from Proposition D.2, $\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$ satisfies

$$\left|\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}})\right| \leq \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\,\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\| \qquad (\forall\, \boldsymbol{y}, \tilde{\boldsymbol{c}} \in \mathbb{R}^n)$$

with respect to the norm $\|\cdot\|$.

(1) Case $\lambda \geq \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$**:** For any $\boldsymbol{y} \in \mathbb{R}^n$,

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) \leq \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}}) + \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\,\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\| \leq \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}}) + \lambda\,\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|$$

holds. Therefore

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\| \leq \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}}) \qquad (\forall\, \boldsymbol{y}).$$

On the other hand, if we take $\boldsymbol{y} = \tilde{\boldsymbol{c}}$, the left side becomes $\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}}) - \lambda\|\tilde{\boldsymbol{c}} - \tilde{\boldsymbol{c}}\| = \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}})$, so the following can be derived:

$$\sup_{\boldsymbol{y}\in\mathbb{R}^n} \{\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|\} = \phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}}).$$

(2) Case $\lambda < \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$**:** From the definition of $L(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$,

$$\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) = \sup_{\boldsymbol{w}\in\mathcal{W}} \left\|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\right\|_*.$$

Since $\mathcal{W}$ is a compact set and the mapping

$$\boldsymbol{w} \ \mapsto\ \left\|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\right\|_*$$

is continuous, there exists a point $\boldsymbol{w}^* \in \mathcal{W}$ achieving the maximum value, and

$$\left\|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*\right\|_* = \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$$

holds.

Next, if we take the unit sphere

$$\mathcal{G} := \{\boldsymbol{g} \in \mathbb{R}^n : \|\boldsymbol{g}\| = 1\},$$

since $\mathcal{G}$ is compact and the mapping

$$\boldsymbol{g} \mapsto \langle \boldsymbol{g}, \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*\rangle$$

is continuous, there exists $\boldsymbol{g}^* \in \mathcal{G}$ achieving the maximum value, satisfying

$$\max_{\boldsymbol{g}\in\mathcal{G}} \boldsymbol{g}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*) = \left\|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*\right\|_* = \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})).$$

For arbitrary $t > 0$, let

$$\boldsymbol{y}_t := \tilde{\boldsymbol{c}} + t\,\boldsymbol{g}^*.$$

From the definition of $\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}$

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) = \max_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{y}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}),$$

we have

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}_t) \;\geq\; \boldsymbol{y}_t^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*) = \tilde{\boldsymbol{c}}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*) + t\;(\boldsymbol{g}^{*\top}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*)) = \tilde{\boldsymbol{c}}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*) + t\,\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})).$$

On the other hand, since $\|\boldsymbol{y}_t - \tilde{\boldsymbol{c}}\| = \|t\,\boldsymbol{v}^*\| = t$,

$$\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}_t) - \lambda\|\boldsymbol{y}_t - \tilde{\boldsymbol{c}}\| \;\geq\; \tilde{\boldsymbol{c}}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*) + t\big(\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) - \lambda\big).$$

Here, since $\lambda < \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$, $\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) - \lambda > 0$, and the right side diverges to $+\infty$ as $t \to \infty$. Therefore

$$\sup_{\boldsymbol{y}\in\mathbb{R}^n}\left\{\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|\right\} = +\infty$$

follows.

From the above, for arbitrary $\lambda \geq 0,\ \tilde{\boldsymbol{c}} \in \mathbb{R}^n$, it is shown that

$$\sup_{\boldsymbol{y}\in\mathbb{R}^n}\{\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|\} = \begin{cases}\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}}), & \lambda \geq \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})),\\ +\infty, & \lambda < \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))\end{cases}$$

holds. Finally, in the dual problem

$$\inf_{\lambda\geq 0}\Big\{\lambda\varepsilon + \mathbb{E}_\mu\big[\sup_{\boldsymbol{y}}\{\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\boldsymbol{y}) - \lambda\|\boldsymbol{y} - \tilde{\boldsymbol{c}}\|\}\big]\Big\},$$

in the range $\lambda < \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$, the right side becomes $+\infty$, and for $\lambda \geq \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$, the right side is $\lambda\varepsilon + \mathbb{E}_\mu[\phi_{\boldsymbol{w}^*(\hat{\boldsymbol{c}})}(\tilde{\boldsymbol{c}})]$, so the minimum value with respect to λ is achieved at $\lambda^\star = \mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}}))$. □

D.6 PROOF OF LEMMA B.2

Proof. First, we show non-negativity. Fix arbitrary $\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$. From $z(\boldsymbol{c}) = \min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{c}^\top\boldsymbol{w}$, for any $\boldsymbol{w} \in \mathcal{W}$,

$$z(\boldsymbol{c}) \leq \boldsymbol{c}^\top\boldsymbol{w} \tag{45}$$

holds. In particular, taking $\boldsymbol{w} = \boldsymbol{w}^*(\hat{\boldsymbol{c}}) \in \mathcal{W}$,

$$\boldsymbol{c}^\top\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}) \geq \boldsymbol{c}^\top\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{c}^\top\boldsymbol{w}^*(\hat{\boldsymbol{c}}) = 0.$$

Therefore, even taking the maximum,

$$\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) = \max_{\boldsymbol{c}\in\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})}\{\boldsymbol{c}^\top\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c})\} \geq 0$$

follows.

Next, we show the upper bound. For arbitrary $\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$,

$$\begin{aligned}\boldsymbol{c}^\top\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}) &= \boldsymbol{c}^\top\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{c}^\top\boldsymbol{w}\\ &= \max_{\boldsymbol{w}\in\mathcal{W}}\big(\boldsymbol{c}^\top\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{c}^\top\boldsymbol{w}\big)\\ &= \max_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}).\end{aligned}$$

From Hölder's inequality, for arbitrary $\boldsymbol{w} \in \mathcal{W}$,

$$\boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) \leq \|\boldsymbol{c}\| \cdot \|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\|_* \tag{46}$$

holds. Therefore

$$\max_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{c}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}) \leq \|\boldsymbol{c}\| \cdot \max_{\boldsymbol{w}\in\mathcal{W}} \|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\|_*.$$

Here, since $\boldsymbol{w}^*(\hat{\boldsymbol{c}}) \in \mathcal{W}$,

$$\max_{\boldsymbol{w}\in\mathcal{W}} \|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\|_* \leq \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{W}} \|\boldsymbol{u} - \boldsymbol{v}\|_* = \mathrm{diam}_*(\mathcal{W}).$$

Therefore, for arbitrary $\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$,

$$\boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}) \leq \|\boldsymbol{c}\| \cdot \mathrm{diam}_*(\mathcal{W})$$

is obtained. Taking the maximum of $\boldsymbol{c} \in \mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})$ on both sides,

$$\mathcal{L}_{\mathrm{rob}}(\hat{\boldsymbol{c}}, \boldsymbol{c}_{\mathrm{obs}}) \leq \Big(\sup_{\boldsymbol{c}\in\mathcal{C}(\boldsymbol{c}_{\mathrm{obs}})} \|\boldsymbol{c}\|\Big) \cdot \mathrm{diam}_*(\mathcal{W}) = R_{\mathcal{C}}(\boldsymbol{c}_{\mathrm{obs}}) \cdot \mathrm{diam}_*(\mathcal{W})$$

becomes. Thus the claim holds. □

D.7 PROOF OF THEOREM B.4

Proof. We show the expected value bound (35) and generalization (34) in order.

Expected Value Bound Let $\mathcal{D} = (\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_N)$ and $\mathcal{D}' = (\tilde{\boldsymbol{s}}'_1, \ldots, \tilde{\boldsymbol{s}}'_N)$ be i.i.d. draws from $\mathcal{P}$. For each $\boldsymbol{\theta}$,

$$R(\boldsymbol{\theta}) = \mathbb{E}[g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}})] = \mathbb{E}_{\mathcal{D}'}\Big[\frac{1}{N}\sum_{i=1}^N g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i)\Big] = \mathbb{E}_{\mathcal{D}'}[\widehat{R}_{\mathcal{D}'}(\boldsymbol{\theta})].$$

Therefore

$$R(\boldsymbol{\theta}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}'}[\widehat{R}_{\mathcal{D}'}(\boldsymbol{\theta})] - \widehat{R}_{\mathcal{D}}(\boldsymbol{\theta}).$$

Taking $\sup_{\boldsymbol{\theta}}$ and expectation with respect to $\mathcal{D}$, by the convexity of sup and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\Big[\sup_{\boldsymbol{\theta}}(R(\boldsymbol{\theta}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{\theta}))\Big] &= \mathbb{E}_{\mathcal{D}}\Big[\sup_{\boldsymbol{\theta}}\big(\mathbb{E}_{\mathcal{D}'}[\widehat{R}_{\mathcal{D}'}(\boldsymbol{\theta})] - \widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})\big)\Big] \\
&\leq \mathbb{E}_{\mathcal{D},\mathcal{D}'}\Big[\sup_{\boldsymbol{\theta}}\big(\widehat{R}_{\mathcal{D}'}(\boldsymbol{\theta}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})\big)\Big] \\
&= \mathbb{E}_{\mathcal{D},\mathcal{D}'}\Big[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^N\big(g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i) - g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i)\big)\Big].
\end{aligned}$$

Here, we introduce independent Rademacher variables $\sigma_1, \ldots, \sigma_N$. Since $\mathcal{D}$ and $\mathcal{D}'$ are identically distributed and independent, the joint distribution of $\{(\tilde{\boldsymbol{s}}_i, \tilde{\boldsymbol{s}}'_i)\}_{i=1}^N$ does not change even if $\tilde{\boldsymbol{s}}_i$ and $\tilde{\boldsymbol{s}}'_i$ are swapped for each i. Due to this symmetry, the expected value does not change even if the sign of the difference $\big(g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i) - g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i)\big)$ is randomly flipped by $\sigma_i \in \{+1, -1\}$, and

$$\mathbb{E}_{\mathcal{D},\mathcal{D}'}\Big[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^N\big(g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i) - g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i)\big)\Big] = \mathbb{E}_{\mathcal{D},\mathcal{D}',\sigma}\Big[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^N\sigma_i\big(g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i) - g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i)\big)\Big]$$

holds.

Also, since $\sup_{\boldsymbol{\theta}}(a_{\boldsymbol{\theta}} + b_{\boldsymbol{\theta}}) \leq \sup_{\boldsymbol{\theta}} a_{\boldsymbol{\theta}} + \sup_{\boldsymbol{\theta}} b_{\boldsymbol{\theta}}$,

$$\sup_{\boldsymbol{\theta}}\sum_{i=1}^N\sigma_i\big(g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i) - g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i)\big) \leq \sup_{\boldsymbol{\theta}}\sum_{i=1}^N\sigma_i g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i) + \sup_{\boldsymbol{\theta}}\sum_{i=1}^N(-\sigma_i)g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i).$$

Therefore

$$\mathbb{E}_{\mathcal{D}}\Big[\sup_{\boldsymbol{\theta}}(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta}))\Big] \leq \mathbb{E}_{\mathcal{D}',\sigma}\Big[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_i \sigma_i g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i)\Big] + \mathbb{E}_{\mathcal{D},\sigma}\Big[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_i(-\sigma_i) g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i)\Big].$$

Since Rademacher variables are identically distributed even if signs are flipped, the two terms are equal, and

$$\mathbb{E}_{\mathcal{D}}\Big[\sup_{\boldsymbol{\theta}}(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta}))\Big] \leq 2\,\mathbb{E}_{\mathcal{D},\sigma}\Big[\sup_{g\in\mathcal{G}}\frac{1}{N}\sum_i \sigma_i g(\tilde{\boldsymbol{s}}_i)\Big] = 2\,\mathfrak{R}_N(\mathcal{G})$$

is obtained. This shows (35).

Generalization First, we directly evaluate the amount of change in $\Phi(\mathcal{D})$ due to replacing one sample point. Let

$$\Phi(\mathcal{D}) := \sup_{\boldsymbol{\theta}\in\Theta}\big(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})\big).$$

Here $\mathcal{D}=(\tilde{\boldsymbol{s}}_1,\ldots,\tilde{\boldsymbol{s}}_N)$ are i.i.d. samples. Let $\mathcal{D}^{(i)}$ be the sample obtained by replacing the i-th element $\tilde{\boldsymbol{s}}_i$ of $\mathcal{D}$ with an independent identically distributed sample $\tilde{\boldsymbol{s}}'_i$:

$$\mathcal{D}^{(i)} := (\tilde{\boldsymbol{s}}_1,\ldots,\tilde{\boldsymbol{s}}_{i-1},\tilde{\boldsymbol{s}}'_i,\tilde{\boldsymbol{s}}_{i+1},\ldots,\tilde{\boldsymbol{s}}_N).$$

At this time,

$$\begin{aligned}|\Phi(\mathcal{D})-\Phi(\mathcal{D}^{(i)})| &= \Big|\sup_{\boldsymbol{\theta}\in\Theta}\big(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})\big)-\sup_{\boldsymbol{\theta}\in\Theta}\big(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}^{(i)}}(\boldsymbol{\theta})\big)\Big| \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta}\Big|\big(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})\big)-\big(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}^{(i)}}(\boldsymbol{\theta})\big)\Big|\end{aligned}$$

holds. Here, in the last inequality, we used that generally

$$\Big|\sup_{\boldsymbol{\theta}} a_{\boldsymbol{\theta}}-\sup_{\boldsymbol{\theta}} b_{\boldsymbol{\theta}}\Big| \leq \sup_{\boldsymbol{\theta}}|a_{\boldsymbol{\theta}}-b_{\boldsymbol{\theta}}|$$

holds.

Furthermore, since $R(\boldsymbol{\theta})=\mathbb{E}[g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}})]$ does not depend on the sample $\mathcal{D}$,

$$\big(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})\big)-\big(R(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}^{(i)}}(\boldsymbol{\theta})\big) = \widehat{R}_{\mathcal{D}^{(i)}}(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})$$

holds. Therefore

$$|\Phi(\mathcal{D})-\Phi(\mathcal{D}^{(i)})| \leq \sup_{\boldsymbol{\theta}\in\Theta}\big|\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}^{(i)}}(\boldsymbol{\theta})\big|$$

is obtained.

Here, for arbitrary $\boldsymbol{\theta}$,

$$\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}^{(i)}}(\boldsymbol{\theta}) = \frac{1}{N}\big(g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}_i)-g_{\boldsymbol{\theta}}(\tilde{\boldsymbol{s}}'_i)\big),$$

so from the assumption $0\leq g_{\boldsymbol{\theta}}(\cdot)\leq B$,

$$\big|\widehat{R}_{\mathcal{D}}(\boldsymbol{\theta})-\widehat{R}_{\mathcal{D}^{(i)}}(\boldsymbol{\theta})\big| \leq \frac{B}{N}$$

follows. Therefore

$$|\Phi(S)-\Phi(S^{(i)})| \leq \frac{B}{N}$$

is shown.

Therefore, from McDiarmid's inequality [McDiarmid, 1989], for arbitrary $t>0$,

$$\Pr\big(\Phi(S)-\mathbb{E}[\Phi(S)]\geq t\big) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^N (B/N)^2}\Big) = \exp\Big(-\frac{2Nt^2}{B^2}\Big)$$

holds. Therefore, for arbitrary $\eta_1 \in (0,1)$, with probability at least $1-\eta_1$,

$$\Phi(S) \leq \mathbb{E}[\Phi(S)] + B\sqrt{\frac{\log(1/\eta_1)}{2N}} \tag{47}$$

holds.

Next, regarding the empirical Rademacher complexity $\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G})$, we also evaluate the amount of change due to replacing one sample point.

Let $\mathcal{D}^{(i)}$ be the sample obtained by replacing the i-th element of $\mathcal{D} = (\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_N)$ with an independent identically distributed sample $\tilde{\boldsymbol{s}}_i'$:

$$\mathcal{D}^{(i)} := (\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_{i-1}, \tilde{\boldsymbol{s}}_i', \tilde{\boldsymbol{s}}_{i+1}, \ldots, \tilde{\boldsymbol{s}}_N).$$

At this time, it suffices to evaluate

$$\left|\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) - \widehat{\mathfrak{R}}_{\mathcal{D}^{(i)}}(\mathcal{G})\right| = \left|\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j)\right] - \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j^{(i)})\right]\right|.$$

Here $\sigma = (\sigma_1, \ldots, \sigma_N)$ are Rademacher variables.

Using the linearity of expectation and $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$,

$$\left|\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) - \widehat{\mathfrak{R}}_{\mathcal{D}^{(i)}}(\mathcal{G})\right| \leq \mathbb{E}_\sigma\left[\left|\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j) - \sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j^{(i)})\right|\right].$$

Here, let

$$\psi_\sigma(\mathcal{D}) := \sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j),$$

then

$$\left|\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) - \widehat{\mathfrak{R}}_{\mathcal{D}^{(i)}}(\mathcal{G})\right| \leq \mathbb{E}_\sigma\left[\,|\psi_\sigma(\mathcal{D}) - \psi_\sigma(\mathcal{D}^{(i)})|\,\right]$$

holds. Thus, we upper bound $|\psi_\sigma(\mathcal{D}) - \psi_\sigma(\mathcal{D}^{(i)})|$.

Since for any sequences of real numbers $\{a_g\}$ and $\{b_g\}$,

$$\left|\sup_g a_g - \sup_g b_g\right| \leq \sup_g |a_g - b_g|$$

holds, we have

$$\begin{aligned}
|\psi_\sigma(\mathcal{D}) - \psi_\sigma(\mathcal{D}^{(i)})| &= \left|\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j) - \sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j^{(i)})\right| \\
&\leq \sup_{g\in\mathcal{G}}\left|\frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j) - \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j^{(i)})\right|.
\end{aligned}$$

Here, since the difference between $\mathcal{D}$ and $\mathcal{D}^{(i)}$ is only in the i-th element, for any $g \in \mathcal{G}$,

$$\frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j) - \frac{1}{N}\sum_{j=1}^N \sigma_j g(\tilde{\boldsymbol{s}}_j^{(i)}) = \frac{1}{N}\sigma_i\big(g(\tilde{\boldsymbol{s}}_i) - g(\tilde{\boldsymbol{s}}_i')\big)$$

holds. From the assumption $0 \leq g(\cdot) \leq B$,

$$\left|\frac{1}{N}\sigma_i\big(g(\tilde{\boldsymbol{s}}_i) - g(\tilde{\boldsymbol{s}}_i')\big)\right| \leq \frac{1}{N}\big|g(\tilde{\boldsymbol{s}}_i) - g(\tilde{\boldsymbol{s}}_i')\big| \leq \frac{B}{N}.$$

Therefore,

$$|\psi_\sigma(\mathcal{D}) - \psi_\sigma(\mathcal{D}^{(i)})| \leq \frac{B}{N}$$

follows. Thus, we obtain

$$\left|\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) - \widehat{\mathfrak{R}}_{\mathcal{D}^{(i)}}(\mathcal{G})\right| \leq \mathbb{E}_\sigma\left[\,|\psi_\sigma(\mathcal{D}) - \psi_\sigma(\mathcal{D}^{(i)})|\,\right] \leq \frac{B}{N}.$$

Therefore, by McDiarmid's inequality, for any $t > 0$,

$$\Pr\left(\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G})] \geq t\right) \leq \exp\left(-\frac{2Nt^2}{B^2}\right)$$

holds. Thus, for any $\eta_2 \in (0,1)$, with probability at least $1-\eta_2$,

$$\mathbb{E}_{\mathcal{D}}[\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G})] \leq \widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) + B\sqrt{\frac{\log(1/\eta_2)}{2N}} \tag{48}$$

holds. That is,

$$\mathfrak{R}_N(\mathcal{G}) \leq \widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) + B\sqrt{\frac{\log(1/\eta_2)}{2N}}$$

Finally, in (47), by using $\mathbb{E}[\Phi(\mathcal{D})] \leq 2\mathfrak{R}_N(\mathcal{G})$ and further using (48) to upper bound $\mathfrak{R}_N(\mathcal{G})$ by $\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G})$, specifically, evaluating the right-hand side of (47) as

$$\mathbb{E}[\Phi(\mathcal{D})] \leq 2\mathfrak{R}_N(\mathcal{G}) \leq 2\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) + 2B\sqrt{\frac{\log(1/\eta_2)}{2N}},$$

we obtain

$$\Phi(\mathcal{D}) \leq 2\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) + 2B\sqrt{\frac{\log(1/\eta_2)}{2N}} + B\sqrt{\frac{\log(1/\eta_1)}{2N}}.$$

Here, setting $\eta_1 = \eta_2 = \delta/2$, we have

$$2B\sqrt{\frac{\log(1/\eta_2)}{2N}} + B\sqrt{\frac{\log(1/\eta_1)}{2N}} = 3B\sqrt{\frac{\log(2/\delta)}{2N}}.$$

In this case, the probability that (47) and (48) hold simultaneously is at least $1-(\eta_1+\eta_2) = 1-\delta$. Thus, with probability at least $1-\delta$,

$$\Phi(\mathcal{D}) \leq 2\,\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) + 3B\sqrt{\frac{\log(2/\delta)}{2N}}$$

is obtained. This is (34).

□

D.8 PROOF OF COROLLARY B.5

Proof. From Theorem B.4, for any $\boldsymbol{\theta}$,

$$R(\boldsymbol{\theta}) \leq \widehat{R}_{\mathcal{D}}(\boldsymbol{\theta}) + 2\,\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}) + 3B\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Applying this to $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ proves the claim. □

D.9 PROOF OF LEMMA C.2

Proof. First, we show $\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c}) \geq 0$. Since $z(\boldsymbol{c}) = \min_{\boldsymbol{w} \in \mathcal{W}} \boldsymbol{c}^\top \boldsymbol{w}$, for any $\boldsymbol{w} \in \mathcal{W}$, $z(\boldsymbol{c}) \leq \boldsymbol{c}^\top \boldsymbol{w}$ holds. In particular, taking $\boldsymbol{w} = \boldsymbol{w}^*(\hat{\boldsymbol{c}}) \in \mathcal{W}$, we have

$$\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c}) = \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - z(\boldsymbol{c}) \geq \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{c}^\top \boldsymbol{w}^*(\hat{\boldsymbol{c}}) = 0.$$

Next, we show the upper bound of Reg. Similar to the proof of Lemma B.2, for any $\boldsymbol{c}$,

$$\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c}) \leq \|\boldsymbol{c}\| \operatorname{diam}_*(\mathcal{W}).$$

From Assumption C.1, since $\|\boldsymbol{c}\| \leq \overline{R}_{\mathcal{C}}$,

$$\mathrm{Reg}(\hat{\boldsymbol{c}}, \boldsymbol{c}) \leq \overline{R}_{\mathcal{C}} \operatorname{diam}_*(\mathcal{W}).$$

Furthermore, by definition, $\mathrm{Lip}(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x})))$ is

$$\mathrm{Lip}\left(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x}))\right) = \sup_{\boldsymbol{w} \in \mathcal{W}} \|\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x})) - \boldsymbol{w}\|_*$$

and

$$0 \leq \mathrm{Lip}\left(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x}))\right) \leq \operatorname{diam}_*(\mathcal{W})$$

holds. Therefore,

$$0 \leq \varepsilon \,\mathrm{Lip}\left(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x}))\right) \leq \varepsilon \operatorname{diam}_*(\mathcal{W}).$$

Summing these up,

$$0 \leq g_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{c}) = \mathrm{Reg}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{c}) + \varepsilon \,\mathrm{Lip}\left(\boldsymbol{w}^*(f_{\boldsymbol{\theta}}(\boldsymbol{x}))\right) \leq (\overline{R}_{\mathcal{C}} + \varepsilon) \operatorname{diam}_*(\mathcal{W}) = B_{\mathrm{dro}}.$$

Thus the claim holds. □

D.10 PROOF OF THEOREM C.3

Proof. The proof follows the same framework as the proof of Theorem B.4. The only differences are

$$\mathcal{G} \;\to\; \mathcal{G}_{\mathrm{dro}}, \qquad B \;\to\; B_{\mathrm{dro}}.$$

By Lemma C.2, $g_{\boldsymbol{\theta}}$ falls within $[0, B_{\mathrm{dro}}]$ almost surely, so the change due to the replacement of one sample point is bounded by B_{dro}/N. This yields the generalization bound with high probability ([Bartlett and Mendelson, 2002, McDiarmid, 1989]). □

D.11 PROOF OF COROLLARY C.4

Proof. From Theorem C.3, for any $\boldsymbol{\theta}$,

$$R_{\mathrm{dro}}(\boldsymbol{\theta}) \leq \widehat{R}_{\mathrm{dro},\mathcal{D}}(\boldsymbol{\theta}) + 2\,\widehat{\mathfrak{R}}_{\mathcal{D}}(\mathcal{G}_{\mathrm{dro}}) + 3B_{\mathrm{dro}}\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Applying this to $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ proves the claim. □

E DETAILS OF OPTIMIZATION PROBLEMS AND ALGORITHMS

E.1 DETAILS OF OPTIMIZATION PROBLEMS

Specific formulations of the optimization problems used in the experiments are shown below.

E.1.1 Shortest Path (SP)

We find the shortest path from a start node $s \in V$ to a target node $t \in V$ on a directed graph $G = (V, A)$. For each edge $a = (u, v) \in A$, a cost c_a is given, and the variable w_a represents whether to use edge a. The SP in PyEPO is formulated as follows:

$$\min_{\boldsymbol{w}} \quad \sum_{(u,v)\in A} c_{uv}\, w_{uv} \tag{49}$$

$$\text{s.t.} \quad \sum_{(v,u)\in A} w_{vu} - \sum_{(u,v)\in A} w_{uv} = \begin{cases} -1 & (u = s), \\ 1 & (u = t), \\ 0 & (u \in V \setminus \{s, t\}), \end{cases} \tag{50}$$

$$0 \le w_{uv} \le 1 \qquad (\forall (u, v) \in A). \tag{51}$$

E.1.2 Traveling Salesman Problem (TSP)

Consider TSP on a vertex set $V = \{1, \ldots, m\}$ (in our experiments, $m = 8$). We introduce an undirected edge set $E = \{\{i, j\} \mid 1 \le i < j \le m\}$, where the variable $w_{ij} \in \{0, 1\}$ indicates whether edge $\{i, j\}$ is included in the tour. We use the following formulation:

$$\min_{\boldsymbol{w}} \quad \sum_{1\le i<j\le m} c_{ij}\, w_{ij} \tag{52}$$

$$\text{s.t.} \quad \sum_{j\in V:\, j\neq i} w_{\min\{i,j\},\, \max\{i,j\}} = 2 \qquad (\forall i \in V), \tag{53}$$

$$\sum_{\substack{1\le i<j\le m \\ i\in S,\ j\in S}} w_{ij} \le |S| - 1 \qquad (\forall S \subset V,\ 2 \le |S| \le m - 1), \tag{54}$$

$$w_{ij} \in \{0, 1\} \qquad (\forall \{i, j\} \in E). \tag{55}$$

E.1.3 Knapsack Problem (KP)

Consider a multi-dimensional ($L = 2$ in our experiments) 0–1 knapsack problem for a set of items $\{1, \ldots, n\}$ ($n = 16$). The variable $x_i \in \{0, 1\}$ indicates whether item i is selected. Given values c_i, resource consumptions $a_{\ell i}$, and capacities b_ℓ, KP is given by the following maximization problem:

$$\max_{\boldsymbol{x}} \quad \sum_{i=1}^{n} c_i x_i \tag{56}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} a_{\ell i} x_i \le b_\ell \qquad (\ell = 1, \ldots, L), \tag{57}$$

$$x_i \in \{0, 1\} \qquad (i = 1, \ldots, n). \tag{58}$$

E.1.4 Formulation of Energy Scheduling Problem

In the energy scheduling problem, we optimize the demand (power consumption) allocation over $T = 24$ periods (hourly). For each time $t \in \{1, \ldots, T\}$, let the decision variable $w_t \in \mathbb{R}$ be the demand at that time. Let $\boldsymbol{p}^{\text{lo}}, \boldsymbol{p}^{\text{sch}}, \boldsymbol{p}^{\text{up}} \in \mathbb{R}^T$ be the lower bound of demand, baseline demand, and upper bound of demand, respectively.

The coefficient vector of the objective function $\boldsymbol{c} \in \mathbb{R}^T$ represents the unit cost of demand (e.g., electricity price) at each time, assumed to be generated or predicted from covariates $\boldsymbol{x} \in \mathbb{R}^p$. The energy scheduling problem in PyEPO is formulated as:

$$\boldsymbol{w}^*(\boldsymbol{c}) \in \arg\min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{c}^\top \boldsymbol{w}, \qquad \mathcal{W} := \Big\{ \boldsymbol{w} \in \mathbb{R}^T \;\Big|\; \boldsymbol{p}^{\text{lo}} \le \boldsymbol{w} \le \boldsymbol{p}^{\text{up}},\ \mathbf{1}^\top \boldsymbol{w} = \mathbf{1}^\top \boldsymbol{p}^{\text{sch}} \Big\}. \tag{59}$$

Here, $\mathbf{1} \in \mathbb{R}^T$ is a vector of all ones. The constraint $\mathbf{1}^\top \boldsymbol{w} = \mathbf{1}^\top \boldsymbol{p}^{\text{sch}}$ ensures that the total demand over 24 hours matches the total baseline demand.

E.2 DETAILS OF LEARNING ALGORITHMS

In this section, we describe the details of the learning algorithms in our proposed method.

E.2.1 Learning Algorithms Robust to Observation Errors

In learning using the robust regret loss described in Section 3.1, when using SPO+ [Elmachtoub and Grigas, 2022], the gradient of the SPO+ loss with $\boldsymbol{c}^*$ regarded as the true coefficient vector,

$$G_{\mathrm{SPO+}}(\hat{\boldsymbol{c}}; \boldsymbol{c}^*) := 2\Big(\boldsymbol{w}^*(\boldsymbol{c}^*) - \boldsymbol{w}^*\big(2\hat{\boldsymbol{c}} - \boldsymbol{c}^*\big)\Big) \tag{60}$$

is used as an approximation of $\nabla_{\hat{\boldsymbol{c}}}\mathcal{L}_{\mathrm{rob}}$. When using FY loss [Blondel et al., 2020], using the optimal solution with regularization term Ω, $\hat{\boldsymbol{w}}_\Omega(\hat{\boldsymbol{c}}) = \arg\max_{\boldsymbol{w}\in\mathcal{W}}\{\langle\hat{\boldsymbol{c}}, \boldsymbol{w}\rangle - \Omega(\boldsymbol{w})\}$,

$$G_{\mathrm{FY}}(\hat{\boldsymbol{c}}; \boldsymbol{c}^*) := \hat{\boldsymbol{w}}_\Omega(\hat{\boldsymbol{c}}) - \boldsymbol{w}^*(\boldsymbol{c}^*) \tag{61}$$

is used as the gradient. Here $\boldsymbol{w}^*(\boldsymbol{c}^*)$ is the optimal solution for $\boldsymbol{c}^*$.

Summarizing the above procedure, the learning algorithm by the proposed method iterates the following.

1. Based on the current model output $\hat{\boldsymbol{c}}$, solve the maximization problem of the robust regret loss to find the worst-case coefficient vector $\boldsymbol{c}^*$.
2. For the obtained $\boldsymbol{c}^*$, compute the gradient $G(\hat{\boldsymbol{c}}; \boldsymbol{c}^*)$ of the surrogate loss with respect to $\hat{\boldsymbol{c}}$.
3. Update the model parameter $\boldsymbol{\theta}$ by backpropagation using this gradient G:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \cdot \left(\frac{\partial\hat{\boldsymbol{c}}}{\partial\boldsymbol{\theta}}\right)^\top G(\hat{\boldsymbol{c}}; \boldsymbol{c}^*)$$

Here, η is the learning rate. This allows learning a predictive model that makes robust decisions even for the worst case within the uncertainty set.

E.2.2 Learning Algorithms Robust to Distribution Shifts

The details of the learning algorithm using the distributionally robust regret loss described in Section 3.2 are as follows.

The learning algorithm using distributionally robust regret iterates the following.

1. Based on the current model output $\hat{\boldsymbol{c}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$, find the decision $\boldsymbol{w}^*(\hat{\boldsymbol{c}})$, and evaluate the distributionally robust loss

   $$\mathcal{L}_{\mathrm{dro}}(\boldsymbol{x}) = \mathbb{E}_{\tilde{\boldsymbol{c}}\sim P_{\tilde{\boldsymbol{c}}|\boldsymbol{x}}}\big[\mathrm{Reg}(\hat{\boldsymbol{c}}, \tilde{\boldsymbol{c}})\big] + \varepsilon\,\mathrm{Lip}\big(\boldsymbol{w}^*(\hat{\boldsymbol{c}})\big)$$

   Here $\mathrm{Lip}(\boldsymbol{w}^*(\hat{\boldsymbol{c}})) = \sup_{\boldsymbol{w}\in\mathcal{W}}\|\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}\|_*$.
2. Compute surrogate gradients for the two terms in the above equation. Specifically, for the regret term, compute $G_{\mathrm{reg}}(\hat{\boldsymbol{c}}; \tilde{\boldsymbol{c}})$ using existing surrogate gradients such as SPO+ or FY loss. For the Lipschitz term, using the optimal solution

   $$(\boldsymbol{w}^*, \boldsymbol{g}^*) \in \operatorname*{arg\,max}_{(\boldsymbol{w},\boldsymbol{g})\in\mathcal{W}\times\mathcal{G}} \boldsymbol{g}^\top(\boldsymbol{w}^*(\hat{\boldsymbol{c}}) - \boldsymbol{w}),$$

   compute its subgradient (or approximate gradient) $G_{\mathrm{lip}}(\hat{\boldsymbol{c}}; \boldsymbol{g}^*)$. Combining these, we use

   $$G_{\mathrm{dro}}(\hat{\boldsymbol{c}}) := G_{\mathrm{reg}}(\hat{\boldsymbol{c}}) + \varepsilon\, G_{\mathrm{lip}}(\hat{\boldsymbol{c}})$$

   as the approximate gradient of the distributionally robust loss.
3. Update the model parameter $\boldsymbol{\theta}$ by backpropagation using the obtained gradient $G_{\mathrm{dro}}(\hat{\boldsymbol{c}})$:

   $$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \cdot \left(\frac{\partial\hat{\boldsymbol{c}}}{\partial\boldsymbol{\theta}}\right)^\top G_{\mathrm{dro}}(\hat{\boldsymbol{c}}).$$

Here, η is the learning rate. This allows learning a predictive model that suppresses the deterioration of regret even when the conditional distribution of the coefficient vector is perturbed in the sense of Wasserstein distance.

F DETAILS OF EXPERIMENTAL SETTINGS

F.1 COMMON SETTINGS

Common Settings for Datasets and Learning In our experiments, the following common settings were used for all optimization problems.

- **Data Splitting**: The total number of samples N_{total} was split into training (N_{train}), validation (N_{val}), and test (N_{test}) sets. Specifically, we set $N_{\mathrm{train}} = 100$, $N_{\mathrm{val}} = 100$, and $N_{\mathrm{test}} = 1000$.
- **Dimension of Covariates**: $p = 5$.
- **Polynomial Degree**: A polynomial model of degree $d = 4$ was used for data generation (see Section F.2 for details).
- **Learning Settings**: Adam [Kingma and Ba, 2015] was used as the optimization algorithm, with a learning rate of 10^{-2}, ℓ_2 regularization coefficient of 10^{-4}, batch size of 64, and 200 epochs.
- **Size of Optimization Problems**: For SP, we used a 5×5 grid graph; for TSP, the number of vertices $m = 8$; for KP, the number of items $n = 16$ and the number of capacity constraints $L = 2$.

Settings for Perturbation Parameters **Parameters for Observation Error (Parameter Noise)**: For sparse perturbation (Sparse), the ratio of selected dimensions was $\gamma_{\mathrm{frac}} = 0.5$. The shift sizes δ were as follows:

- SP: $\delta = 5.0$
- TSP: $\delta = 28.0$
- KP: $\delta = 22.0$

Parameters for Distribution Shift (Distribution Noise): The parameters for each problem setting are as follows.

- SP: $d_\gamma = 4$, $\alpha_\gamma = 0.8$, $\sigma_{\mathrm{global}} = 1.0$.
- TSP: $d_\gamma = 4$, $\alpha_\gamma = 1.0$, $\sigma_{\mathrm{global}} = 0.30$.
- KP: $d_\gamma = 4$, $\alpha_\gamma = 0.6$, $\sigma_{\mathrm{global}} = 0.50$.

Settings for Real Data Experiments In the experiments using real data, the total number of data points was 362, consisting of 100 training samples, 100 validation samples, and 162 test samples.

Details of Evaluation Metrics To assess robustness against worst-case scenarios, we report the Conditional Value-at-Risk (CVaR_α) at confidence level $\alpha = 0.9$. Specifically, let $\tilde{r}_{(1)} \leq \cdots \leq \tilde{r}_{(N_{\mathrm{test}})}$ denote the sorted relative regrets on the test set. We compute CVaR_α as the average of the worst $(1-\alpha)$-fraction of regrets:

$$\mathrm{CVaR}_\alpha(\boldsymbol{R}) := \frac{1}{N_{\mathrm{test}} - k_\alpha + 1} \sum_{j=k_\alpha}^{N_{\mathrm{test}}} \tilde{r}_{(j)}, \tag{62}$$

where $k_\alpha := \lceil \alpha N_{\mathrm{test}} \rceil$.

F.2 DETAILS OF SYNTHETIC DATA GENERATION

For an n-dimensional coefficient vector, we fix a 0–1 random matrix $B \in \{0,1\}^{n \times p}$, and generate the true coefficient vector $\boldsymbol{c}_i \in \mathbb{R}^n$ calculated from the covariate $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, I_p)$ by the following procedure. Here, $\mathcal{N}(\mathbf{0}, I_p)$ denotes the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix I_p. First, we compute the base vector $\tilde{\boldsymbol{c}}_i$ as

$$\tilde{\boldsymbol{c}}_i = \left(\frac{B\boldsymbol{x}_i}{\sqrt{p}} + 3\mathbf{1} \right)^d + \mathbf{1} \tag{63}$$

and set the final coefficient vector $\boldsymbol{c}_i$ as

$$\boldsymbol{c}_i = \frac{1}{3.5^d}\, \tilde{\boldsymbol{c}}_i \circ \boldsymbol{\epsilon}_i. \tag{64}$$

Here, $d = 4$ is the degree of the polynomial, $\circ$ denotes element-wise product, and $\mathbf{1}$ is a vector of all ones. Also, $\boldsymbol{\epsilon}_i \in \mathbb{R}^n$ is a multiplicative noise term, where each element follows a uniform distribution $\mathrm{Unif}[1-\eta, 1+\eta]$.

F.3 DETAILS OF DISTRIBUTION SHIFT GENERATION

We describe the detailed procedure for generating distribution shifts. Starting from coefficient vectors $\boldsymbol{c}_i$ generated from the noiseless distribution, we add noise according to a mixture perturbation model that combines (i) a covariate-dependent bias term and (ii) covariate-independent random noise. For the covariate-dependent perturbation, we define a polynomial map $g_\gamma : \mathbb{R}^p \to \mathbb{R}^n$ using a binary random matrix $B \in \{0,1\}^{n\times p}$ and degree $d_\gamma \in \mathbb{N}$, and set

$$\boldsymbol{c}_{i,\mathrm{obs}} = \boldsymbol{c}_i + \alpha_\gamma\, g_\gamma(\boldsymbol{x}_i).$$

For the covariate-independent perturbation, we sample

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}\big(\mathbf{0},\ \sigma^2_{\mathrm{global}} I_n\big)$$

and set

$$\boldsymbol{c}_{i,\mathrm{obs}} = \boldsymbol{c}_i + \boldsymbol{\varepsilon}_i.$$

We switch between these two perturbations with probability π_{mix} to obtain the final observed coefficient vector. The specific parameter values used in the experiments (e.g., degree d_γ, coefficient α_γ, noise variance σ_{global}) are listed in Section F.1.

F.4 HYPERPARAMETER CANDIDATES

In the hyperparameter search performed using the validation dataset in our experiments, the following parameter sets were considered as candidates.

- SP:

$$\kappa \in \{2.0,\ 7.5,\ 12.5,\ 20.0,\ 30.0\}, \tag{65}$$
$$\rho \in \{1.0,\ 3.0,\ 5.0,\ 7.0,\ 9.0\}. \tag{66}$$

- KP:

$$\kappa \in \{1.5,\ 4.0,\ 9.0,\ 15.0,\ 25.0\}, \tag{67}$$
$$\rho \in \{10.0,\ 14.0,\ 22.0,\ 28.0,\ 48.0\}. \tag{68}$$

- TSP:

$$\kappa \in \{1.0,\ 3.0,\ 8.0,\ 18.0,\ 30.0,\ 50.0\}, \tag{69}$$
$$\rho \in \{10.0,\ 24.0,\ 28.0,\ 35.0,\ 42.0,\ 65.0\}. \tag{70}$$

In the experiments on robustness against observation errors, the parameter of the budget uncertainty set $\mathcal{C}_\Gamma$ in our proposed method was set to $\Gamma = n/2$, and the identity matrix was used for the covariance matrix Σ of the ellipsoidal uncertainty set $\mathcal{C}_{\mathrm{ell}}$. The size parameters ρ and κ of the uncertainty sets were selected from the candidate sets above using the validation dataset.

In the experiments on robustness against distribution shifts, the Wasserstein ball radius ε was selected from the set $\{0.25,\ 0.5,\ 1.0,\ 2.0\}$ (for KP only $\{0.05, 0.1, 0.25,\ 0.5,\ 1.0,\ 2.0\}$), and determined based on regret on the validation data.

F.5 COMPUTATIONAL ENVIRONMENT

All numerical experiments in this study were conducted under the following computational environment.

- OS: macOS 15.3
- Processor: Apple M4
- RAM: 16GB
- Language: Python 3.9.6
- Optimization Software: Gurobi 12.0.3

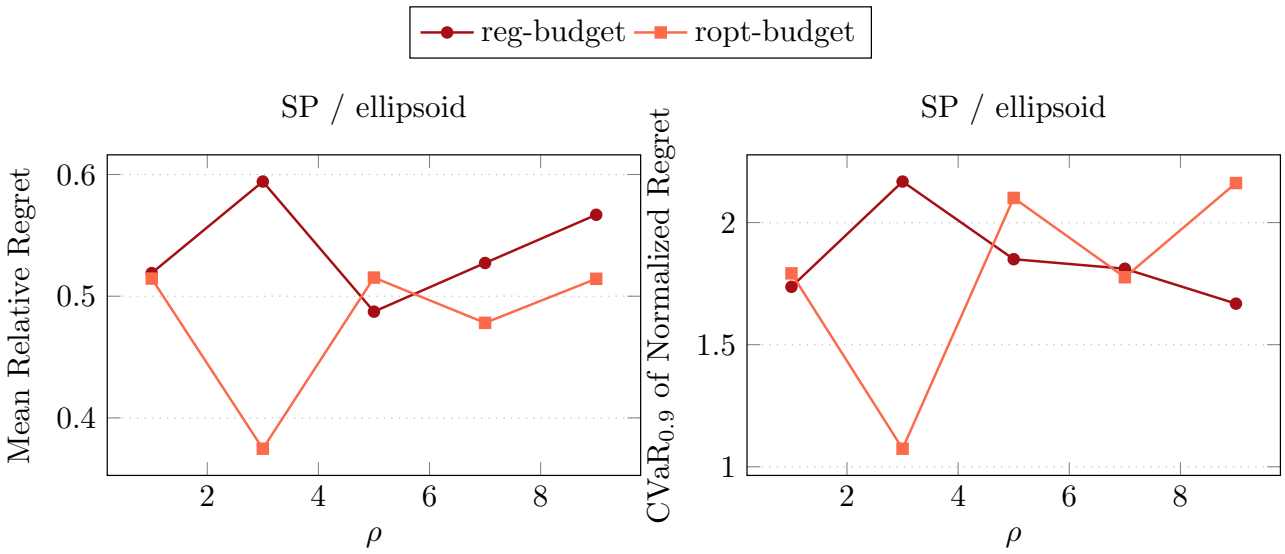

Figure 8: Sensitivity of mean relative regret and $\text{CVaR}_{0.9}$ for `ropt-budget` in SP (under ellipsoidal perturbation) (Parameter ρ).

G ADDITIONAL EXPERIMENTAL RESULTS

G.1 SENSITIVITY ANALYSIS

In this section, we describe the results of sensitivity analysis regarding the size parameters (ρ, κ) of the uncertainty sets. As shown in Figure 8, for SP (under ellipsoidal perturbation), a trade-off can be confirmed: robustness is insufficient when ρ is too small, while excessive conservativeness worsens regret when ρ is too large. As a result, $\text{CVaR}_{0.9}$ is minimized around $\rho = 3$.

Similarly, as can be seen from Figure 9, in TSP as well, regret worsens if κ is either too small or too large, and both mean relative regret and $\text{CVaR}_{0.9}$ are best around $\kappa = 18$. Note that since ρ and κ are parameters corresponding to uncertainty sets of different shapes, simple comparison on the same numerical scale is not possible. Therefore, appropriate candidate sets must be set for each problem setting and selected based on validation data. From the above results, it was shown that the proposed method can improve both mean relative regret and $\text{CVaR}_{0.9}$ in situations where observation errors exist. On the other hand, since the improvement effect depends on the choice of the uncertainty set and the setting of hyperparameters, an appropriate verification process according to the scale of errors assumed during operation is important.

Figure 10 presents a sensitivity analysis of the mean relative regret in KP. We observe a clear trade-off in the Wasserstein radius ε: when ε is too small, the method behaves similarly to `emp`, whereas an excessively large ε leads to overly conservative decisions. The choice of ground norm is also important; for example, the uncertainty set can expand rapidly under ℓ_∞, suggesting that the norm and radius should be selected to match the problem characteristics. Overall, these results confirm that distributionally robust DFL with appropriately chosen norm and radius is effective under distribution shift.

Furthermore, Figure 11 shows the results of sensitivity analysis of $\text{CVaR}_{0.9}$ with respect to the Wasserstein radius and norm in KP. Similar to the mean relative regret, although a tendency to become overly conservative is seen when the radius ε is excessive, it was confirmed that $\text{CVaR}_{0.9}$ can be effectively reduced by selecting an appropriate ε. Also, the ℓ_∞ norm has higher sensitivity to parameters compared to other norms, suggesting that appropriate parameter setting is more important.

H RELATED WORK

In this section, we review previous research related to this study. First, in Section H.1, we describe two-stage methods that separate learning and optimization. Next, in Section H.2, we explain robust optimization and distributionally robust optimization considering the uncertainty of coefficient vectors. Then, in Section H.3, we overview the research trends of Decision-Focused Learning (DFL), and finally in Section H.4, we summarize existing research on robust DFL closely

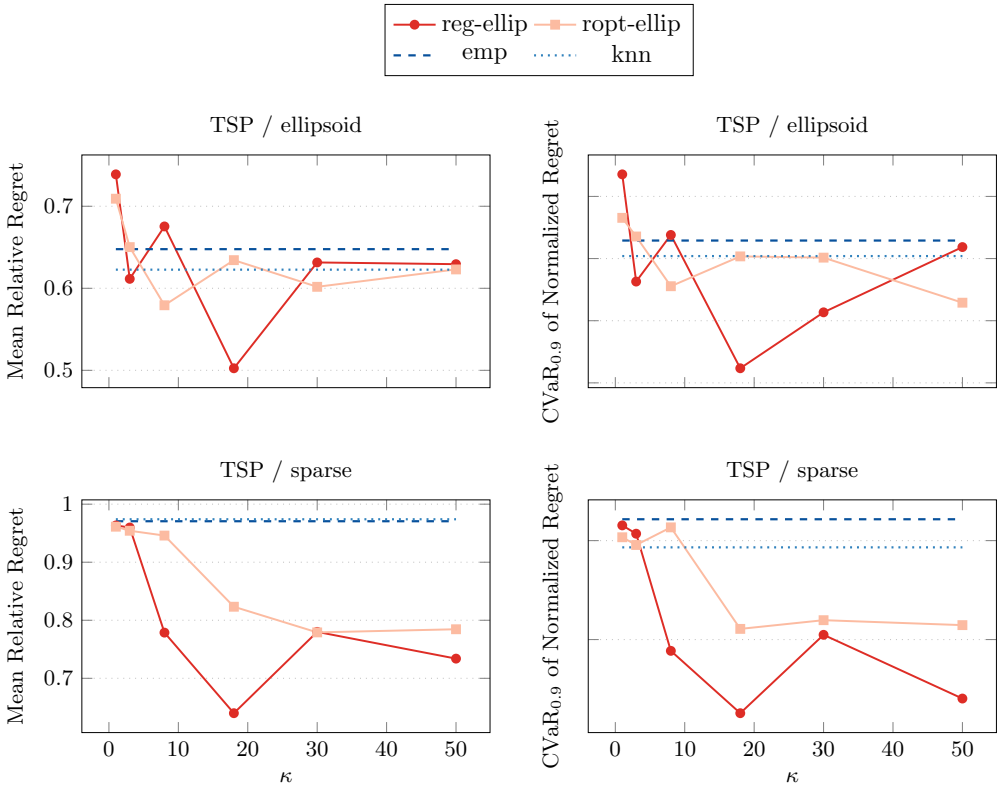

Figure 9: Sensitivity of mean relative regret and $\text{CVaR}_{0.9}$ of the loss using the ellipsoidal uncertainty set (`reg-ellip`) in TSP (Parameter κ).

related to this study.

H.1 TWO-STAGE METHODS SEPARATING LEARNING AND OPTIMIZATION

When the coefficient vector of the objective function in a mathematical optimization problem is unknown, its value needs to be predicted using observable covariates. Predict-then-Optimize takes a two-stage procedure: first predicting the coefficient vector from covariates using a predictive model, and then solving the subsequent optimization problem using the obtained predicted value. Examples include demand prediction and order quantity determination in the newsvendor problem [Huber et al., 2019, Rios et al., 2015] and prescriptive analysis based on conditional expectation estimation [Bertsimas and Kallus, 2020], which are major methods in contextual optimization [Sadana et al., 2025]. However, it has been pointed out that the minimization of the loss function (such as mean squared error) in the first learning stage does not necessarily coincide with the improvement of decision quality (i.e., minimization of regret) in the subsequent optimization problem [Sadana et al., 2025].

H.2 OPTIMIZATION PROBLEMS CONSIDERING UNCERTAINTY OF COEFFICIENT VECTORS

Robust optimization and distributionally robust optimization are known as methods to directly model the uncertainty of the coefficient vector and suppress performance degradation in the worst case. In robust optimization, it is assumed that the coefficient vector takes a value within a certain uncertainty set, and optimization is performed assuming the worst case within that set [Ben-Tal et al., 2009, Bertsimas et al., 2011, Bertsimas and Sim, 2004]. On the other hand, DRO assumes that the true distribution of the coefficient vector is unknown, and minimizes the worst-case expected value over an uncertainty set of distributions (such as sets based on moment conditions or Wasserstein distance) [Delage and Ye, 2010, Wiesemann et al., 2014, Mohajerin Esfahani and Kuhn, 2018]. Although these are usually optimization frameworks that do not include the learning process of predictive models, they provide fundamental ideas for dealing with uncertainty.

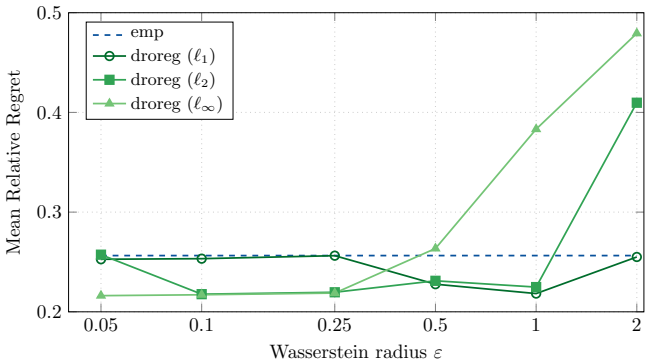

Figure 10: Sensitivity of mean relative regret to Wasserstein radius and norm in KP.

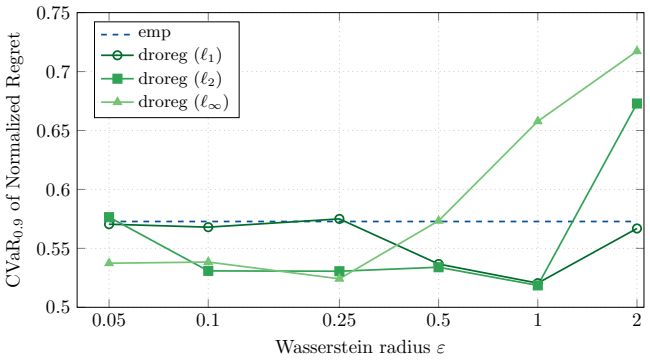

Figure 11: Sensitivity of $\text{CVaR}_{0.9}$ to Wasserstein radius and norm in KP.

H.3 DECISION-FOCUSED LEARNING

To address the mismatch between learning and optimization mentioned above, Decision-Focused Learning (DFL) has been proposed, which learns a model to directly improve the quality of the final decision (e.g., regret). DFL methods are broadly classified into two categories.

H.3.1 Methods Based on Implicit Differentiation and Continuous Relaxation

One is to treat the optimal solution of the optimization problem as a differentiable layer in a neural network and compute gradients using implicit differentiation or continuous relaxation. These include OptNet which incorporates quadratic programming as a layer [Amos and Kolter, 2017], CVXPY Layers providing convex optimization layers [Agrawal et al., 2019], and DDP simplifying dynamic programming [Mensch and Blondel, 2018]. Methods using continuous relaxation or cutting plane methods have also been proposed for MIP (Mixed Integer Programming) [Wilder et al., 2019, Ferber et al., 2020, Donti et al., 2017].

H.3.2 Methods Based on Surrogate Loss and Surrogate Gradient

The second approach is to design surrogate loss functions or surrogate gradients for regret to deal with the discontinuity of the optimal solution or when the solver is a black box. Representative methods include SPO+ loss giving a convex upper bound of regret [Elmachtoub and Grigas, 2022] and Fenchel-Young loss based on Fenchel duality [Blondel et al., 2020]. Perturbation Gradient methods approximating directional derivatives with finite differences [Huang and Gupta, 2024], methods learning the shape of the loss function [Shah et al., 2022], and gradient estimation for black-box solvers [Vlastelica et al., 2020] have also been proposed.

H.4 ROBUST DECISION-FOCUSED LEARNING

Recently, research has been progressing on enhancing the robustness of DFL against data and environmental uncertainty. For example, Schutte et al. [2024] proposed RO loss based on the concept of robust optimization, risk-averse top-k loss, and k-NN loss which performs smoothing using neighboring data to address observation errors and data scarcity. Ma et al. [2024] proposed a method to incorporate a layer solving the DRO problem into a neural network. Furthermore, Wang et al. [2026] proposed Gen-DFL, which learns the conditional distribution of coefficient vectors using a generative model and minimizes CVaR on the learned distribution.