

Integrated Learning and Robust Optimization

Cheng Tan¹, Yuchen Mao², Shuming Wang¹, and Huan Xu^{3,*}

¹School of Economics & Management, University of Chinese Academy of Sciences, China
Emails: tancheng241@mails.ucas.ac.cn, wangshuming@ucas.edu.cn

²Department of Statistics, Rutgers University, United States
Email: yuchen.mao@rutgers.edu

³Antai College of Economics and Management, Shanghai Jiao Tong University, China
Email: xuhuan_antai@sjtu.edu.cn

Many operational decisions require solving a linear program whose cost vector is unknown at decision time and must be predicted from contextual information. Because prediction and decision are only weakly aligned, the emerging integrated learning and optimization (ILO) paradigm trains the predictor through the downstream problem, judging a prediction by the decision it induces. However, predictions are inevitably imprecise, so robustness often enters the decision stage. To address this issue, we propose an integrated learning and robust optimization (ILRO) framework, where a robust decision problem is used both to define the training problem (termed the $\mathcal{RSPO}$ loss problem), and to produce the deployed decision. Thus, this framework simultaneously achieves both robustness and learning–decision alignment. To tackle its computational challenges, we develop a convex surrogate, $\mathcal{RSPO}_+$, and characterize when it is Fisher consistent. Moreover, the $\mathcal{RSPO}$ loss possesses informative gradients, allowing us to develop first-order computational methods. We also derive finite-sample excess risk bounds for both $\mathcal{RSPO}$ and $\mathcal{RSPO}_+$ predictors. Numerical experiments on transportation and portfolio problems, in comparison with multiple benchmarks, show the advantage in decision quality of the proposed framework. The gain is more pronounced for scenarios with limited samples, high-dimensional decisions, and model misspecification.

Key words: contextual stochastic optimization; decision-focused learning; robust optimization; surrogate loss; Fisher consistency; excess risk bounds

1. Introduction

Many operational decisions are made after contextual information is observed but before the relevant cost parameters are realized. In transportation planning, arc costs depend not only on observable weather, congestion, and demand signals at the time of routing, but also on unforeseen events—such as sudden traffic accidents or real-time weather shifts—that are revealed only after the vehicle has departed. Similarly, in portfolio allocation, asset returns are influenced by market covariates observed prior to the investment, yet the actual returns are determined by post-decision market shocks that are unknown at the time of allocation. Hence, such applications are naturally modeled as contextual stochastic optimization problems, where the decision is adapted to observed contextual information while the uncertain cost parameters are realized only after the decision is made.

* Corresponding author.

In this paper, we focus on a fundamental and widely used class of contextual stochastic programs with a linear objective and a context-independent feasible region. Specifically, after observing a context vector $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p$, the decision-maker chooses a feasible solution $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$, where $\mathcal{Z}$ is a convex compact feasible set. The realized cost is linear in the decision, with a random coefficient vector $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^d$ whose conditional distribution depends on the observed context $\mathbf{x}$. Under a risk-neutral objective, the benchmark decision solves

$$\min_{\mathbf{z} \in \mathcal{Z}} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\mathbf{y}^\top \mathbf{z}] = \min_{\mathbf{z} \in \mathcal{Z}} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\bar{\mathbf{y}}]^\top \mathbf{z}. \quad (1)$$

Thus, if the conditional mean $\bar{\mathbf{y}}(\mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\mathbf{y}]$ were known, the problem would reduce to a deterministic optimization problem with this conditional mean as the cost vector. In practice, however, the conditional distribution of $\mathbf{y}$, and hence its conditional mean, is unknown and must be inferred from data. Given historical samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [N]}$, where $[N] = \{1, \dots, N\}$, a standard approach is to estimate the conditional mean of $\mathbf{y}$ and then optimize using the estimate. This predict-then-optimize (PTO) pipeline is modular and easy to implement, but it is *not* designed directly for decision quality. Indeed, as emphasized in [Elmachtoub and Grigas \(2022\)](#), prediction accuracy and decision quality can be weakly aligned: two predictors with similar statistical accuracy may induce remarkably different optimal decisions.

The recent literature on contextual and prescriptive optimization formalizes this issue and develops methods that use covariates directly for decision making ([Ban and Rudin 2019](#), [Bertsimas and Kallus 2020](#)). These approaches shift the learning cost from purely statistical accuracy toward downstream operational performance, thereby providing a more decision-aware alternative to the classical PTO paradigm. Integrated learning and optimization (ILO), also called decision-focused learning, takes a further step by training predictive models through the downstream optimization problem itself ([Donti et al. 2017](#), [Wilder et al. 2019](#), [Elmachtoub and Grigas 2022](#)). Rather than evaluating a prediction solely by its discrepancy from the realized cost vector, ILO assesses predictions through the decisions they induce and the corresponding objective values. It therefore provides a framework for training prediction models that are directly tailored for downstream decision quality.

To illustrate the basic ILO framework in the present linear-objective setting, consider a parametric prediction class $\{\mathbf{g}_\theta : \theta \in \Theta\}$, where $\mathbf{g}_\theta(\mathbf{x})$ predicts the cost vector conditional on the observed context $\mathbf{x}$. For a given context $\mathbf{x}$, the predicted cost vector $\hat{\mathbf{y}} = \mathbf{g}_\theta(\mathbf{x})$ is passed to the nominal decision map, that is

$$\mathbf{z}^*(\hat{\mathbf{y}}) \in \arg \min_{\mathbf{z} \in \mathcal{Z}} \hat{\mathbf{y}}^\top \mathbf{z}, \quad (2)$$

which returns a decision that is optimal under the predicted cost vector $\hat{\mathbf{y}}$. For each training sample $(\mathbf{x}_i, \mathbf{y}_i)$, the model first predicts the cost vector $\mathbf{g}_\theta(\mathbf{x}_i)$, which induces the downstream decision

$\mathbf{z}^*(\mathbf{g}_\theta(\mathbf{x}_i))$. The quality of this decision is then evaluated using the realized cost vector $\mathbf{y}_i$, yielding the realized decision cost $\mathbf{y}_i^\top \mathbf{z}^*(\mathbf{g}_\theta(\mathbf{x}_i))$. Accordingly, the empirical ILO problem selects the model parameter by minimizing the average realized cost of the decisions induced by its predictions:

$$\theta_N^* \in \arg \min_{\theta \in \Theta} \frac{1}{N} \sum_{i \in [N]} \mathbf{y}_i^\top \mathbf{z}^*(\mathbf{g}_\theta(\mathbf{x}_i)). \quad (3)$$

After training, for a new context $\mathbf{x}$, the learned model produces the predicted cost vector $\mathbf{g}_{\theta_N^*}(\mathbf{x})$, and the corresponding prescription is obtained by solving

$$\mathbf{z}^*(\mathbf{g}_{\theta_N^*}(\mathbf{x})) \in \arg \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{g}_{\theta_N^*}(\mathbf{x})^\top \mathbf{z}. \quad (4)$$

The formulations in (3)–(4) capture the central idea of decision-focused learning: the prediction model is trained not according to prediction accuracy alone, but according to the downstream performance of the decisions induced by its predictions. Nevertheless, the framework continues to rely on a nominal linear program and does not explicitly account for uncertainty in the predicted costs. Moreover, in linear optimization, the nominal decision map may be set-valued and unstable with respect to small perturbations in the predicted cost vector. These limitations motivate our **I**ntegrated **L**earning and **R**obust **O**ptimization (ILRO) framework introduced below.

1.1. Motivation and Our Approach

The ILO framework at deployment still commits fully to the nominal decision map evaluated at a single point estimate $\hat{\mathbf{y}} = \mathbf{g}_\theta(\mathbf{x})$, so decision quality hinges on the gap $\hat{\mathbf{y}} - \bar{\mathbf{y}}(\mathbf{x})$ —and decision-focused training does not close this gap. Sampling variability persists in the data-scarce regimes that motivate ILO in the first place; misspecification and omitted covariates leave a bias that no training criterion removes; distributional shift—drift in the covariates or in the conditional law of $\mathbf{y}$ given $\mathbf{x}$ —biases even a decision-optimally trained predictor at deployment; and the realized cost stays random regardless. Worse, the nominal decision map in ILO amplifies these gaps: it is piecewise constant, so near a switching boundary a small prediction error sends the decision to a distant vertex. Robustness must therefore enter the decision stage itself. By accepting a controlled loss in nominal optimality, robust optimization hedges against a neighborhood of plausible cost vectors rather than a single prediction (Ben-Tal and Nemirovski 1999, Bertsimas and Sim 2004, Ben-Tal et al. 2009, Lu and Shen 2021).

To incorporate robustness in the ILO framework, the most direct way is post hoc: train the predictor under the loss induced by the nominal decision map, then replace it with its robust counterpart at deployment. This pipeline is modular and easy to implement, but it optimizes the predictor for a decision rule that is never used—the model is trained for the nominal decision map yet deployed through a different, robust one, and a predictor that excels for the former need

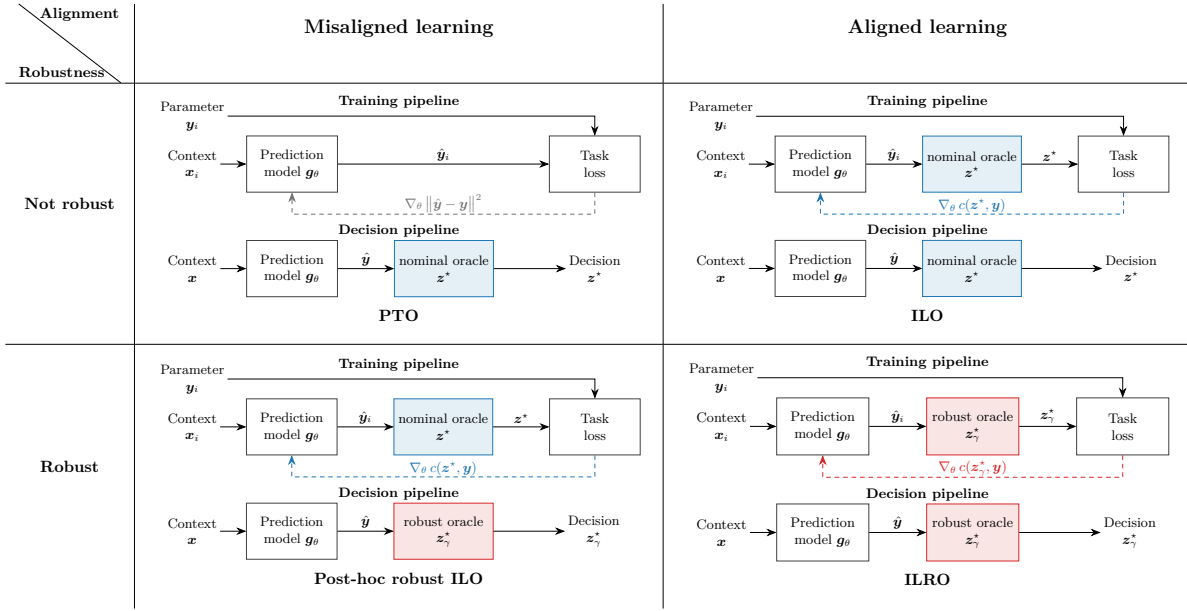


Figure 1 Pipelines of four methods.

not perform well for the latter. This mismatch motivates an integrated framework in which the same robust decision map generates the deployed decisions and defines the training loss. Figure 1 compares the four resulting pipelines—PTO, ILO, post-hoc robust ILO, and the proposed ILRO—and Table 1 locates them along the two design dimensions of robust decision-making and learning–decision alignment. ILRO is the only framework that captures both design dimensions.

Table 1 Comparison of different frameworks.

Method	Robust decision-making	Learning–decision alignment
PTO	✗	✗
ILO	✗	✓
Post-hoc robust ILO	✓	✗
ILRO	✓	✓

In this paper, we formalize the ILRO framework by developing the following robust learning–decision aligned model components. For a predicted cost vector $\hat{\mathbf{y}}$ and a radius ϑ , the robust decision problem

$$\arg \min_{\mathbf{z} \in \mathcal{Z}} \max_{\|\mathbf{u}\|_2 \leq \vartheta} \{(\hat{\mathbf{y}} + \mathbf{u})^\top \mathbf{z}\}$$

is equivalent to

$$\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \arg \min_{\mathbf{z} \in \mathcal{Z}} \hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2, \quad (\text{ILRO-DECISION})$$

up to a change of hyperparameter ($\vartheta \leftrightarrow \gamma$). See Remark 1 for a detailed explanation. Given a prediction class $\{g_\theta : \theta \in \Theta\}$, ILRO trains the predictor through this robust decision map:

$$\theta_{\gamma, N}^* \in \arg \min_{\theta \in \Theta} \frac{1}{N} \sum_{i \in [N]} \mathbf{y}_i^\top \mathbf{z}_\gamma^*(g_\theta(\mathbf{x}_i)). \quad (\text{ILRO-LEARNING})$$

For a new context $\mathbf{x}$, the deployed decision is then

$$\mathbf{z}_\gamma^*(\mathbf{g}_{\theta^*_{\gamma,N}}(\mathbf{x})) \in \arg \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{z}^\top \mathbf{g}_{\theta^*_{\gamma,N}}(\mathbf{x}) + \frac{\gamma}{2} \|\mathbf{z}\|_2^2.$$

ILRO differs from the ILO framework in that the former replaces the nominal linear program with a robust decision problem and uses this same decision problem to define the learning objective. Thus, ILRO aligns learning and deployment by training the predictor for the same robust decision rule. Throughout the paper, we take the hyperparameter γ that reflects the robustness attitude of the decision-maker as exogenously given.

In addition to aligning learning and decision within the robustness modeling, we also emphasize that compared with ILO, the ILRO framework offers a *structural advantage*—continuity of the solution mapping $\mathbf{z}_\gamma^*(\cdot)$. Notice that the nominal decision map in ILO is typically set-valued (e.g., for a polyhedral feasible set), and its selected solution is piecewise constant as a function of the predicted cost vector, providing little useful gradient information for directly optimizing the ILO objective. As illustrated in Figure 2, two nearby predictions lying on opposite sides of a boundary, such as $\hat{\mathbf{y}}^-$ and $\hat{\mathbf{y}}^+$, select different extreme points, inducing a discontinuous change in the selected decision. In contrast, our framework smooths the sharp geometric structure of the nominal linear program—as it admits the projection representation $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \Pi_{\mathcal{Z}}(-\hat{\mathbf{y}}/\gamma)$ —and hence replaces abrupt extreme-point selection with a continuous decision map. This benefits both the algorithm design of the optimization procedure and the analysis of the decision-focused learning performance.

These alignment and structural properties make ILRO a promising framework. Because the training loss is now defined through the robust decision problem in [ILRO-DECISION](#), it evaluates predictions via the realized cost of a decision given by the optimization problem [ILRO-DECISION](#), which is generally *nonconvex* in the prediction—as is the ILO objective of the nominal decision-focused learning that shares this difficulty. Nevertheless, incorporating the robustness actually adds no difficulty of its own here; instead, it offers the appealing structural advantage that mitigates the difficulty. We therefore develop two complementary learning approaches: a tractable convex surrogate, and a gradient-based method that descends the original objective directly by exploiting the differentiability of the robust decision map, an option that the nominal *SPO* loss does not admit. We then establish theoretical guarantees for the resulting loss and its surrogate, and evaluate the framework through numerical experiments.

1.2. Summary of Main Contributions

The main contributions are summarized as follows.

An integrated learning and robust optimization framework. We develop an ILRO framework for contextual linear optimization problems with context-dependent objective coefficients, in

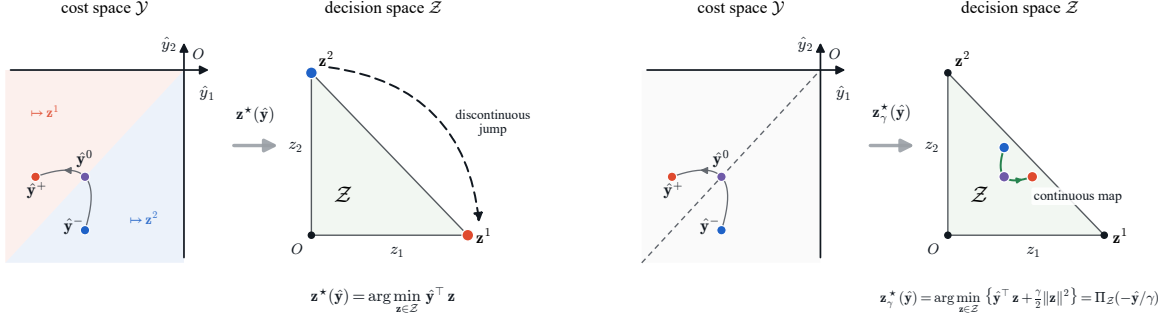


Figure 2 Comparison of solution paths of nominal and robust decision maps. The left panel shows the nominal decision map: nearby predictions $\hat{y}^-$ and $\hat{y}^+$ are mapped to different extreme points of $\mathcal{Z}$, producing a discontinuous decision map $z^*(\cdot)$. The right panel shows the robust decision map: nearby predictions are mapped to nearby feasible decisions, yielding a continuous decision map $z_\gamma^*(\cdot)$.

which the same robust decision problem generates the deployed decisions and defines the training loss—the $\mathcal{RSPO}$ loss. The deployed decision problem is equivalent to a robust optimization problem up to a change of hyperparameter (Remark 1), where γ represents the robustness attitude of the decision-maker. The ILRO framework thus simultaneously achieves decision robustness and the alignment of learning and decision objectives.

Two complementary training schemes. To solve the resulting learning problem, we develop two complementary training schemes. First, we construct a convex surrogate loss, $\mathcal{RSPO}_+$, which upper-bounds the $\mathcal{RSPO}$ loss, and is convex in the prediction. Moreover, $\mathcal{RSPO}_+$ admits an explicit gradient, in contrast to $\mathcal{SPO}_+$, which only admits subgradients in general. For polyhedral feasible sets, its empirical minimization admits a finite-dimensional reformulation (Theorem 1), which becomes a convex program under linear predictors and standard convex regularizers. Second, we exploit the differentiability of the robust decision map, whose Jacobian we derive in closed form, to refine the $\mathcal{RSPO}_+$ solution by descending the original $\mathcal{RSPO}$ objective directly via a local first-order scheme. This is unavailable for the nominal $\mathcal{SPO}$ loss, whose decision map is piecewise constant and carries no useful first-order information.

Fisher consistency: sufficient conditions. In contrast to the $\mathcal{SPO}$ setting, where Fisher consistency of the surrogate holds under mild distributional conditions for every feasible region, the population minimizer of the $\mathcal{RSPO}_+$ surrogate is given by $2\bar{\mathbf{y}}(\mathbf{x})$ (with $a = 1$), whereas the set of target-risk minimizers explicitly depends on γ . We show that Fisher consistency holds whenever γ does not exceed a threshold $\bar{\gamma}(\mathbf{x})$ defined by an explicitly computable linear program (Theorem 2), and we show that when Fisher consistency fails, surrogate training can converge to a predictor that remains suboptimal under the target $\mathcal{RSPO}$ risk, resulting in a nonvanishing asymptotic bias. As a

byproduct, our analysis at the boundary $\gamma = 0$ over bounded polyhedra (Corollary 2) also enriches the Fisher consistency results of nominal $\mathcal{SPO}_+$ established by Elmachoub and Grigas (2022).

Theoretical guarantees. We establish finite-sample excess-risk guarantees for both target and surrogate predictors. For the $\mathcal{RSPO}$ predictor, we derive a meta bound that depends on the robust decision problem through its Lipschitz factor of order $1/\gamma$ and the prediction class through its Rademacher complexity (Theorem 3); specializing the latter yields $N^{-1/2}$ rates (up to logarithmic factors) for four representative classes, including infinite-dimensional ones (Appendix D). For the $\mathcal{RSPO}_+$ predictor, within the Fisher-consistent regime, the target excess risk converges at the rate $N^{-1/4}$ (Theorem 4), improving to $N^{-1/2}$ under an extra quadratic growth condition (Corollary 3). Both analyses leverage the Lipschitz property of the prediction loss, which follows from the projection representation of the robust decision map. Our bounds scale as $1/\gamma$, which means that when the formulation is more robust (larger γ), fewer samples are required to achieve the same excess risk bound. This suggests that, when data are limited, the decision-maker should seek more robust decisions.

Empirical validation and insights. Numerical experiments on capacitated transportation and portfolio optimization problems—with benchmarks of standard $\mathcal{SPO}_+$, post-hoc robustification, and least-squares—show that $\mathcal{RSPO}_+$ attains the lowest or a comparable normalized decision loss across the reported settings, with its clearest advantages under scenarios of limited samples, high-dimensional decisions, and stronger model misspecification. Two findings merit emphasis. First, the decision gains of $\mathcal{RSPO}_+$ frequently occur despite *higher* relative prediction loss than the benchmarks, consistent with an aligned training loss prioritizing decision-relevant accuracy over aggregate prediction accuracy. Second, its comparison with the post-hoc robust benchmark isolates the value of alignment itself: both pipelines tune over the same class of robust decision maps and differ only in whether the training loss is defined through that decision map. For practitioners, these results indicate that robustness should be embedded in the learning objective itself rather than appended after a predictor has been trained for a nominal problem. Moreover, our empirical results show that gradient-based refinement of the $\mathcal{RSPO}_+$ solution yields a further improvement in out-of-sample decision quality.

Organization. The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 formalizes the ILRO framework: it establishes the well-posedness and projection representation of the robust decision map, defines the $\mathcal{RSPO}$ loss through this map, and states the empirical learning problem. Section 4 develops two complementary training schemes: the convex surrogate $\mathcal{RSPO}_+$ loss, together with its finite-dimensional reformulation (Section 4.1) and Fisher consistency result characterizing the admissible range of the robustness parameter γ (Section 4.2),

and a gradient-based method that directly refines the empirical $\mathcal{RSP}\mathcal{O}$ objective (Section 4.3). Section 5 establishes finite-sample guarantees: excess risk bounds for the $\mathcal{RSP}\mathcal{O}$ predictor via a meta generalization bound (Section 5.1) and target excess risk bounds for the $\mathcal{RSP}\mathcal{O}_+$ predictor within the Fisher-consistent regime (Section 5.2). Section 6 reports numerical experiments on capacitated transportation and portfolio optimization problems, and Section 7 concludes. The electronic companion collects the technical lemmas and proofs (Appendices A and B), the stochastic gradient-descent algorithm (Appendix C), concrete excess risk bounds for four representative hypothesis classes (Appendix D), and additional experimental details (Appendix E).

2. Literature Review

The learning–decision pipelines introduced in Section 1 belong to the contextual stochastic optimization literature surveyed by [Sadana et al. \(2025\)](#). To position ILRO, we first distinguish predict-then-optimize methods from integrated learning and optimization according to whether the prediction model is trained independently of or through the downstream problem. We then discuss robust optimization and efforts to incorporate robustness into learning–decision pipelines.

2.1. Predict-then-Optimize Methods

Predict-then-optimize methods first fit a prediction model using a statistical criterion and then pass its output to the downstream problem. In the linear-objective setting considered here, a point predictor estimates the conditional mean cost vector. Other implementations approximate the conditional distribution from contextual data. [Bertsimas and Kallus \(2020\)](#) use nonparametric methods to construct context-dependent weights for a weighted sample average approximation and establish asymptotic optimality for their k -nearest-neighbor and kernel-based prescriptions. [Kannan et al. \(2025\)](#) combine fitted predictions with empirical residuals to generate scenarios for sample average approximation and provide asymptotic and finite-sample guarantees.

Theoretical studies have compared PTO and ILO in terms of downstream regret. For contextual linear optimization, [Hu et al. \(2022\)](#) compare the naive plug-in approach with its integrated counterpart over the same prediction class. Under correct specification of that class and a condition limiting near-dual-degeneracy, they show that the plug-in approach can attain faster regret rates. For a general class of nonlinear stochastic optimization problems with parametric distribution models, [Elmachtoub et al. \(2023\)](#) compare PTO with ILO. They show that PTO has stochastically smaller asymptotic regret when the model class is well specified and sufficient data are available, whereas integrated estimation can perform better under misspecification. These results show that the relative performance of the two approaches depends on problem structure and model specification.

2.2. Integrated Learning and Optimization

Integrated learning and optimization, also referred to as decision-focused learning, trains the prediction model through the downstream optimization problem rather than through a separate statistical criterion. For contextual linear optimization, our work builds most directly on [Elmachtoub and Grigas \(2022\)](#), who introduce the $\mathcal{SPO}$ loss, derive its convex upper bound $\mathcal{SPO}_+$, and establish Fisher consistency under distributional and geometric conditions. Subsequent work develops statistical guarantees for both the target and surrogate losses. [El Balghiti et al. \(2023\)](#) derive generalization bounds for the $\mathcal{SPO}$ loss using margin and combinatorial-complexity arguments. For $\mathcal{SPO}_+$, [Liu and Grigas \(2021\)](#) establish risk bounds and calibration results that transfer surrogate excess risk to $\mathcal{SPO}$ excess risk. Beyond these loss-specific results, [Ho-Nguyen and Kilinc-Karzan \(2022\)](#) identify conditions under which prediction losses yield asymptotic consistency and nonasymptotic guarantees for optimization risk. The key difference between the ILO and ILRO frameworks is that the former uses a nominal decision problem, whereas the latter uses a robust decision problem. At the same time, ILRO essentially preserves the learning–decision alignment of the ILO framework, as each framework uses its deployed decision rule to define the training loss. Replacing the nominal decision map with a robust one leads to a distinct learning problem, for which we provide a comprehensive computational and statistical analysis. On the computational side, we construct the convex $\mathcal{RSPO}_+$ surrogate and exploit the differentiability of the robust decision map to develop a first-order method for optimizing the $\mathcal{RSPO}$ loss directly. On the statistical side, we characterize an explicit robustness threshold for the surrogate’s Fisher consistency and establish finite-sample excess-risk guarantees for both predictors.

Another line develops gradient-based methods for solving ILO models through implicit differentiation under suitable regularity conditions or through smoothing techniques. [Amos and Kolter \(2017\)](#) differentiate the KKT system of a quadratic program, and [Donti et al. \(2017\)](#) use this implicit-differentiation approach to train probabilistic models through stochastic programs. Among smoothing methods, [Berthet et al. \(2020\)](#) use stochastic perturbations to construct a differentiable expected optimizer, whereas [Wilder et al. \(2019\)](#) add quadratic smoothing to a continuous relaxation of a linear combinatorial problem to obtain training gradients. Overall, these methods still focus on solving ILO problems with a nominal downstream decision map and use differentiation or smoothing to mitigate computational difficulties in training. We refer to [Mandi et al. \(2024\)](#) for a survey of decision-focused learning.

2.3. Robust Optimization in Learning–Decision Pipelines

Robust optimization addresses parameter uncertainty by optimizing against realizations in a prescribed uncertainty set ([Ben-Tal and Nemirovski 1999](#), [Bertsimas and Sim 2004](#), [Ben-Tal et al.](#)

2009) and has been used across operations management (Lu and Shen 2021). Most relevant here are models with uncertain objective coefficients, including discrete optimization and network flows with cost uncertainty (Bertsimas and Sim 2003) and portfolio models with uncertain return and risk parameters (Goldfarb and Iyengar 2003). When uncertainty concerns a probability distribution rather than individual parameter realizations, distributionally robust optimization (DRO) optimizes against the worst-case distribution in an ambiguity set consistent with the available information (Delage and Ye 2010, Wiesemann et al. 2014, Gao and Kleywegt 2023, Kuhn et al. 2025).

Several studies incorporate robustness into ILO from different perspectives. Chenreddy and Delage (2024) and Yeh et al. (2025) use downstream decision-making losses to learn uncertainty sets for conditional robust optimization rather than point predictors of uncertain objective coefficients. For objective-coefficient predictors, Schutte et al. (2024) modify the $\mathcal{SPO}$ loss by replacing the benchmark of the nominal clairvoyant optimal value with a robust counterpart. Costa and Iyengar (2023) develop a portfolio-specific DRO pipeline based on a ϕ -divergence ball. Im et al. (2025) extend $\mathcal{SPO}$ to robust constraints by replacing its fixed feasible set with a robust feasible set that enforces the constraints for every constraint-parameter realization in a separately constructed uncertainty set. They provide a convex surrogate for the resulting loss and establish the surrogate’s Fisher consistency with respect to that loss. ILRO instead robustifies the predicted objective coefficients, leading to different computational and statistical challenges.

In addition, some studies introduce robustness into the downstream decision problem within a PTO pipeline. Kannan et al. (2024) first fit a prediction model and then use its residuals to construct an empirical distribution and corresponding ambiguity sets for the downstream DRO problem. Sim et al. (2025) develop a residual-based robust satisficing framework using a fitted regression model, followed by a fortification step to account for parameter-estimation uncertainty. Further examples include Hu et al. (2025) and Mao et al. (2026). In these approaches, predictor fitting remains separate from downstream robustness.

As the studies above illustrate, robustness is important for downstream decision-making when predictions are imprecise. It is therefore natural to incorporate downstream robustness into ILO. To the best of our knowledge, ILRO is the first attempt to do so while maintaining learning–decision alignment in the presence of objective-coefficient prediction errors. We further address the computational challenges and study the statistical properties of the ILRO framework.

3. Model

Recall that in the ILRO framework proposed in Section 1, the robust decision map **ILRO-DECISION** generates the deployed decision, and the learning problem **ILRO-LEARNING** trains the predictor through that same decision map. What **ILRO-LEARNING** minimizes is the average realized cost

$\mathbf{y}_i^\top \mathbf{z}_\gamma^*(\mathbf{g}_\theta(\mathbf{x}_i))$ of the induced decisions. This section formalizes this criterion as a loss function on prediction–realization pairs, following the route taken by [Elmachtoub and Grigas \(2022\)](#) for the nominal decision map.

We begin by recalling their construction. For the ILO framework with a linear cost function, [Elmachtoub and Grigas \(2022\)](#) propose the smart predict-then-optimize ($\mathcal{SPO}$) loss, defined as

$$\ell_{\mathcal{SPO}}(\hat{\mathbf{y}}, \mathbf{y}) := \max_{\mathbf{z} \in \mathbf{Z}^*(\hat{\mathbf{y}})} \mathbf{y}^\top \mathbf{z} - v^*(\mathbf{y}),$$

where $\mathbf{Z}^*(\mathbf{y}) = \arg \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{y}^\top \mathbf{z}$ denotes the optimal solution set. When this set is a singleton, or when a fixed selection rule is understood, we denote the corresponding optimizer by $\mathbf{z}^*(\mathbf{y})$. The associated optimal value is denoted by $v^*(\mathbf{y}) = \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{y}^\top \mathbf{z}$. The loss function characterizes the discrepancy between a predicted parameter $\hat{\mathbf{y}}$ and the true parameter $\mathbf{y}$ through their induced decisions. The inner maximization selects the worst decision from the optimal set under $\hat{\mathbf{y}}$ and evaluates its performance under $\mathbf{y}$, thereby quantifying the decision loss caused by prediction error. Based on the $\mathcal{SPO}$ loss, the empirical risk minimization problem takes the form

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i \in [N]} \ell_{\mathcal{SPO}}(\mathbf{g}_\theta(\mathbf{x}_i), \mathbf{y}_i) = \min_{\theta \in \Theta} \frac{1}{N} \sum_{i \in [N]} \left(\max_{\mathbf{z} \in \mathbf{Z}^*(\mathbf{g}_\theta(\mathbf{x}_i))} \mathbf{y}_i^\top \mathbf{z} - v^*(\mathbf{y}_i) \right),$$

which is equivalent to the learning part of the ILO framework up to the choice of the optimal decision. Since the $\mathcal{SPO}$ loss is generally nonconvex and computationally difficult to minimize, [Elmachtoub and Grigas \(2022\)](#) propose $\mathcal{SPO}_+$ as a convex surrogate.

Following a similar spirit, we define a robust counterpart of the $\mathcal{SPO}$ loss that evaluates predictions through the robust decision map in [ILRO-DECISION](#). Lemma 1 presents several elementary properties of that map, which formalize the structural advantage discussed in Section 1 and are used repeatedly in Sections 4 and 5.

LEMMA 1 (Well-Posedness). *Let $\gamma > 0$ and let $\mathcal{Z} \subseteq \mathbb{R}^d$ be a nonempty compact convex set. Then, for every $\mathbf{y}_0 \in \mathbb{R}^d$, the objective of the robust decision problem in [ILRO-DECISION](#) is strongly convex in $\mathbf{z}$, so the problem admits a unique minimizer $\mathbf{z}_\gamma^*(\mathbf{y}_0)$. Moreover, completing the square yields the projection representation*

$$\mathbf{z}_\gamma^*(\mathbf{y}_0) = \Pi_{\mathcal{Z}} \left(-\frac{\mathbf{y}_0}{\gamma} \right), \tag{5}$$

where $\Pi_{\mathcal{Z}}(\cdot)$ denotes the Euclidean projection onto $\mathcal{Z}$. Thus, $\mathbf{z}_\gamma^*(\cdot)$ is Lipschitz continuous with constant $1/\gamma$ and differentiable almost everywhere.

Since for given $\mathbf{y}_0$, the minimizer $\mathbf{z}_\gamma^*(\mathbf{y}_0)$ is unique when $\gamma > 0$, we henceforth write $=$ in place of $\in$. We also write

$$v_\gamma^*(\mathbf{y}_0) := \min_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}_0^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 \right\}$$

for the corresponding optimal value.

DEFINITION 1 (*RSPOL* LOSS). Given a vector prediction $\hat{\mathbf{y}}$ and a realized cost vector $\mathbf{y}$, the *RSPOL* loss (short for Robust *SPOL* loss) is defined by

$$\ell_{\mathcal{RSPOL}}(\hat{\mathbf{y}}, \mathbf{y}) := \mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) - v^*(\mathbf{y}).$$

The *RSPOL* loss is well-defined, since the robust decision map $\mathbf{z}_\gamma^*(\cdot)$ is single-valued by Lemma 1. It measures the excess realized cost of the robust decision induced by $\hat{\mathbf{y}}$ under the realized cost $\mathbf{y}$, relative to a benchmark being the *nominal* clairvoyant optimal value $v^*(\mathbf{y})$. Notice that minimizing the empirical *RSPOL* risk is equivalent to the learning problem *ILRO-LEARNING*, regardless of whether the benchmark is taken as $v^*(\mathbf{y})$ or $v_\gamma^*(\mathbf{y})$. We intentionally choose $v^*(\mathbf{y})$ as the benchmark, since it ensures $\ell_{\mathcal{RSPOL}} \geq 0$, which facilitates our analysis of the excess risk in Section 5.¹

Given an i.i.d. sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [N]}$, we define the empirical *RSPOL* learning problem as

$$\min_{\mathbf{g} \in \mathcal{G}} \frac{1}{N} \sum_{i \in [N]} \ell_{\mathcal{RSPOL}}(\mathbf{g}(\mathbf{x}_i), \mathbf{y}_i) + \lambda \Omega(\mathbf{g}), \quad (6)$$

where $\mathcal{G} = \{\mathbf{g}_\theta : \theta \in \Theta\}$ is the hypothesis class and $\lambda \Omega(\mathbf{g})$ is the regularization term with $\lambda \geq 0$. Here, $\Omega : \mathcal{G} \rightarrow [0, \infty)$ is a convex regularizer and for a parametric class we write $\Omega(\mathbf{g}_\theta)$.

Problem (6) completes the ILRO formulation by training the predictor according to the same robust decision map used at deployment. The resulting problem is nonconvex, as is the *SPOL* objective it generalizes. We will discuss the computational aspects in the next section. We now turn to some discussions on the components of the ILRO framework.

REMARK 1 (ROBUSTNESS INTERPRETATION OF *ILRO-DECISION*). We start from the classical robust problem with a ball uncertainty set below

$$\min_{\mathbf{z} \in \mathcal{Z}} \max_{\|\mathbf{u}\|_2 \leq \vartheta} \{(\hat{\mathbf{y}} + \mathbf{u})^\top \mathbf{z}\},$$

which is equivalent to

$$\min_{\mathbf{z} \in \mathcal{Z}} \hat{\mathbf{y}}^\top \mathbf{z} + \vartheta \|\mathbf{z}\|_2.$$

Both this problem and *ILRO-DECISION* penalize the magnitude $\|\mathbf{z}\|_2$ of the decision, and the two families are equivalent at the level of solution paths: for any $\vartheta \geq 0$, every solution of the problem above solves *ILRO-DECISION* for some $\gamma \in [0, \infty]$, and vice versa. Moreover, along this correspondence a larger γ corresponds to a larger ϑ , thereby yielding a more robust formulation of the downstream decision problem. We formalize these claims in Lemma EC.1 in Appendix A. When $\gamma = 0$, *ILRO-DECISION* reduces to the nominal ILO counterpart. $\square$

¹ In contrast, if $v_\gamma^*(\mathbf{y})$ is chosen as the benchmark, the loss may take negative values as it measures the regret against a conservative benchmark.

REMARK 2 (ROBUSTIFICATION THROUGH A GENERAL STRONGLY CONVEX PENALTY). The ℓ_2 ball in Remark 1 is not essential. Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be closed and μ -strongly convex with $\mu > 0$ and let ϕ^* be its convex conjugate. Since $\max_{\mathbf{u}} \{\mathbf{u}^\top \mathbf{z} - \phi^*(\mathbf{u})\} = \phi(\mathbf{z})$, the decision map

$$\mathbf{z}_\phi^*(\hat{\mathbf{y}}) := \arg \min_{\mathbf{z} \in \mathcal{Z}} \max_{\mathbf{u} \in \mathbb{R}^d} \{(\hat{\mathbf{y}} + \mathbf{u})^\top \mathbf{z} - \phi^*(\mathbf{u})\} = \arg \min_{\mathbf{z} \in \mathcal{Z}} \{\hat{\mathbf{y}}^\top \mathbf{z} + \phi(\mathbf{z})\}$$

is also an exact robust counterpart of the nominal problem, with an adversary perturbing the predicted cost by $\mathbf{u}$ at the price $\phi^*(\mathbf{u})$. The choice $\phi = \frac{\gamma}{2} \|\cdot\|_2^2$, hence $\phi^* = \frac{1}{2\gamma} \|\cdot\|_2^2$, recovers **ILRO-DECISION**, which leads to another robust optimization interpretation of **ILRO-DECISION**.

We remark that all our results related to **ILRO-DECISION** apply to this general robustification model with minimal modifications ($1/\gamma$ replaced by $1/\mu$). Notice that ϕ is μ -strongly convex, which implies $\nabla \phi^*$ is $(1/\mu)$ -Lipschitz. Thus, if ϕ^* is conic representable, ϕ is twice differentiable, and L -smooth, all technical results hold. The details are straightforward yet tedious, and hence we omit them. Accordingly, we model robustness through the ℓ_2 ball throughout the paper, as it captures all essential properties of the framework. $\square$

REMARK 3 (STRUCTURAL PROPERTY OF THE LEARNING OBJECTIVE). Before concluding this section, we present a structural property of the empirical learning problem with $\mathcal{RSPOL}$ loss, which may be of independent interest.

PROPOSITION 1 (**Continuous Piecewise Affine Objective**). *Let $\gamma > 0$, let $\mathcal{Z} = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}$ be a nonempty bounded polyhedron with $\mathbf{A} \in \mathbb{R}^{m \times d}$, $\mathbf{b} \in \mathbb{R}^m$, and let the prediction class be linear, $\mathbf{g}_\theta(\mathbf{x}) = \mathbf{B}\mathbf{x}$ with $\theta = \mathbf{B} \in \mathbb{R}^{d \times p}$. Then the unregularized ($\lambda = 0$) empirical $\mathcal{RSPOL}$ objective*

$$\mathbf{B} \mapsto \frac{1}{N} \sum_{i \in [N]} [\mathbf{y}_i^\top \mathbf{z}_\gamma^*(\mathbf{B}\mathbf{x}_i) - v^*(\mathbf{y}_i)]$$

is continuous and piecewise affine: there is a finite family of full-dimensional polyhedra covering $\mathbb{R}^{d \times p}$ on the interior of each of which the objective is affine, with a constant gradient.

To better understand this structural result, consider the case of $\mathcal{SPOL}$. Here, since $\mathbf{z}^*(\mathbf{B}\mathbf{x}_i)$ belongs to the finite set of extreme points, the resulting objective is piecewise constant and hence discontinuous (except in the trivial degenerate case). Thus, Proposition 1 explains why the ILRO framework is more amenable to first-order methods in computation, since its objective is continuous and has informative gradient almost everywhere. Moreover, the piecewise affine structure may inspire design of approximation algorithms, which we leave for future study. $\square$

4. Training Schemes and Fisher Consistency

In this section, we develop two complementary schemes to address the computational difficulty of ILRO. The first scheme, in Section 4.1, replaces the target loss with a convex and differentiable surrogate, the $\mathcal{RSP}\mathcal{O}_+$ loss. We develop the Fisher consistency theory of this surrogate in Section 4.2, which identifies the regime of γ where minimizing the surrogate also minimizes the target risk.

The second scheme, in Section 4.3, exploits the structural advantage of the robust decision map discussed in Section 1—the decision map is Lipschitz continuous and differentiable almost everywhere—to apply gradient-descent-type algorithms directly to the empirical $\mathcal{RSP}\mathcal{O}$ learning problem (6). We can also initialize the gradient approach using the surrogate solution.

4.1. Convex Surrogate: $\mathcal{RSP}\mathcal{O}_+$ Loss

Throughout Sections 4.1 and 4.2, we let $\gamma > 0$ and let $\mathcal{Z} \subseteq \mathbb{R}^d$ be a nonempty compact convex set, unless otherwise specified. We construct a convex surrogate for the generally nonconvex $\mathcal{RSP}\mathcal{O}$ loss in three steps: we first reformulate the target loss as the limit of a family of penalized problems indexed by a scalar $a > 0$; we then fix a finite a , which yields an upper bound; finally, we linearize the concave value function $v_\gamma^*(\cdot)$ at the realized cost, which leads to our proposed surrogate loss that is convex with respect to the prediction.

For the first step, fix $\hat{\mathbf{y}}$ and $\mathbf{y}$ and define

$$q(a) := \max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - a \hat{\mathbf{y}}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + a v_\gamma^*(\hat{\mathbf{y}}), \quad a > 0. \quad (7)$$

The following proposition connects the function q to the target loss.

PROPOSITION 2 (Limiting Representation). *For fixed $\hat{\mathbf{y}}$ and $\mathbf{y}$, the function $q(\cdot)$ is convex and nonincreasing on $(0, \infty)$, and the $\mathcal{RSP}\mathcal{O}$ loss admits the representation*

$$\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) = \lim_{a \rightarrow +\infty} q(a) - v^*(\mathbf{y}).$$

Since q is nonincreasing, fixing any finite $a > 0$ in Proposition 2 yields the upper bound $\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) \leq q(a) - v^*(\mathbf{y})$, which completes the second step. This bound, however, is not yet convex in the prediction: the maximization term is convex in $\hat{\mathbf{y}}$, but the term $a v_\gamma^*(\hat{\mathbf{y}})$ is concave. The third step removes the concave term by linearization. Since the robust problem has a unique minimizer, $v_\gamma^*(\cdot)$ is differentiable with $\nabla v_\gamma^*(\mathbf{y}) = \mathbf{z}_\gamma^*(\mathbf{y})$ by Danskin's theorem, and concavity gives $v_\gamma^*(\hat{\mathbf{y}}) \leq v_\gamma^*(\mathbf{y}) + \mathbf{z}_\gamma^*(\mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y})$. Combining the three steps,

$$\begin{aligned} \ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) &\leq \max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - a \hat{\mathbf{y}}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + a v_\gamma^*(\hat{\mathbf{y}}) - v^*(\mathbf{y}) \\ &\leq \max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - a \hat{\mathbf{y}}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + a \left\{ v_\gamma^*(\mathbf{y}) + \mathbf{z}_\gamma^*(\mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y}) \right\} - v^*(\mathbf{y}) \\ &= \max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - a \hat{\mathbf{y}}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + a \left\{ \mathbf{z}_\gamma^*(\mathbf{y})^\top \hat{\mathbf{y}} + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(\mathbf{y})\|_2^2 \right\} - v^*(\mathbf{y}), \end{aligned}$$

where the equality uses $v_\gamma^*(\mathbf{y}) = \mathbf{y}^\top \mathbf{z}_\gamma^*(\mathbf{y}) + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(\mathbf{y})\|_2^2$. The resulting expression is convex in $\hat{\mathbf{y}}$ and upper-bounds the $\mathcal{RSP}\mathcal{O}$ loss, which leads to the following convex surrogate.

DEFINITION 2 ($\mathcal{RSP}\mathcal{O}_+$ LOSS). Given a vector prediction $\hat{\mathbf{y}}$, a realized cost vector $\mathbf{y}$, and a fixed scalar $a > 0$, the $\mathcal{RSP}\mathcal{O}_+$ loss is defined as

$$\ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y}) := \max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - a \hat{\mathbf{y}}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + a \left\{ \mathbf{z}_\gamma^*(\mathbf{y})^\top \hat{\mathbf{y}} + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(\mathbf{y})\|_2^2 \right\} - v^*(\mathbf{y}).$$

REMARK 4 (ROLE OF THE PARAMETER a). The two approximation steps introduce two gaps with opposite behavior in a . The penalty gap $q(a) - \lim_{a' \rightarrow \infty} q(a')$ vanishes as $a \rightarrow +\infty$, whereas the linearization gap equals a times the (a -independent) Bregman gap $v_\gamma^*(\mathbf{y}) + \mathbf{z}_\gamma^*(\mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y}) - v_\gamma^*(\hat{\mathbf{y}}) \geq 0$ and therefore grows linearly in a . The tightness of $\ell_{\mathcal{RSP}\mathcal{O}_+}$ is thus not monotone in a , and no single choice dominates pointwise. Section 4.2 singles out $a = 1$ on statistical grounds. $\square$

REMARK 5 (RELATION TO THE $\mathcal{SP}\mathcal{O}_+$ CONSTRUCTION). Setting $\gamma = 0$ in Definition 2 yields $\max_{\mathbf{z} \in \mathcal{Z}} \{\mathbf{y}^\top \mathbf{z} - a \hat{\mathbf{y}}^\top \mathbf{z}\} + a \mathbf{z}^*(\mathbf{y})^\top \hat{\mathbf{y}} - v^*(\mathbf{y})$, which for $a = 2$ is precisely the $\mathcal{SP}\mathcal{O}_+$ loss of [Elmachtoub and Grigas \(2022\)](#). The $\mathcal{RSP}\mathcal{O}_+$ loss is therefore the exact robust counterpart of the $\mathcal{SP}\mathcal{O}_+$ loss, while for $\gamma > 0$ the quadratic term additionally renders the loss differentiable, as shown next. $\square$

We next collect the analytical properties of the surrogate.

PROPOSITION 3. *Given a fixed $\mathbf{y} \in \mathcal{Y}$, it holds that:*

- (i) $0 \leq \ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) \leq \ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y})$ for all $\hat{\mathbf{y}}$;
- (ii) $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ is convex on $\mathbb{R}^d$;
- (iii) $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ is differentiable everywhere, with gradient at $\hat{\mathbf{y}}$ given by

$$a \left(\mathbf{z}_\gamma^*(\mathbf{y}) - \mathbf{z}_\gamma^* \left(\hat{\mathbf{y}} - \frac{1}{a} \mathbf{y} \right) \right);$$

- (iv) $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ is $a \cdot D_{\mathcal{Z}}$ -Lipschitz continuous, where $D_{\mathcal{Z}} := \sup_{\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}} \|\mathbf{z}_1 - \mathbf{z}_2\|_2$ denotes the diameter of the feasible set $\mathcal{Z}$.

The differentiability in (iii) follows from the uniqueness induced by the strong convexity of the decision-problem objective, in contrast to the generally nondifferentiable $\mathcal{SP}\mathcal{O}_+$ loss. This proposition is used repeatedly in the risk analysis of Section 5.

Having established the convexity and differentiability of the $\mathcal{RSP}\mathcal{O}_+$ loss, we now use it to define a surrogate learning problem. Replacing the generally nonconvex $\mathcal{RSP}\mathcal{O}$ loss in Problem (6) with $\mathcal{RSP}\mathcal{O}_+$ yields the surrogate empirical risk minimization (ERM) problem:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i \in [N]} \ell_{\mathcal{RSP}\mathcal{O}_+}(\mathbf{g}_\theta(\mathbf{x}_i), \mathbf{y}_i) + \lambda \Omega(\mathbf{g}_\theta), \quad (8)$$

where $\{\mathbf{g}_\theta : \theta \in \Theta\}$ is the prediction class, $\Omega(\cdot)$ is a regularizer on the prediction model, and $\lambda \geq 0$ controls the regularization strength. The surrogate ERM can be optimized by first-order methods

based on the gradient derived in Proposition 3, because when $\mathbf{g}_\theta$ is linear in θ , the problem (8) is convex in θ . Note that this is consistent with the convexity requirements for empirical risk minimization of the $\mathcal{RSP}\mathcal{O}_+$ loss and standard statistical losses such as squared loss. Moreover, for a general prediction class $\{\mathbf{g}_\theta : \theta \in \Theta\}$, whose members are differentiable in θ (e.g., multilayer perceptron), the problem can be trained by (stochastic) gradient algorithms.

In addition to first-order methods, we can also resort to a reformulation approach. For a bounded polyhedral feasible set, Problem (8) admits an explicit single-level reformulation through duality.

THEOREM 1 (Reformulation of the $\mathcal{RSP}\mathcal{O}_+$ ERM). *Suppose $\mathcal{Z} = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}$ is a nonempty bounded polyhedron, and fix $a > 0$ and $\gamma > 0$. Then Problem (8) is equivalent to*

$$\begin{aligned} \min_{\theta, \mathbf{p}_i} \quad & \frac{1}{N} \sum_{i \in [N]} \left\{ \frac{1}{2a\gamma} \|\mathbf{y}_i - a\mathbf{g}_\theta(\mathbf{x}_i) + \mathbf{A}^\top \mathbf{p}_i\|_2^2 - \mathbf{b}^\top \mathbf{p}_i + a\mathbf{z}_\gamma^*(\mathbf{y}_i)^\top \mathbf{g}_\theta(\mathbf{x}_i) + \frac{a\gamma}{2} \|\mathbf{z}_\gamma^*(\mathbf{y}_i)\|_2^2 - v^*(\mathbf{y}_i) \right\} + \lambda \Omega(\mathbf{g}_\theta) \\ \text{s.t.} \quad & \mathbf{p}_i \in \mathbb{R}_+^m, i \in [N], \theta \in \Theta, \end{aligned}$$

in the sense that the two problems have the same optimal value and θ^* is optimal for Problem (8) if and only if there exist $\{\mathbf{p}_i^*\}$ such that $(\theta^*, \{\mathbf{p}_i^*\})$ is optimal for the reformulation.

Similarly to Problem (8), the reformulation as given by Theorem 1 is convex when $\mathbf{g}_\theta(\mathbf{x})$ is linear in θ . For example, when $\mathbf{g}_\theta(\mathbf{x}) = \mathbf{B}\mathbf{x}$ and $\Omega(\mathbf{g}_\theta) = \|\mathbf{B}\|_1$ or $\Omega(\mathbf{g}_\theta) = \|\mathbf{B}\|_F^2$, it can be formulated as a convex quadratic program using standard epigraph reformulations.

To examine whether minimizing this tractable surrogate is statistically aligned with the target loss, we next study the Fisher consistency of $\mathcal{RSP}\mathcal{O}_+$ with respect to $\mathcal{RSP}\mathcal{O}$.

4.2. Fisher Consistency

We now examine the optimality validity of the $\mathcal{RSP}\mathcal{O}_+$ loss by asking whether minimizing this surrogate population risk also minimizes the target $\mathcal{RSP}\mathcal{O}$ risk. As we show below, this property—known as Fisher consistency (see, e.g., [Elmachtoub and Grigas 2022](#))—need not hold for all values of γ . We therefore characterize the regime of γ under which it holds. We begin with the formal definition.

DEFINITION 3. Let $\mathbb{P}$ denote the joint distribution of $(\mathbf{x}, \mathbf{y})$, and let $\mathcal{G}_{\text{all}}$ denote the class of all measurable functions mapping $\mathcal{X}$ to $\mathbb{R}^d$. A loss function $\ell(\cdot, \cdot)$ is said to be *Fisher consistent* with respect to the $\mathcal{RSP}\mathcal{O}$ loss if every minimizer of

$$\min_{\mathbf{g} \in \mathcal{G}_{\text{all}}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}} [\ell(\mathbf{g}(\mathbf{x}), \mathbf{y})] \tag{9}$$

is also a minimizer of

$$\min_{\mathbf{g} \in \mathcal{G}_{\text{all}}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}} [\ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{g}(\mathbf{x}), \mathbf{y})]. \tag{10}$$

Fisher consistency is a population-level property over the unrestricted class of measurable predictors. It rules out any intrinsic bias from replacing $\ell_{\mathcal{RSP}\mathcal{O}}$ with $\ell_{\mathcal{RSP}\mathcal{O}_+}$ when data are unlimited. Because the minimization in (9) and (10) is over $\mathcal{G}_{\text{all}}$, the interchangeability principle (e.g., Rockafellar and Wets 1998, Theorem 14.60) reduces both problems to pointwise minimization: for any loss ℓ ,

$$\min_{g \in \mathcal{G}_{\text{all}}} \mathbb{E}_{(x, y) \sim \mathbb{P}} [\ell(g(x), y)] = \mathbb{E}_{x \sim \mathbb{P}_x} \left[\min_{\hat{y} \in \mathbb{R}^d} \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [\ell(\hat{y}, y)] \right],$$

where $\mathbb{P}_x$ and $\mathbb{P}_{y|x}$ denote the marginal and conditional distributions. All statements in this subsection are understood pointwise, and the population statements follow by integrating over $\mathbb{P}_x$. For a given x , define the sets of minimizers of the pointwise target and surrogate risks,

$$\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* := \arg \min_{\hat{y} \in \mathbb{R}^d} \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [\ell_{\mathcal{RSP}\mathcal{O}}(\hat{y}, y)], \quad \mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^* := \arg \min_{\hat{y} \in \mathbb{R}^d} \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [\ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{y}, y)].$$

Fisher consistency then amounts to the inclusion $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^* \subseteq \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$ holding for $\mathbb{P}_x$ -almost every x . We characterize the two sets in turn—the target set in Proposition 4, the surrogate set in Proposition 5 and Corollary 1—and then determine when the inclusion holds.

PROPOSITION 4. *Suppose that $\mathbb{E}_{y \sim \mathbb{P}_{y|x}} [\|y\|_2] < \infty$ and the nominal problem $\min_{z \in \mathcal{Z}} \bar{y}^\top z$ at the conditional mean $\bar{y} = \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y]$ admits a unique optimal solution $z^*(\bar{y})$, and write $\mathcal{N}_{\mathcal{Z}}(z^*(\bar{y}))$ for the normal cone to $\mathcal{Z}$ at $z^*(\bar{y})$. Then*

$$\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* = -\mathcal{N}_{\mathcal{Z}}(z^*(\bar{y})) - \gamma z^*(\bar{y}). \quad (11)$$

In particular, if $\mathcal{Z} = \{z \in \mathbb{R}^d : \mathbf{A}z \geq \mathbf{b}\}$ is a nonempty bounded polyhedron with $\mathbf{A} \in \mathbb{R}^{m \times d}$, then $-\mathcal{N}_{\mathcal{Z}}(z^(\bar{y})) = \{\mathbf{A}^\top \lambda : \lambda \in \Lambda\}$, and*

$$\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* = \{\mathbf{A}^\top \lambda - \gamma z^*(\bar{y}) : \lambda \in \Lambda\},$$

where $\Lambda := \{\lambda \geq 0 : \lambda_i = 0, \forall i \in [m] \setminus \mathcal{I}\}$ and $\mathcal{I} \subseteq [m]$ collects the indices of the constraints $\mathbf{a}_i^\top z \geq b_i$ active at $z^(\bar{y})$, with $\mathbf{a}_i^\top$ being the i -th row of $\mathbf{A}$.*

Proposition 4 shows that the set of target-risk minimizers is a closed convex cone—the negative normal cone at the nominal decision—translated by $-\gamma z^*(\bar{y})$. As a limiting case when $\gamma = 0$, the translation vanishes and the set is a cone, hence invariant under positive scaling of the prediction. We also note that $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$ is typically large and unbounded, since distinct predictions can induce the same robust decision.

It remains to locate the surrogate minimizers, which we do via the first-order optimality condition of the convex $\mathcal{RSP}\mathcal{O}_+$ conditional risk.

PROPOSITION 5. *Suppose that $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\|\mathbf{y}\|_2] < \infty$. Then*

$$\mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^* = \left\{ \hat{\mathbf{y}} \in \mathbb{R}^d : \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\mathbf{z}_\gamma^*(\mathbf{y})] = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[\mathbf{z}_\gamma^* \left(\hat{\mathbf{y}} - \frac{1}{a} \mathbf{y} \right) \right] \right\}.$$

REMARK 6 (NO CONTINUITY ASSUMPTION ON COST DISTRIBUTION). In contrast to the $\mathcal{SPO}_+$ setting, no continuity assumption on the distribution of $\mathbf{y}$ is needed in Proposition 5, since $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ is differentiable everywhere when $\gamma > 0$ as shown in Proposition 3. $\square$

Proposition 5 characterizes the surrogate minimizers through a balance condition: $\hat{\mathbf{y}}$ must match, in expectation, the two robust decisions appearing in the gradient of Proposition 3. To solve this condition in closed form we exploit central symmetry of the conditional distribution, which pairs $\mathbf{y}$ with $2\bar{\mathbf{y}} - \mathbf{y}$. The argument $\hat{\mathbf{y}} - \frac{1}{a}\mathbf{y}$ matches this pairing exactly when the coefficient of $\mathbf{y}$ is one, which is why the choice $a = 1$ anticipated in Remark 4 (Section 4.1) yields a closed-form minimizer.

COROLLARY 1. *Given $a = 1$, suppose that $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\|\mathbf{y}\|_2] < \infty$ and that the distribution of $\mathbf{y}$ admits a density that is positive on an open set containing $2\bar{\mathbf{y}} + \gamma\mathcal{Z}$ and is centrally symmetric about its mean $\bar{\mathbf{y}} := \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\mathbf{y}]$. Then $2\bar{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^*$. Furthermore, if the interior of $\mathcal{Z}$ is nonempty, then $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^* = \{2\bar{\mathbf{y}}\}$.*

Combining the two characterizations shows that Fisher consistency reduces to a single membership condition: does the surrogate optimizer $2\bar{\mathbf{y}}$ belong to the translated cone (11)? By the scale-invariance discussion following Proposition 4, when $\gamma = 0$, if the conditional mean $\bar{\mathbf{y}}$ belongs to the cone, then so does $2\bar{\mathbf{y}}$. Based on this fact, for $\gamma > 0$, the cone is translated further along $-\mathbf{z}^*(\bar{\mathbf{y}})$, while the surrogate optimizer remains $2\bar{\mathbf{y}}$, so the membership condition holds up to a threshold. We quantify this threshold in the next result. For a given $\mathbf{x}$, define

$$\bar{\gamma}(\mathbf{x}) := \sup \left\{ \gamma \geq 0 : 2\bar{\mathbf{y}} + \gamma \mathbf{z}^*(\bar{\mathbf{y}}) \in -\mathcal{N}_{\mathcal{Z}}(\mathbf{z}^*(\bar{\mathbf{y}})) \right\} \in [0, +\infty]. \quad (12)$$

When $\mathcal{Z}$ is a bounded polyhedron, (12) is the linear program

$$\sup \left\{ \gamma \geq 0 : 2\bar{\mathbf{y}} = \mathbf{A}^\top \boldsymbol{\lambda} - \gamma \mathbf{z}^*(\bar{\mathbf{y}}), \boldsymbol{\lambda} \in \boldsymbol{\Lambda} \right\}$$

over $(\gamma, \boldsymbol{\lambda})$, and is thus computable from the problem data.

THEOREM 2 (**Fisher Consistency**). *Set $a = 1$ and suppose the assumptions of Proposition 4 and of Corollary 1 hold for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$, and $\text{int } \mathcal{Z} \neq \emptyset$. Let $\bar{\gamma}(\mathbf{x})$ be defined by (12). Then:*

- (i) *The set in (12) is a closed interval containing 0, so $\bar{\gamma}(\mathbf{x})$ is well-defined and is attained whenever it is finite. If $0 < \gamma \leq \bar{\gamma}(\mathbf{x})$ for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$, then for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$ the unique minimizer $2\bar{\mathbf{y}}(\mathbf{x})$ of the pointwise $\mathcal{RSP}\mathcal{O}_+$ risk is also a minimizer of the pointwise $\mathcal{RSP}\mathcal{O}$ risk. That is, the $\mathcal{RSP}\mathcal{O}_+$ loss is Fisher consistent with the $\mathcal{RSP}\mathcal{O}$ loss.*

(ii) Suppose in addition that $\mathcal{Z} = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}$ is a nonempty bounded polyhedron. The threshold is strictly positive:

$$\bar{\gamma}(\mathbf{x}) \geq \frac{2 \operatorname{dist}(\bar{\mathbf{y}}(\mathbf{x}), \partial\mathcal{C}(\mathbf{x}))}{\|\mathbf{z}^*(\bar{\mathbf{y}}(\mathbf{x}))\|_2} > 0, \quad \mathcal{C}(\mathbf{x}) := -\mathcal{N}_{\mathcal{Z}}(\mathbf{z}^*(\bar{\mathbf{y}}(\mathbf{x}))),$$

where $\partial\mathcal{C}(\mathbf{x})$ denotes the boundary of $\mathcal{C}(\mathbf{x})$ and, when $\mathbf{z}^*(\bar{\mathbf{y}}(\mathbf{x})) = \mathbf{0}$, one has $\bar{\gamma}(\mathbf{x}) = +\infty$ and the bound holds trivially. In particular, the regime in (i) is nonempty for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$.

Theorem 2 essentially provides sufficient conditions for Fisher consistency. This is not only important to the analysis of the population limit scenario, but also a fundamental requirement for finite-sample performance guarantees, as we elaborate on in Section 5.2. The mechanism behind part (ii) is the same scale invariance identified after Proposition 4: uniqueness of the nominal solution forces $\bar{\mathbf{y}}(\mathbf{x})$ into the *interior* of the cone $\mathcal{C}(\mathbf{x})$; the interior of a cone is invariant under positive scaling, so $2\bar{\mathbf{y}}(\mathbf{x})$ lies in the interior as well—at distance $2 \operatorname{dist}(\bar{\mathbf{y}}(\mathbf{x}), \partial\mathcal{C}(\mathbf{x}))$ from the boundary—and the translation $-\gamma\mathbf{z}^*(\bar{\mathbf{y}}(\mathbf{x}))$ must travel at least that far before the membership can fail.

The following example illustrates both sets of $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$ and $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^*$ in the theorem and shows that the restriction on γ is essential: if $\gamma > \bar{\gamma}(\mathbf{x})$ for some $\mathbf{x}$, the unique surrogate minimizer $2\bar{\mathbf{y}}(\mathbf{x})$ lies outside $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$ and induces a decision that is suboptimal under the $\mathcal{RSP}\mathcal{O}$ criterion—a discrepancy that persists in the population limit.

EXAMPLE 1. Set $a = 1$. Fix $\mathbf{x}$ and let $\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\bar{\mathbf{y}}, \sigma^2 \mathbf{I}_2)$ with $\bar{\mathbf{y}} = (1, \frac{3}{2})^\top$ and any $\sigma > 0$, so that the density is positive on all of $\mathbb{R}^2$ and centrally symmetric about $\bar{\mathbf{y}}$, and the assumptions of Corollary 1 hold. Let the feasible set be

$$\mathcal{Z} := \left\{ \mathbf{z} = (z_1, z_2) : z_1 - z_2 \leq 1, \ z_1 \geq -1, \ z_1, z_2 \leq 0 \right\}.$$

The nominal solution is $\mathbf{z}^*(\bar{\mathbf{y}}) = (-1, -2)^\top$, at which the active constraints have normal vectors $(-1, 1)^\top$ and $(1, 0)^\top$. Therefore, the threshold of (12) evaluates to $\bar{\gamma} = 3/2$. For a comparison with the nominal situation, we also compute the sets $\mathcal{Y}_{\mathcal{SP}\mathcal{O}}^*$ and $\mathcal{Y}_{\mathcal{SP}\mathcal{O}_+}^*$, where $\mathcal{Y}_{\mathcal{SP}\mathcal{O}}^* = \arg \min_{\hat{\mathbf{y}} \in \mathbb{R}^d} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{SP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y})]$ and $\mathcal{Y}_{\mathcal{SP}\mathcal{O}_+}^* = \arg \min_{\hat{\mathbf{y}} \in \mathbb{R}^d} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{SP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y})]$. Figure 3 displays the two sets, together with $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$ and $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^* = \{2\bar{\mathbf{y}}\}$. For $\gamma = 1 \leq \bar{\gamma}$ the surrogate minimizer $2\bar{\mathbf{y}}$ lies in $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$ and Fisher consistency holds; for $\gamma = 2 > \bar{\gamma}$ it lies outside and the property fails, even though the distribution of $\mathbf{y}$ is continuous and symmetric. In both panels $\mathcal{Y}_{\mathcal{SP}\mathcal{O}_+}^* \subseteq \mathcal{Y}_{\mathcal{SP}\mathcal{O}}^*$, so the nominal surrogate remains Fisher consistent regardless of γ . $\square$

We close the subsection with a byproduct of the analysis by setting $\gamma = 0$. For a bounded polyhedron, uniqueness of the nominal solution forces $\bar{\mathbf{y}}$ into the interior of the cone $\mathcal{C}(\mathbf{x})$, hence $2\bar{\mathbf{y}}$ as well by scale-invariance. Therefore, Fisher consistency holds for the nominal surrogate with $a = 1$, as presented in Corollary 2. Notice that this result *complements* the Fisher consistency established

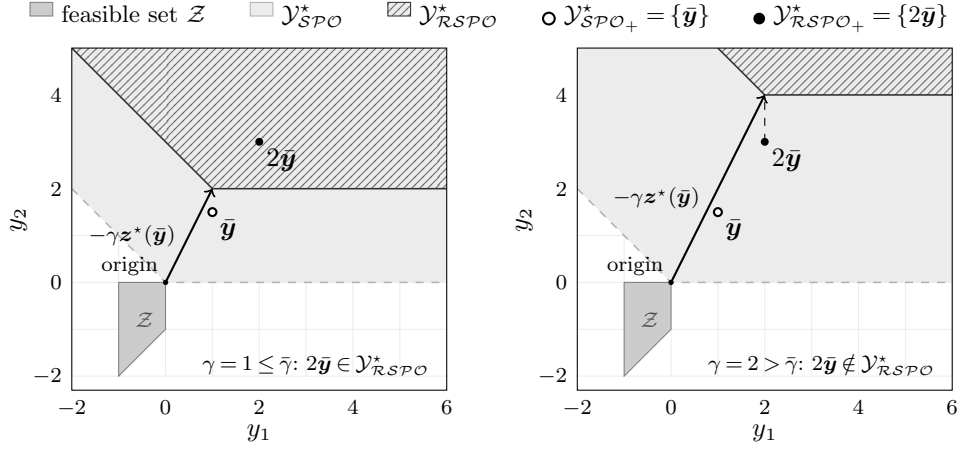


Figure 3 Illustration of Example 1. $\mathcal{Y}_{RSP}^*$ (hatched) is the negative normal cone $-\mathcal{N}_{\mathcal{Z}}(\mathbf{z}^*(\bar{\mathbf{y}}))$ translated by $-\gamma \mathbf{z}^*(\bar{\mathbf{y}})$ (solid arrow), so it moves with γ while the surrogate minimizer stays at $2\bar{\mathbf{y}}$. Dashed versus solid boundaries mark the open set $\mathcal{Y}_{SPO}^*$ against the closed set $\mathcal{Y}_{RSP}^*$. **Left:** $\gamma = 1 \leq \bar{\gamma}$, Fisher consistency holds; **right:** $\gamma = 2 > \bar{\gamma}$, Fisher consistency fails.

for the SPO_+ loss by [Elmachtoub and Grigas \(2022\)](#), which concerns $a = 2$ over general compact convex feasible sets, whereas Corollary 2 shows that, under analogous distributional conditions, the $a = 1$ variant holds over bounded polyhedra.

COROLLARY 2 (Fisher Consistency of SPO_+ with $a = 1$ and Polyhedrality). *Suppose that $\mathcal{Z}$ is a bounded polyhedron, and that for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$, the conditional distribution of $\mathbf{y}$ is centrally symmetric about $\bar{\mathbf{y}}$ with $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\|\mathbf{y}\|_2] < \infty$ and has a density positive on an open set containing $2\bar{\mathbf{y}}$. Further assume that the interior of $\mathcal{Z}$ is nonempty and the nominal problem $\min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z}$ admits a unique optimal solution. Then the SPO_+ loss with $a = 1$ is Fisher consistent with the SPO loss.*

4.3. Gradient Descent Method

Sections 4.1 and 4.2 have developed the convex surrogate scheme. We now turn to gradient descent methods, for which we begin by deriving the Jacobian of the decision map, based on which we sketch the algorithm.

The Jacobian of the decision map. Recall Lemma 1, which shows the decision map $\mathbf{z}_\gamma^*(\mathbf{y}_0) = \Pi_{\mathcal{Z}}(-\mathbf{y}_0/\gamma)$ is Lipschitz continuous with constant $1/\gamma$ and differentiable almost everywhere. To obtain an implementable algorithm, we need the derivative of the decision map. For a polyhedral feasible set it is available in closed form at points satisfying two standard regularity conditions: the linear independence constraint qualification (LICQ) and strict complementarity.

PROPOSITION 6. *Let $\gamma > 0$ and let $\mathcal{Z} = \{\mathbf{z} : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}$ be a nonempty bounded polyhedron. For a given $\mathbf{y}_0$, let $S_{\mathbf{y}_0} = \{i : \mathbf{a}_i^\top \mathbf{z}_\gamma^*(\mathbf{y}_0) = b_i\}$ denote the indices of active constraints at $\mathbf{z}_\gamma^*(\mathbf{y}_0)$ and $\mathbf{A}_{S_{\mathbf{y}_0}}$*

denote the submatrix of $\mathbf{A}$ containing only the rows $\mathbf{a}_i$ with indices $i \in S_{\mathbf{y}_0}$. Assume that (i) $\mathbf{A}_{S_{\mathbf{y}_0}}$ has full row rank (LICQ), and (ii) strict complementarity holds at $\mathbf{z}_\gamma^*(\mathbf{y}_0)$, i.e., the unique Karush–Kuhn–Tucker (KKT) multiplier $\boldsymbol{\lambda}^0$ associated with $\mathbf{z}_\gamma^*(\mathbf{y}_0)$ satisfies $\lambda_i^0 > 0$ for all $i \in S_{\mathbf{y}_0}$. Then there exists a neighborhood of $\mathbf{y}_0$ on which the active set of $\mathbf{z}_\gamma^*(\cdot)$ is identically $S_{\mathbf{y}_0}$ and $\mathbf{z}_\gamma^*(\cdot)$ is affine. In particular, $\mathbf{z}_\gamma^*(\cdot)$ is differentiable at $\mathbf{y}_0$, and its Jacobian is

$$\nabla \mathbf{z}_\gamma^*(\mathbf{y}_0) = \frac{1}{\gamma} (\mathbf{A}_{S_{\mathbf{y}_0}}^\top (\mathbf{A}_{S_{\mathbf{y}_0}} \mathbf{A}_{S_{\mathbf{y}_0}}^\top)^{-1} \mathbf{A}_{S_{\mathbf{y}_0}} - \mathbf{I}). \quad (13)$$

It is noted that the Jacobian (13) equals $-\frac{1}{\gamma}(\mathbf{I} - \mathbf{P}_{S_{\mathbf{y}_0}})$, where $\mathbf{P}_{S_{\mathbf{y}_0}} := \mathbf{A}_{S_{\mathbf{y}_0}}^\top (\mathbf{A}_{S_{\mathbf{y}_0}} \mathbf{A}_{S_{\mathbf{y}_0}}^\top)^{-1} \mathbf{A}_{S_{\mathbf{y}_0}}$ is the orthogonal projection onto the row space of $\mathbf{A}_{S_{\mathbf{y}_0}}$. It is thus a scaled projection onto the tangent space of the active face, with operator norm at most $1/\gamma$, consistent with Lemma 1. Three implementation points follow. First, since the expression depends on $\mathbf{A}_{S_{\mathbf{y}_0}}$ only through its row space, the inverse may be replaced by the Moore–Penrose pseudoinverse when $\mathbf{A}_{S_{\mathbf{y}_0}}$ is rank deficient; at such points, and at points where strict complementarity fails, (13) should be read as the derivative of the affine piece selected by the current active set rather than as a Jacobian of $\mathbf{z}_\gamma^*$. Second, in practice Jacobian–vector products are computed by implicitly differentiating the KKT system of the robust decision problem **ILRO-DECISION**, which avoids forming (13) explicitly. Third, this also accommodates feasible sets described through auxiliary variables: if $\mathcal{Z}$ is represented by a polyhedral lift in $(\mathbf{z}, \mathbf{w})$, the formula applies in the lifted space and the required derivative is read off the $\mathbf{z}$ -block. The portfolio experiment of Section 6.2 uses exactly this device for the ℓ_1 risk constraint; see Appendices E.1 and E.3 for more details.

REMARK 7 (RELATION TO DIFFERENTIABLE OPTIMIZATION LAYERS). Differentiating through a strongly convex quadratic program via its optimality conditions is standard (Amos and Kolter 2017, Agrawal et al. 2019), and quadratic smoothing of a linear program for this purpose is introduced by Wilder et al. (2019). We highlight that the difference lies in the modeling aspect: in the existing literature the quadratic term is a training-time smoothing device that is discarded at deployment, where the nominal decision map is restored, so the derivative is taken through a decision rule that is never implemented. In the ILRO framework the same γ governs training and deployment, so (13) is the derivative of the decision rule actually used, and descending the resulting objective decreases the regret that is actually incurred. $\square$

Algorithm. Based on the explicit Jacobian of the robust decision map, we minimize the regularized empirical $\mathcal{RSP\mathcal{O}}$ objective

$$\hat{L}_N(\theta) := \frac{1}{N} \sum_{i \in [N]} [\mathbf{y}_i^\top \mathbf{z}_\gamma^*(\mathbf{g}_\theta(\mathbf{x}_i)) - v^*(\mathbf{y}_i)] + \lambda \Omega(\mathbf{g}_\theta) \quad (14)$$

over a parametric class $\{\mathbf{g}_\theta : \theta \in \Theta\}$ with $\mathbf{g}_\theta$ differentiable in θ , where $\Omega(\cdot)$ is the prediction-model regularizer and $\lambda \geq 0$ its weight. Since $v^*(\mathbf{y}_i)$ does not depend on θ , the chain rule gives the per-sample gradient $[\nabla_\theta \mathbf{g}_\theta(\mathbf{x}_i)]^\top \mathbf{q}_i$ with $\mathbf{q}_i = \nabla \mathbf{z}_\gamma^*(\mathbf{g}_\theta(\mathbf{x}_i))^\top \mathbf{y}_i$. We initialize $\theta^{(0)}$ at the $\mathcal{RSP}\mathcal{O}_+$ minimizer, and apply standard stochastic gradient descent coupled with an Armijo test (Nocedal and Wright 2006) to decrease the objective sufficiently. The details of the algorithm are deferred to Appendix C.

In summary, this section provides two complementary routes to the empirical $\mathcal{RSP}\mathcal{O}$ learning problem (6): the convex $\mathcal{RSP}\mathcal{O}_+$ surrogate, which is globally solvable for linear predictors over polyhedral feasible sets and targets the correct population predictor when the robustness parameter γ is below the specific threshold; and the gradient method, which descends the $\mathcal{RSP}\mathcal{O}$ objective directly for any $\gamma > 0$, and can be used to refine the surrogate solution locally. The next section analyzes the statistical guarantees of the two predictors defined by the target and surrogate losses.

5. Theoretical Guarantees

This section is devoted to the finite-sample performance guarantees of the two predictors, namely, the $\mathcal{RSP}\mathcal{O}$ predictor and the $\mathcal{RSP}\mathcal{O}_+$ predictor. In both cases, we are interested in the population $\mathcal{RSP}\mathcal{O}$ risk, since it is the target loss that measures the realized cost of the deployed robust decision.

Let $\mathbb{P}$ denote the population distribution of $(\mathbf{x}, \mathbf{y})$ and fix $\gamma > 0$. We define the following quantities for a predictor $\mathbf{g}$. For a concise exposition, we suppress two types of dependence throughout this section. First, all quantities implicitly depend on γ through the losses $\ell_{\mathcal{RSP}\mathcal{O}}$ and $\ell_{\mathcal{RSP}\mathcal{O}_+}$. Second, sample-based quantities, including empirical risks, estimators, and Rademacher complexities, depend on the sample size N . We make these dependencies explicit only when they are relevant to the interpretation. The population risks under the target and surrogate losses are

$$R(\mathbf{g}) := \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}}[\ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{g}(\mathbf{x}), \mathbf{y})], \quad R_+(\mathbf{g}) := \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}}[\ell_{\mathcal{RSP}\mathcal{O}_+}(\mathbf{g}(\mathbf{x}), \mathbf{y})],$$

respectively. Given i.i.d. samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [N]}$, their empirical counterparts are

$$\hat{R}(\mathbf{g}) := \frac{1}{N} \sum_{i \in [N]} \ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{g}(\mathbf{x}_i), \mathbf{y}_i), \quad \hat{R}_+(\mathbf{g}) := \frac{1}{N} \sum_{i \in [N]} \ell_{\mathcal{RSP}\mathcal{O}_+}(\mathbf{g}(\mathbf{x}_i), \mathbf{y}_i),$$

respectively. For a class $\mathcal{G}$ of measurable predictors from $\mathcal{X}$ to $\mathbb{R}^d$, we set

$$R^*(\mathcal{G}) := \inf_{\mathbf{g} \in \mathcal{G}} R(\mathbf{g}), \quad R_+^*(\mathcal{G}) := \inf_{\mathbf{g} \in \mathcal{G}} R_+(\mathbf{g}).$$

Recall $\mathcal{G}_{\text{all}}$ denotes the class of all measurable functions from $\mathcal{X}$ to $\mathbb{R}^d$, and define the lowest achievable target and surrogate risks by $R^* := R^*(\mathcal{G}_{\text{all}})$ and $R_+^* := R_+^*(\mathcal{G}_{\text{all}})$, respectively. Throughout the section we analyze the empirical $\mathcal{RSP}\mathcal{O}$ and $\mathcal{RSP}\mathcal{O}_+$ risk minimizers without regularization, corresponding to $\lambda = 0$ in the formulations (6) and (8):

$$\hat{\mathbf{g}}_{\mathcal{RSP}\mathcal{O}} \in \arg \min_{\mathbf{g} \in \mathcal{G}} \hat{R}(\mathbf{g}), \quad \hat{\mathbf{g}}_{\mathcal{RSP}\mathcal{O}_+} \in \arg \min_{\mathbf{g} \in \mathcal{G}} \hat{R}_+(\mathbf{g}),$$

which we assume to exist.

The two predictors are measured against different benchmarks, and the distinction drives the analysis that follows. The $\mathcal{RSP}\mathcal{O}$ predictor minimizes an empirical version of the very risk we wish to control, so the natural benchmark is the best predictor available in the hypothesis class, and the object of interest is the within-class excess risk $R(\hat{\mathbf{g}}_{\mathcal{RSP}\mathcal{O}}) - R^*(\mathcal{G})$, which is studied in Section 5.1. The $\mathcal{RSP}\mathcal{O}_+$ predictor, by contrast, minimizes a different risk, so to obtain a target-risk guarantee, we naturally exploit Fisher consistency. The latter is a statement about pointwise minimizers over all measurable predictors as opposed to a restricted function class $\mathcal{G}$. Its benchmark is therefore R^* , and the object of interest is the target excess risk $R(\hat{\mathbf{g}}_{\mathcal{RSP}\mathcal{O}_+}) - R^*$, which is studied in Section 5.2.

5.1. Excess Risk Bounds of $\mathcal{RSP}\mathcal{O}$ Predictor

The excess risk $R(\hat{\mathbf{g}}_{\mathcal{RSP}\mathcal{O}}) - R^*(\mathcal{G})$ is controlled by a uniform deviation bound over the hypothesis class, and the analysis therefore reduces to measuring its complexity. We carry this out in two steps. We first bound the deviation by the Rademacher complexity of the function class induced by the loss, and then transfer that bound to the prediction class itself, using the Lipschitz continuity of the decision map.

Given a vector-valued hypothesis class $\mathcal{G}$, its d -dimensional multivariate Rademacher complexity, and its empirical counterpart conditional on a sample $\mathbf{x}^N = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, are

$$\mathfrak{R}_d(\mathcal{G}) := \mathbb{E}_{\sigma, \mathbf{X}} \left[\sup_{\mathbf{g} \in \mathcal{G}} \frac{1}{N} \sum_{i \in [N]} \sigma_i^\top \mathbf{g}(\mathbf{x}_i) \right], \quad \hat{\mathfrak{R}}_d(\mathcal{G}; \mathbf{x}^N) := \mathbb{E}_{\sigma} \left[\sup_{\mathbf{g} \in \mathcal{G}} \frac{1}{N} \sum_{i \in [N]} \sigma_i^\top \mathbf{g}(\mathbf{x}_i) \right],$$

where $\sigma_i \in \{+1, -1\}^d$ are independent Rademacher random vectors (Maurer 2016, Bartlett and Mendelson 2002) with $\sigma = (\sigma_i)_{i \in [N]}$, and $\mathbf{X}$ is the underlying population for the sample $\mathbf{x}^N$, so that $\mathfrak{R}_d(\mathcal{G}) = \mathbb{E}_{\mathbf{X}}[\hat{\mathfrak{R}}_d(\mathcal{G}; \mathbf{x}^N)]$. We next define the loss-induced function class.

DEFINITION 4 (LOSS-INDUCED FUNCTION CLASS). Let $\mathcal{G} \subseteq \{\mathbf{g} : \mathcal{X} \rightarrow \mathbb{R}^d\}$ and let $\ell : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a loss function. The loss-induced function class associated with ℓ and $\mathcal{G}$ is

$$\mathcal{H}_\ell = \{h_{\mathbf{g}} : (\mathbf{x}, \mathbf{y}) \mapsto \ell(\mathbf{g}(\mathbf{x}), \mathbf{y}) \mid \mathbf{g} \in \mathcal{G}\},$$

and its Rademacher complexity is

$$\mathfrak{R}_1(\mathcal{H}_\ell) := \mathbb{E}_{(\sigma_i)_{i \in [N]}, (\mathbf{X}, \mathbf{Y})} \left[\sup_{\mathbf{g} \in \mathcal{G}} \frac{1}{N} \sum_{i \in [N]} \sigma_i \ell(\mathbf{g}(\mathbf{x}_i), \mathbf{y}_i) \right],$$

where σ_i are independent Rademacher random variables and $\mathbf{Y}$ is the underlying population for the sample $(\mathbf{y}_1, \dots, \mathbf{y}_N)$.

To ease the analysis we impose the bounded objective-gap condition

$$b = \sup_{\mathbf{y} \in \mathcal{Y}} \left\{ \max_{\mathbf{z} \in \mathcal{Z}} \mathbf{y}^\top \mathbf{z} - \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{y}^\top \mathbf{z} \right\} < \infty,$$

which measures the largest difference in realized objective value between two feasible decisions under the same cost vector. Since $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \mathcal{Z}$, it implies the uniform bound $0 \leq \ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) \leq b$ for every $\hat{\mathbf{y}} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathcal{Y}$. The first step is then a standard consequence of the symmetrization argument of [Bartlett and Mendelson \(2002\)](#).

PROPOSITION 7. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sup_{\mathbf{g} \in \mathcal{G}} \left| R(\mathbf{g}) - \hat{R}(\mathbf{g}) \right| \leq 2\mathfrak{R}_1(\mathcal{H}_{\ell_{\mathcal{RSP}\mathcal{O}}}) + b \sqrt{\frac{\log(2/\delta)}{2N}}.$$

The second step relates the complexity of $\mathcal{H}_{\ell_{\mathcal{RSP}\mathcal{O}}}$ to that of $\mathcal{G}$. This is possible because the robust decision map is a projection, so the loss is Lipschitz in the prediction. In particular, for every fixed $\mathbf{y}$, $\ell_{\mathcal{RSP}\mathcal{O}}(\cdot, \mathbf{y})$ is $(\|\mathbf{y}\|_2/\gamma)$ -Lipschitz, because $\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) - v^*(\mathbf{y})$ and $\mathbf{z}_\gamma^*(\cdot)$ is $1/\gamma$ -Lipschitz by Lemma 1. This yields the following result by applying the vector contraction inequality of [Maurer \(2016\)](#).

LEMMA 2. *Suppose that $\|\mathbf{y}\|_2 \leq r$ for all $\mathbf{y} \in \mathcal{Y}$. Then $\mathfrak{R}_1(\mathcal{H}_{\ell_{\mathcal{RSP}\mathcal{O}}}) \leq \frac{\sqrt{2}r}{\gamma} \mathfrak{R}_d(\mathcal{G})$.*

Combining the two steps gives the central result of this subsection.

THEOREM 3 (**Meta Generalization Bound**). *For $\gamma > 0$, suppose that $\|\mathbf{y}\|_2 \leq r$ for all $\mathbf{y} \in \mathcal{Y}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$R(\hat{\mathbf{g}}_{\mathcal{RSP}\mathcal{O}}) - R^*(\mathcal{G}) \leq \frac{4\sqrt{2}r}{\gamma} \mathfrak{R}_d(\mathcal{G}) + b \sqrt{\frac{2\log(2/\delta)}{N}}.$$

Theorem 3 separates the two sources of statistical difficulty. The robust decision map enters only through the Lipschitz factor r/γ , and the richness of the prediction class only through $\mathfrak{R}_d(\mathcal{G})$. Deriving a guarantee for a specific hypothesis class therefore reduces to the single task of bounding $\mathfrak{R}_d(\mathcal{G})$. Appendix D carries this out for four representative classes—bounded affine predictors, polynomial-discrimination classes, classes with bounded Dudley entropy integral, and vector-valued RKHS balls. Table 2 collects the resulting bounds: across the four classes the excess risk decays at the parametric rate $N^{-1/2}$, up to a logarithmic factor in the polynomial-discrimination case, with the robustness parameter entering uniformly through $1/\gamma$.

REMARK 8 (COMPARISON WITH $\mathcal{SP}\mathcal{O}$ RISK BOUNDS). The argument leading to Theorem 3 is viable because the robust decision map is smooth. Vector contraction requires the loss to be Lipschitz in the prediction, which is satisfied by $\ell_{\mathcal{RSP}\mathcal{O}}(\cdot, \mathbf{y})$. This is not the case for $\ell_{\mathcal{SP}\mathcal{O}}(\cdot, \mathbf{y})$, as it is discontinuous. Generalization bounds for $\mathcal{SP}\mathcal{O}$ are therefore obtained by controlling the induced

Table 2 Instantiations of the meta generalization bound. The third column reports the complexity bound available for each class and the last column the leading term of the resulting bound on the excess risk $R(\hat{\mathbf{g}}_{\text{RSPO}}) - R^*(\mathcal{G})$.

Hypothesis class	Condition imposed	Complexity bound $\mathfrak{N}_d(\mathcal{G}) \leq$	Excess risk bound (leading term)
Bounded affine (Example EC.1)	$\mathbf{g}_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}\boldsymbol{\varphi}(\mathbf{x})$ with $\ \boldsymbol{\varphi}(\mathbf{x})\ _2 \leq \kappa$ and $\ \mathbf{W}\ _F \leq B$	$B\kappa\sqrt{\frac{d}{N}}$	$\frac{4\sqrt{2}rB\kappa}{\gamma}\sqrt{\frac{d}{N}}$
Polynomial discrimination (Example EC.2)	Order ν in the sense of Definition EC.1, with $D_{\mathcal{G}}(\mathbf{x}^N) \leq B_1$	$B_1\sqrt{\frac{2\nu\log(N+1)}{N}}$	$\frac{8rB_2}{\gamma}\sqrt{\frac{\nu\log(N+1)}{N}}$
Dudley entropy integral (Example EC.3)	Entropy integral bounded by $J_{\mathcal{G}}$ uniformly over samples	$C\frac{J_{\mathcal{G}}}{\sqrt{N}}$	$\frac{4\sqrt{2}Cr}{\gamma}\frac{J_{\mathcal{G}}}{\sqrt{N}}$
RKHS ball (Example EC.4)	$\sum_j \ g_j\ _{\mathcal{H}}^2 \leq R_{\mathcal{H}}^2$ with $\sup_{\mathbf{x}} K(\mathbf{x}, \mathbf{x}) \leq \kappa^2$	$\kappa R_{\mathcal{H}}\sqrt{\frac{d}{N}}$	$\frac{4\sqrt{2}r\kappa R_{\mathcal{H}}}{\gamma}\sqrt{\frac{d}{N}}$

Note. All bounds are stated under the assumptions $\gamma > 0$ and $\|\mathbf{y}\|_2 \leq r$ for every $\mathbf{y} \in \mathcal{Y}$. Each excess risk bound holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$ and reports only the leading term; the additive term $b\sqrt{2\log(2/\delta)/N}$, common to all four classes, is omitted. Details and proofs are given in Appendix D.

loss class directly, through an elegant analysis of margin conditions and combinatorial complexity measures tailored to the feasible region (El Balghiti et al. 2023). However, this comes at the cost that the resulting bounds depend on the complexity and dimensionality of the feasible set $\mathcal{Z}$. $\square$

The fact that our bounds depend on the order of $1/\gamma$ highlights the relationship between the amount of data and the robustness requirement: when the formulation is more robust (larger γ), fewer samples are needed to achieve the same excess risk bound. This suggests that, when data are limited, the decision-maker should seek more robust decisions.

5.2. Excess Risk Bounds of $\mathcal{RSP}\mathcal{O}_+$ Predictor

Now we turn to the excess risk $R(\hat{\mathbf{g}}_{\text{RSPO}_+}) - R^*$ of the surrogate predictor. A natural first attempt is to compare $\hat{\mathbf{g}}_{\text{RSPO}_+}$ with $\hat{\mathbf{g}}_{\text{RSPO}}$ and invoke the bounds of Section 5.1. However, $\hat{\mathbf{g}}_{\text{RSPO}}$ is the minimizer of a nonconvex optimization problem (typically NP-hard), whereas $\hat{\mathbf{g}}_{\text{RSPO}_+}$ is obtained by minimizing a convex surrogate problem (typically tractable), under finite samples and a restricted function class. Therefore, the computational complexity barrier prevents the two predictors from being close in general. In contrast, Fisher consistency results dictate that, in the population limit and when the function class consists of all measurable functions, the surrogate minimizer is also a target-risk minimizer. Thus, we derive the excess risk bounds through Fisher consistency.

To ease our discussion, we introduce the population “oracle predictor”

$$\mathbf{g}_0(\mathbf{x}) := 2\bar{\mathbf{y}}(\mathbf{x}) = 2\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\mathbf{y}], \quad (15)$$

which by Corollary 1 is the unique pointwise minimizer of the surrogate risk when $a = 1$ under the assumptions of this corollary.

The analysis can be roughly divided into three steps. First, using a standard generalization argument, we bound the excess risk of the surrogate $R_+(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R_+^*$. Next, by the local strong convexity, this implies $\widehat{\mathbf{g}}_{\text{RSPO}_+}$ is close to $\mathbf{g}_0$. Finally, applying the Lipschitz property of $R(\cdot)$ and the fact that $R(\mathbf{g}_0) = R^*$ due to Fisher consistency, we obtain the bound on target excess risk $R(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R^*$. We begin by providing necessary definitions and conditions.

Pointwise risks and the conditions. Throughout, $\mathcal{B}(\mathbf{c}, r) := \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u} - \mathbf{c}\|_2 \leq r\}$ denotes the closed Euclidean ball. For a fixed context $\mathbf{x}$, define the pointwise target and surrogate risks as

$$R(\mathbf{c}; \mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\ell_{\text{RSPO}}(\mathbf{c}, \mathbf{y})], \quad R_+(\mathbf{c}; \mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\ell_{\text{RSPO}_+}(\mathbf{c}, \mathbf{y})], \quad (16)$$

respectively. Since $\mathcal{Z}$ is compact, $R_{\mathcal{Z}} := \max_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z}\|_2 < \infty$ and

$$|\ell_{\text{RSPO}}(\mathbf{c}, \mathbf{y})| \leq 2R_{\mathcal{Z}}\|\mathbf{y}\|_2 \quad \text{for every } \mathbf{c} \in \mathbb{R}^d, \quad (17)$$

so whenever $\mathbb{E}\|\mathbf{y}\|_2 < \infty$ every target risk appearing below is finite and the tower property $R(\mathbf{g}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}}[R(\mathbf{g}(\mathbf{x}); \mathbf{x})]$ applies to each. Our analysis leverages the oracle predictor $\mathbf{g}_0$, and hence requires the following condition about $\mathbf{g}_0$.

Condition 1 (Oracle Predictor Condition)

- (i) $\mathbf{g}_0$ is measurable and, for almost every $\mathbf{x}$, $\mathbf{g}_0(\mathbf{x})$ minimizes $R_+(\cdot; \mathbf{x})$ over $\mathbb{R}^d$;
- (ii) for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$, $R_+(\cdot; \mathbf{x})$ is convex on $\mathbb{R}^d$ and μ -strongly convex on the closed ball $\mathcal{B}(\mathbf{g}_0(\mathbf{x}), \rho)$, which is centered at $\mathbf{g}_0(\mathbf{x})$ with radius ρ . Here, $\mu, \rho > 0$ do not depend on $\mathbf{x}$;
- (iii) $\mathbf{g}_0(\mathbf{x}) \in \mathcal{Y}_{\text{RSPO}}^*(\mathbf{x})$ for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$;
- (iv) $\|\mathbf{g}_0(\mathbf{x})\|_2 \leq \bar{\beta}$ almost surely for some $0 < \bar{\beta} < \infty$.

Condition 1(i) and condition 1(iii) are implied by the results established in this paper and we provide sufficient conditions for (ii) in Lemma EC.3. With $a = 1$, Condition 1(i) holds under the assumptions of Corollary 1. Condition 1(ii) is a local curvature requirement. We provide an example where it holds by Lemma EC.3 in Appendix A. The example, roughly speaking, requires that the distribution of $\mathbf{y}$ must place enough mass on a region centered at $\mathbf{g}_0(\mathbf{x})$ obtained by rescaling the feasible set. Since Lemma EC.3 is stated for a fixed $\mathbf{x}$, Condition 1(ii) follows with $\mu = \text{ess inf}_{\mathbf{x}} p_0(\mathbf{x})/\gamma$ provided $\text{ess inf}_{\mathbf{x}} p_0(\mathbf{x}) > 0$. Condition 1(iii) holds by Theorem 2 whenever $\mathcal{Z}$ is a bounded polyhedron with nonempty interior and $\gamma \leq \bar{\gamma}(\mathbf{x})$ for almost every $\mathbf{x}$, which a fixed $\gamma > 0$ satisfies as soon as $\gamma \leq \text{ess inf}_{\mathbf{x}} \bar{\gamma}(\mathbf{x})$. When conditions (i)–(iii) hold, $\mathbf{g}_0(\mathbf{x})$ is the unique minimizer of $R_+(\cdot; \mathbf{x})$ due to strong convexity, and hence Fisher consistency holds by definition. Condition (iv) holds trivially when $\mathbf{y}$ is uniformly bounded.

Bounds on surrogate excess risk. In this step, we bound surrogate excess risk

$$R_+(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R_+^* = [R_+(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R_+^*(\mathcal{G})] + [R_+^*(\mathcal{G}) - R_+^*],$$

using generalization bound for the surrogate empirical risk minimizer. Define the uniform upper bound of the surrogate loss over the prediction image $\mathcal{G}(\mathcal{X}) := \{\mathbf{g}(\mathbf{x}) : \mathbf{g} \in \mathcal{G}, \mathbf{x} \in \mathcal{X}\}$ and the label set $\mathcal{Y}$,

$$\mathfrak{L}(\gamma) := \sup_{\hat{\mathbf{y}} \in \mathcal{G}(\mathcal{X}), \mathbf{y} \in \mathcal{Y}} \ell_{\text{RSPO}_+}(\hat{\mathbf{y}}, \mathbf{y}).$$

REMARK 9. If $\|\mathbf{y}\|_2 \leq r$ on $\mathcal{Y}$ and $\|\hat{\mathbf{y}}\|_2 \leq G_\infty$ on $\mathcal{G}(\mathcal{X})$, the Cauchy–Schwarz inequality yields $\mathfrak{L}(\gamma) \leq 2(r + aG_\infty)R_{\mathcal{Z}} + \frac{a\gamma}{2}R_{\mathcal{Z}}^2$, which implies that $\mathfrak{L}(\gamma)$ grows at most linearly in γ .

PROPOSITION 8. Let $\mathcal{G}$ be a hypothesis class from $\mathcal{X}$ to $\mathbb{R}^d$. Suppose that $\mathcal{Z}$ is bounded with diameter $D_{\mathcal{Z}}$, that $\mathcal{G}(\mathcal{X})$ and $\mathcal{Y}$ are bounded so that $\mathfrak{L}(\gamma) < \infty$, and that $\mathfrak{R}_d(\mathcal{G}) \leq C_0/\sqrt{N}$ for a constant C_0 . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R_+(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R_+^*(\mathcal{G}) \leq \frac{4\sqrt{2}aD_{\mathcal{Z}}C_0}{\sqrt{N}} + \mathfrak{L}(\gamma)\sqrt{\frac{2\log(2/\delta)}{N}}.$$

Proposition 8 controls the generalization error within function class $\mathcal{G}$. However, Fisher consistency is only meaningful over the set of all measurable predictors, and we have to account for the expressiveness of $\mathcal{G}$ itself. Accordingly, for $\delta \in (0, 1)$ we write, at $a = 1$,

$$B_N := \frac{4\sqrt{2}D_{\mathcal{Z}}C_0}{\sqrt{N}} + \mathfrak{L}(\gamma)\sqrt{\frac{2\log(2/\delta)}{N}}, \quad A_{\mathcal{G}}^+ := R_+^*(\mathcal{G}) - R_+^* \geq 0. \quad (18)$$

Under the conditions in Proposition 8, we have

$$R_+(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R_+^* \leq B_N + A_{\mathcal{G}}^+. \quad (19)$$

The first term B_N vanishes at the rate $N^{-1/2}$, and the second term $A_{\mathcal{G}}^+$ is a deterministic property of the pair $(\mathcal{G}, \mathbb{P})$. We briefly discuss when $A_{\mathcal{G}}^+$ vanishes. Under the conditions of Corollary 1, the pointwise minimizer of the surrogate risk is $\mathbf{g}_0(\mathbf{x}) = 2\bar{\mathbf{y}}(\mathbf{x})$, so R_+^* is attained by $\mathbf{g}_0$. Thus, $\mathbf{g}_0 \in \mathcal{G}$ up to $\mathbb{P}_{\mathbf{x}}$ -null sets implies $A_{\mathcal{G}}^+ = 0$. Conversely, if the infimum defining $R_+^*(\mathcal{G})$ is attained by some predictor in $\mathcal{G}$, $A_{\mathcal{G}}^+ = 0$ only if $\mathbf{g}_0 \in \mathcal{G}$ up to null sets due to uniqueness of the pointwise minimizer.

Bounds on target excess risk. Based on (19), we bound the distance between $\widehat{\mathbf{g}}_{\text{RSPO}_+}$ and $\mathbf{g}_0$ by the local strong convexity condition. Then, we derive the bound of excess risk $R(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R^*$, by applying Lipschitz property of $R(\cdot)$ and the fact that $R(\mathbf{g}_0) = R^*$ due to Fisher consistency, which leads to the following theorem. The detailed steps are elaborated in its proof.

THEOREM 4 (Target Excess Risk: General Case). Fix $a = 1$ and $\gamma > 0$, and suppose Condition 1 and the assumptions of Proposition 8 hold. Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R^* \leq \frac{\bar{\beta}}{2\gamma} \left[\sqrt{\frac{2(B_N + A_{\mathcal{G}}^+)}{\mu}} + \frac{2(B_N + A_{\mathcal{G}}^+)}{\mu\rho} \right], \quad (20)$$

where μ, ρ are the curvature constants of Condition 1(ii), $\bar{\beta}$ is the envelope constant of Condition 1(iv), and $B_N, A_{\mathcal{G}}^+$ are defined in (18). If, in addition, $\mathbf{g}_0 \in \mathcal{G}$ up to $\mathbb{P}_{\mathbf{x}}$ -null sets and the structural constants $D_{\mathcal{Z}}, C_0, \mathfrak{L}(\gamma), \mu, \rho, \bar{\beta}, \gamma$ do not depend on N , then

$$R(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R^* = \mathcal{O}_p(N^{-1/4}).^2 \quad (21)$$

This is consistent with the excess risk bounds of the $\mathcal{RSPO}$ predictor derived in Section 5.1: given the structural constants, as the formulation is more robust (larger γ), fewer samples are needed to achieve the same excess risk bound of the $\mathcal{RSPO}_+$ predictor, by noting that the term $\mathfrak{L}(\gamma)$ in B_N grows at most linearly in γ (Remark 9). If we further impose a quadratic target upper-growth condition, we can improve the bound on target excess risk from order $N^{-1/4}$ to $N^{-1/2}$, as the following corollary shows.

COROLLARY 3 (Target Excess Risk: Fast Rate). Suppose the assumptions of Theorem 4 hold, and assume further the following uniform quadratic target-growth condition: there exist constants $L, r_0 > 0$, independent of $\mathbf{x}$, such that, for almost every $\mathbf{x}$,

$$R(\mathbf{c}; \mathbf{x}) - \inf_{\mathbf{c}' \in \mathbb{R}^d} R(\mathbf{c}'; \mathbf{x}) \leq \frac{L}{2} \|\mathbf{c} - \mathbf{g}_0(\mathbf{x})\|_2^2, \quad \text{whenever } \|\mathbf{c} - \mathbf{g}_0(\mathbf{x})\|_2 \leq r_0. \quad (22)$$

Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R^* \leq \frac{B_N + A_{\mathcal{G}}^+}{\min\{\mu/L, \mu\gamma r_0/\bar{\beta}, \mu\gamma\rho/\bar{\beta}\}}. \quad (23)$$

If, in addition, $\mathbf{g}_0 \in \mathcal{G}$ up to $\mathbb{P}_{\mathbf{x}}$ -null sets and the structural constants do not depend on N , then $R(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R^* = \mathcal{O}_p(N^{-1/2})$.

Condition (22) asks that the target risk not grow faster than quadratically near the oracle predictor, complementing from above the lower bound of the growth of surrogate risk from Condition 1(ii). This implies the growth rate of the target risk is at most comparable to that of the surrogate risk.

² Here $\mathcal{O}_p$ denotes stochastic order: $X_N = \mathcal{O}_p(a_N)$ if, for every $\varepsilon > 0$, there exists $M > 0$ such that $\mathbb{P}(|X_N| > Ma_N) < \varepsilon$ for all sufficiently large N .

6. Numerical Experiments

In this section, we examine whether using the same robust decision map to define the training loss and generate deployed decisions improves out-of-sample decision quality. We conduct computational experiments on synthetic instances of two canonical problem classes: capacitated transportation and risk-constrained portfolio optimization.

Synthetic Data-Generation Process. Following a synthetic design similar to that of [Elmachtoub and Grigas \(2022\)](#), we use the same basic data-generating protocol for the two problem classes, where d denotes the dimension of the cost and decision vectors and p the dimension of the context vector. For each experimental configuration, we draw a ground-truth coefficient matrix $\mathbf{B}^* \in \mathbb{R}^{d \times p}$ once, with independent Bernoulli(0.5) entries, and hold it fixed across the 20 replications. For each observation, the context vector $\mathbf{x}_i \in \mathbb{R}^p$ is drawn from a standard multivariate Gaussian distribution; that is, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Given $\mathbf{x}_i$, the cost vector $\mathbf{y}_i \in \mathbb{R}^d$ is generated according to $\mathbf{y}_i = f_{\text{deg}}(\mathbf{B}^* \mathbf{x}_i, \boldsymbol{\epsilon}_i)$, where $\boldsymbol{\epsilon}_i$ denotes the noise term and f_{deg} is polynomial in its first argument. The problem-specific form of f_{deg} and the noise distribution are stated in Sections 6.1 and 6.2. The positive integer parameter, **deg**, determines the polynomial degree. Importantly, since all methods employ a linear hypothesis class for the predictor, **deg** controls the extent of *model misspecification*. Across the numerical experiments, we vary the polynomial degree **deg**, the noise level, the decision (and cost) dimension d , and the training-sample size N to evaluate different aspects of the methods' performance.

Methods Compared. As summarized in Table 1, the four empirical methods instantiate the four learning–decision pipelines introduced in Section 1. All four methods use the same linear prediction class, $\mathbf{g}_\theta(\mathbf{x}) = \mathbf{B}\mathbf{x}$, but differ in the loss used to train the predictor and the decision map used to generate the deployed decision. First, least squares minimizes the squared prediction loss $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$ and uses the nominal decision map. Second, $\mathcal{SPO}_+$ trains the predictor with the nominal $\mathcal{SPO}_+$ surrogate and uses the same nominal decision map for decision making. Third, $\mathcal{SPO}_+$ with robust decisions retains the $\mathcal{SPO}_+$ training loss but uses the robust decision map [ILRO-DECISION](#) to generate deployed decisions. Finally, $\mathcal{RSPO}_+$ trains the predictor by solving (8) and uses the same robust decision map to generate deployed decisions. Thus, among the four methods, $\mathcal{RSPO}_+$ is the only one that combines robust decision-making with learning–decision alignment.

Training, Validation, and Testing. For each replication, we independently generate a new training sample and test sample. We apply the same sample-splitting protocol to all methods. We train the candidate predictors on 70% of the training sample and select their hyperparameters using the remaining 30%. Every method selects the regularization parameter λ from five logarithmically spaced values in $[10^{-3}, 10]$. For $\mathcal{RSPO}_+$ and $\mathcal{SPO}_+$ with robust decisions, we additionally select the robustness parameter γ from five logarithmically spaced values in $[10^{-6}, 10]$. Least squares uses the average prediction loss on the validation data for hyperparameter selection, whereas $\mathcal{SPO}_+$,

$\mathcal{SPO}_+$ with robust decisions, and $\mathcal{RSP}\mathcal{O}_+$ use the average normalized decision loss. We refer to the predictor associated with the selected hyperparameters as the validated predictor. Each validated predictor is evaluated on the corresponding test sample of size $n_{\text{test}} = 100$ in each of the 20 replications.

Performance Metrics. All quantities below are defined within one experimental replication, with the replication index suppressed. For each method m , let $\hat{\mathbf{g}}_m$ denote its validated predictor and let γ_m denote the robustness parameter of the robust decision map used to generate its deployed decision. For notational uniformity, we set $\gamma_m = 0$ for least squares and $\mathcal{SPO}_+$ and write $\mathbf{z}_0^*$ for the decision returned by the nominal decision map (2). For $\mathcal{RSP}\mathcal{O}_+$ and $\mathcal{SPO}_+$ with robust decisions, γ_m is the value of γ selected separately for method m by the validation procedure described above. Under this convention, every method deploys a decision denoted by $\mathbf{z}_{\gamma_m}^*(\hat{\mathbf{g}}_m(\mathbf{x}))$; when $\gamma_m > 0$, this decision is returned by the robust decision map [ILRO-DECISION](#). On the test sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_{\text{test}}}$, we measure prediction accuracy by

$$\text{Relative Prediction Loss}_m = \frac{\sum_{i=1}^{n_{\text{test}}} \|\hat{\mathbf{g}}_m(\mathbf{x}_i) - \mathbf{y}_i\|_2^2}{\sum_{i=1}^{n_{\text{test}}} \|\mathbf{y}_i\|_2^2}.$$

Let $v^*(\mathbf{y}_i) = \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{y}_i^\top \mathbf{z}$ denote the nominal clairvoyant cost under the realized cost vector $\mathbf{y}_i$. We measure decision quality by

$$\text{Normalized Decision Loss}_m = \frac{\sum_{i=1}^{n_{\text{test}}} \{\mathbf{y}_i^\top \mathbf{z}_{\gamma_m}^*(\hat{\mathbf{g}}_m(\mathbf{x}_i)) - v^*(\mathbf{y}_i)\}}{\sum_{i=1}^{n_{\text{test}}} |v^*(\mathbf{y}_i)|}.$$

The normalized decision loss compares the realized cost of each deployed decision with this nominal clairvoyant benchmark. We summarize the replication-level ratios over 20 replications with boxplots; all comparisons below refer to their medians. We use normalized decision loss as the primary performance measure and relative prediction loss as a complementary diagnostic of whether improved decisions coincide with more accurate cost predictions.

6.1. Capacitated Transportation Problem

We first consider a capacitated transportation problem with I supply nodes and J demand nodes. We index supply nodes by $s \in [I]$ and demand nodes by $t \in [J]$, and let Q_s and D_t denote their respective supply capacities and demands, which satisfy $\sum_{s \in [I]} Q_s \geq \sum_{t \in [J]} D_t$ to ensure feasibility. The variable z_{st} denotes the amount shipped from supply node s to demand node t . Given a predicted unit-cost vector $\hat{\mathbf{y}} = (\hat{y}_{st})$, the transportation problem is

$$\mathbf{z}_{\gamma}^*(\hat{\mathbf{y}}) \in \arg \min_{\mathbf{z}} \left\{ \hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 : z_{st} \geq 0, \sum_{t \in [J]} z_{st} \leq Q_s, \sum_{s \in [I]} z_{st} = D_t, \forall s \in [I], t \in [J] \right\}.$$

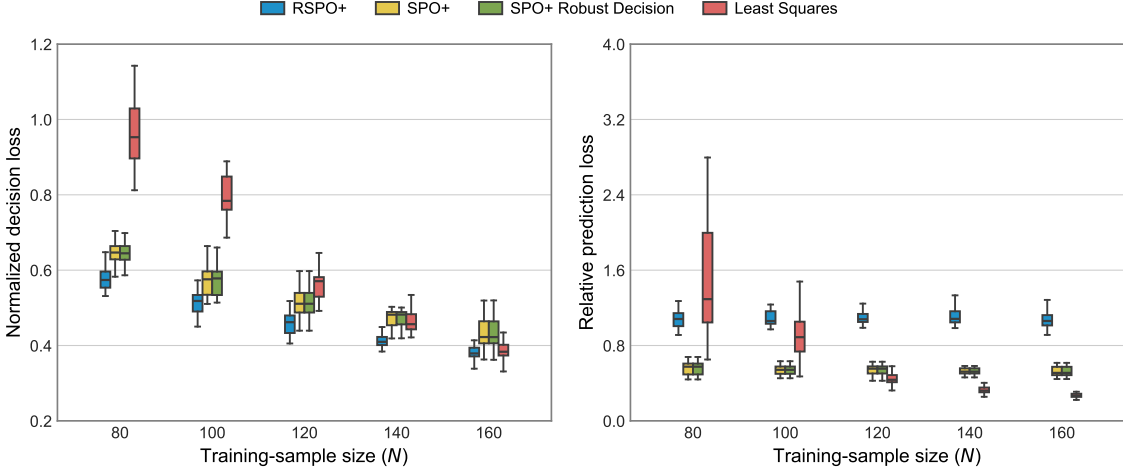


Figure 4 Normalized decision loss (left) and relative prediction loss (right) as the training-sample size N varies on transportation optimization instances, with $p = 60$, $d = 100$, $\text{deg} = 4$, and $\bar{\epsilon} = 0.4$ fixed.

The network has $d = IJ$ arcs, and we identify each arc (s, t) with the coordinate $k = (s - 1)J + t \in [d]$ when forming the cost and decision vectors. For this problem, the unit cost of arc k in observation i is generated as

$$y_{ik} = \left[\left(\frac{1}{\sqrt{p}} (\mathbf{B}^* \mathbf{x}_i)_k + 3 \right)^{\text{deg}} + 1 \right] \epsilon_{ik},$$

where the ϵ_{ik} are independently drawn from $\text{Unif}[1 - \bar{\epsilon}, 1 + \bar{\epsilon}]$. Hence, **deg** controls nonlinearity relative to the linear prediction class and $\bar{\epsilon}$ the noise level.

We organize the transportation results around the four factors considered in our experiments—training-sample size, decision dimension, noise level, and model misspecification—varying one factor at a time while holding the others fixed. Each figure reports normalized decision loss and relative prediction loss to compare downstream decision quality with prediction accuracy.

6.1.1. Effect of Training-Sample Size We first examine how the training-sample size affects out-of-sample performance. We fix $p = 60$, $d = 100$, **deg** = 4, and $\bar{\epsilon} = 0.4$, and vary $N \in \{80, 100, 120, 140, 160\}$. Figure 4 reports the normalized decision loss and relative prediction loss. As N increases, the normalized decision loss of $\mathcal{RSPO}_+$ decreases. It attains the lowest decision loss for $N \leq 140$ and is comparable to the lowest value at $N = 160$. Its margin over $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions is largest for the smaller training samples and becomes smaller as N increases. The prediction panel separates this decision advantage from prediction accuracy: $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions attain lower relative prediction loss, while least squares improves substantially as N increases.

6.1.2. Effect of Decision Dimension We next examine how the decision dimension affects performance by scaling the transportation network. We fix $N = 80$, $p = 60$, **deg** = 4, and $\bar{\epsilon} = 0.4$,

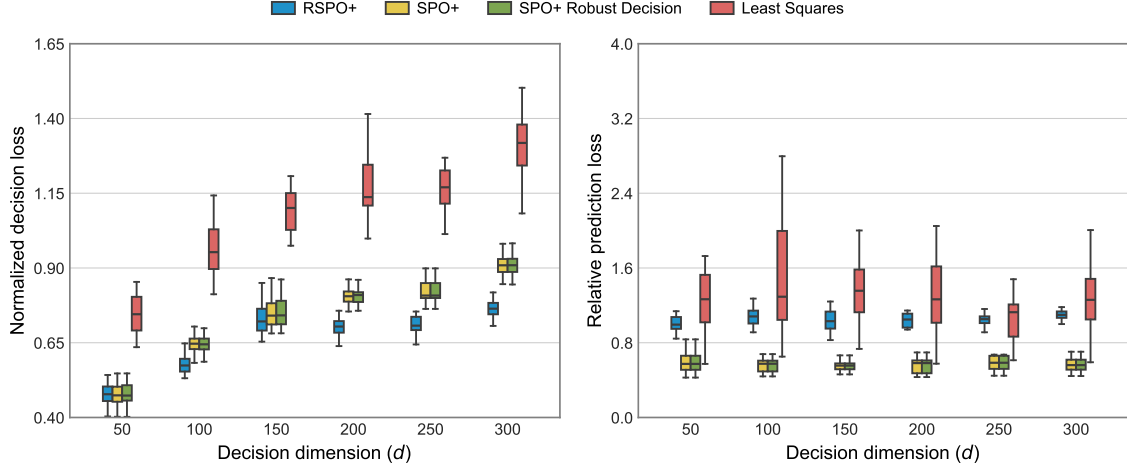


Figure 5 Normalized decision loss (left) and relative prediction loss (right) as the decision dimension d varies on transportation optimization instances, with $N = 80$, $p = 60$, $\text{deg} = 4$, and $\bar{\epsilon} = 0.4$ fixed.

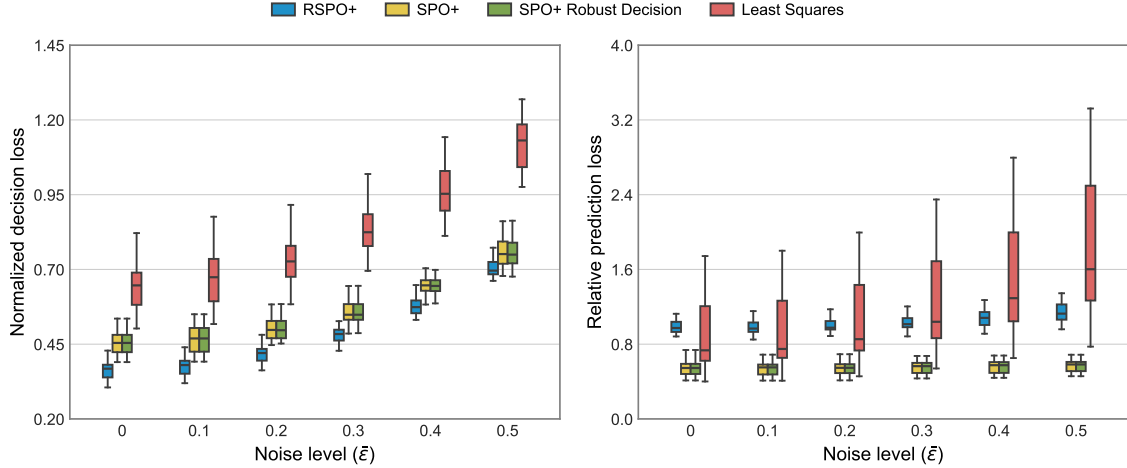


Figure 6 Normalized decision loss (left) and relative prediction loss (right) as the noise level $\bar{\epsilon}$ varies on transportation optimization instances, with $N = 80$, $p = 60$, $d = 100$, and $\text{deg} = 4$ fixed.

and vary $d \in \{50, 100, 150, 200, 250, 300\}$. Figure 5 reports the normalized decision loss and relative prediction loss. As d increases, the normalized decision loss of $\mathcal{RSPO}_+$ increases overall, while its relative advantage becomes more pronounced. $\mathcal{RSPO}_+$, $\mathcal{SPO}_+$, and $\mathcal{SPO}_+$ with robust decisions perform comparably at $d = 50$, and $\mathcal{RSPO}_+$ attains the lowest decision loss for every $d \geq 100$, with its largest margins over $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions at $d \in \{200, 250, 300\}$. Despite having higher relative prediction loss than $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions, $\mathcal{RSPO}_+$ yields lower decision loss throughout $d \geq 100$.

6.1.3. Effect of Noise Level We next examine how the noise level affects performance. We fix $N = 80$, $p = 60$, $d = 100$, and $\text{deg} = 4$, and vary $\bar{\epsilon} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Figure 6 shows that the normalized decision loss of $\mathcal{RSPO}_+$ increases as the noise level rises. It attains the lowest decision loss throughout the reported range, with its largest margins over $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust

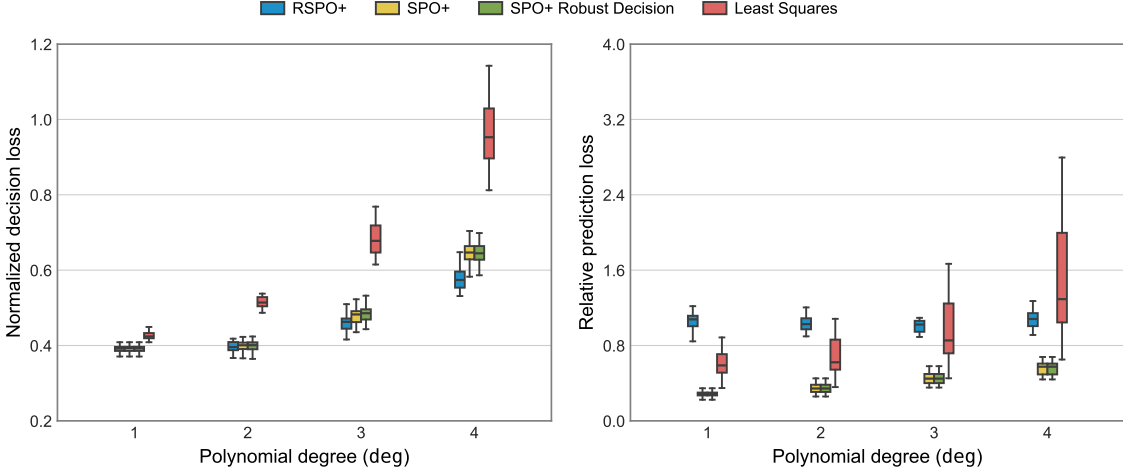


Figure 7 Normalized decision loss (left) and relative prediction loss (right) as model misspecification (deg) varies on transportation optimization instances, with $N = 80$, $p = 60$, $d = 100$, and $\bar{\epsilon} = 0.4$ fixed.

decisions occurring at the lower noise levels. Least squares has the largest decision loss and the widest across-replication variation. $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions attain lower relative prediction loss than $\mathcal{RSPO}_+$, while $\mathcal{RSPO}_+$ attains lower decision loss at every reported noise level.

6.1.4. Effect of Model Misspecification Finally, we examine how model misspecification affects performance. We fix $N = 80$, $p = 60$, $d = 100$, and $\bar{\epsilon} = 0.4$, and vary $\text{deg} \in \{1, 2, 3, 4\}$. Because all methods use a linear prediction class, increasing deg increases the misspecification of the conditional cost function. Figure 7 shows that the normalized decision loss of $\mathcal{RSPO}_+$ increases with misspecification, but less sharply than the losses of the other methods. $\mathcal{RSPO}_+$, $\mathcal{SPO}_+$, and $\mathcal{SPO}_+$ with robust decisions perform comparably at $\text{deg} = 1$; for every nonlinear specification, $\mathcal{RSPO}_+$ attains the lowest decision loss, and its margin over $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions grows with deg . Least squares exhibits the largest deterioration. Despite having higher relative prediction loss than $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions throughout the sweep, $\mathcal{RSPO}_+$ produces better decisions under the nonlinear specifications.

6.1.5. Value of Learning–Decision Alignment To assess the value of learning–decision alignment, we compare $\mathcal{RSPO}_+$ with $\mathcal{SPO}_+$ with robust decisions. Both methods tune their own (λ, γ) , but only $\mathcal{RSPO}_+$ uses the deployed robust decision map to define its training loss. Across Figures 4–7, $\mathcal{RSPO}_+$ attains decision loss that is lower than or comparable to that of $\mathcal{SPO}_+$ with robust decisions. The two methods perform comparably at $d = 50$ and $\text{deg} = 1$, while the difference is more visible at higher decision dimensions and under the nonlinear specifications. These decision gains occur even though $\mathcal{SPO}_+$ with robust decisions attains lower relative prediction loss. Thus, the transportation results support the value of learning–decision alignment.

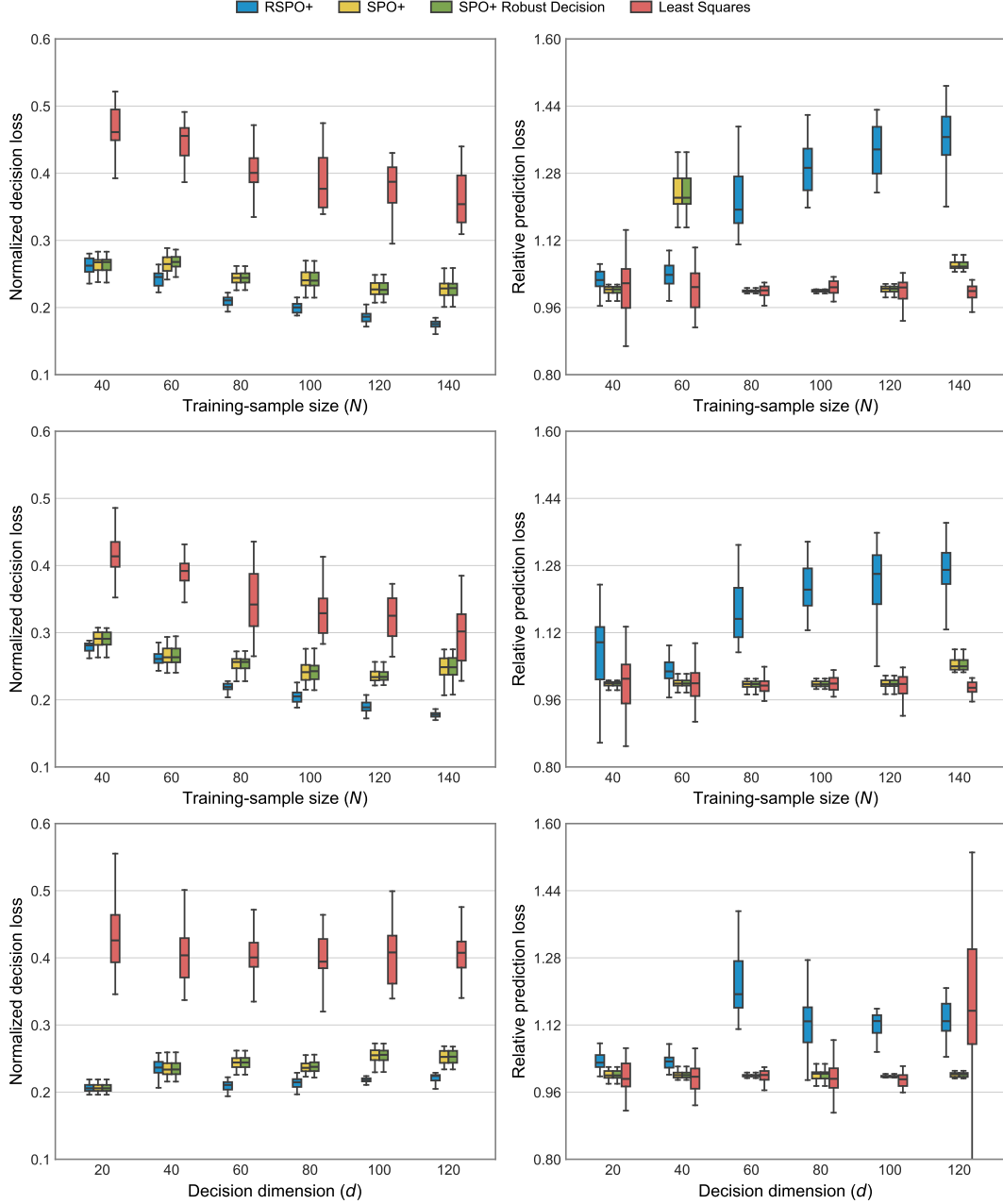


Figure 8 Normalized decision loss (left) and relative prediction loss (right) on portfolio optimization instances. The first and second rows vary the training-sample size N under $\text{deg} = 4$ and $\text{deg} = 1$, respectively, with $p = 80$, $d = 60$, and $\tau = 1$ fixed. The third row varies the number of assets d with $N = 80$, $p = 80$, $\text{deg} = 4$, and $\tau = 1$ fixed.

6.2. Portfolio Optimization Problem

We next study a risk-constrained portfolio problem in which $\mathbf{z}$ specifies the weights assigned to d assets. This experiment examines whether the transportation results extend to a continuous allocation problem with an explicit risk constraint. We generate correlated returns by drawing the entries of a four-factor loading matrix $\mathbf{F} \in \mathbb{R}^{d \times 4}$ independently from $\text{Unif}[-0.0025\tau, 0.0025\tau]$ and setting

$\Sigma = \mathbf{F}\mathbf{F}^\top + (0.01\tau)^2\mathbf{I}_d$. For observation i , the return vector is generated as

$$\mathbf{r}_i = \left(\frac{0.05}{\sqrt{p}} \mathbf{B}^* \mathbf{x}_i + 0.1^{1/\text{deg}} \mathbf{1} \right)^{\text{deg}} + \mathbf{F} \mathbf{l}_i + 0.01\tau \boldsymbol{\epsilon}_i,$$

where $\mathbf{l}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_4)$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ are independent, the power is applied component-wise, and the cost vector is $\mathbf{y}_i = -\mathbf{r}_i$. The terms $\mathbf{F} \mathbf{l}_i$ and $0.01\tau \boldsymbol{\epsilon}_i$ represent common-factor variation and asset-specific noise, respectively, while deg controls misspecification of the common linear prediction class. Given a predicted cost vector $\hat{\mathbf{y}}$, the deployed portfolio solves

$$\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \arg \min_{\mathbf{z}} \left\{ \hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 : \mathbf{z} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{z} \leq 1, \|\Sigma \mathbf{z}\|_1 \leq \beta \right\},$$

where $\|\Sigma \mathbf{z}\|_1$ measures aggregate risk exposure, and β is the risk budget. We use the ℓ_1 risk constraint because its standard epigraph representation is polyhedral, allowing us to apply Theorem 1. Appendix E.1 derives the corresponding learning reformulations, and Appendix E.2 reports the quadratic-risk counterpart.

6.2.1. Effects of Sample Size, Model Misspecification, and Decision Dimension We next examine how performance changes with training-sample size, model misspecification, and the number of assets. Figure 8 summarizes the three experiments. Under $\text{deg} = 4$, the normalized decision loss of $\mathcal{RSP}\mathcal{O}_+$ decreases steadily with N ; it is comparable to $\mathcal{SP}\mathcal{O}_+$ and $\mathcal{SP}\mathcal{O}_+$ with robust decisions at $N = 40$, attains the lowest value for every $N \geq 60$, and develops a more visible margin as N increases. Under $\text{deg} = 1$, $\mathcal{RSP}\mathcal{O}_+$ has the lowest normalized decision loss at every reported N , with a particularly visible margin from $N = 80$ onward. As d varies, $\mathcal{RSP}\mathcal{O}_+$ performs comparably to $\mathcal{SP}\mathcal{O}_+$ and $\mathcal{SP}\mathcal{O}_+$ with robust decisions at $d \in \{20, 40\}$ and attains the lowest normalized decision loss for every $d \geq 60$. Across the three experiments, $\mathcal{SP}\mathcal{O}_+$ and $\mathcal{SP}\mathcal{O}_+$ with robust decisions generally attain lower relative prediction loss, whereas $\mathcal{RSP}\mathcal{O}_+$ attains lower or comparable normalized decision loss.

6.2.2. Value of Learning–Decision Alignment To assess the value of learning–decision alignment, we compare $\mathcal{RSP}\mathcal{O}_+$ with $\mathcal{SP}\mathcal{O}_+$ with robust decisions. Both methods tune their own (λ, γ) , but only $\mathcal{RSP}\mathcal{O}_+$ uses the deployed robust decision map to define its training loss. In Figure 8, $\mathcal{RSP}\mathcal{O}_+$ attains lower or comparable decision loss, with a more visible difference as the sample size and portfolio dimension increase. Thus, the portfolio results reinforce the transportation results and support the value of learning–decision alignment in a continuous allocation problem.

6.2.3. Value of Gradient-Based Refinement Finally, we assess the decision improvement obtained by directly minimizing the empirical $\mathcal{RSP}\mathcal{O}$ objective (14). We first solve the $\mathcal{RSP}\mathcal{O}_+$ empirical risk minimization problem (8) and use the resulting predictor to initialize Algorithm 1.

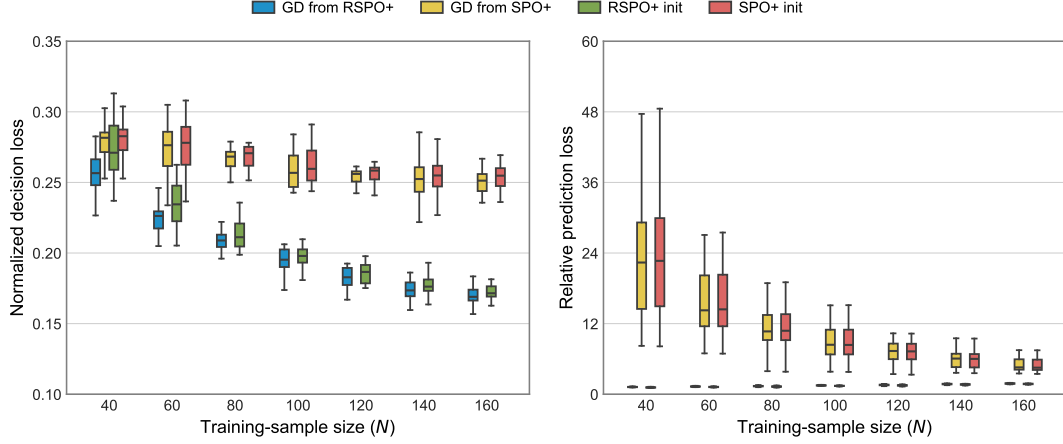


Figure 9 Test decision loss (left) and relative prediction loss (right) before and after gradient-based $RSPO$ refinement as the training-sample size N varies on portfolio optimization instances, with $p = 80$, $d = 40$, $\deg = 4$, and $\tau = 1$ fixed.

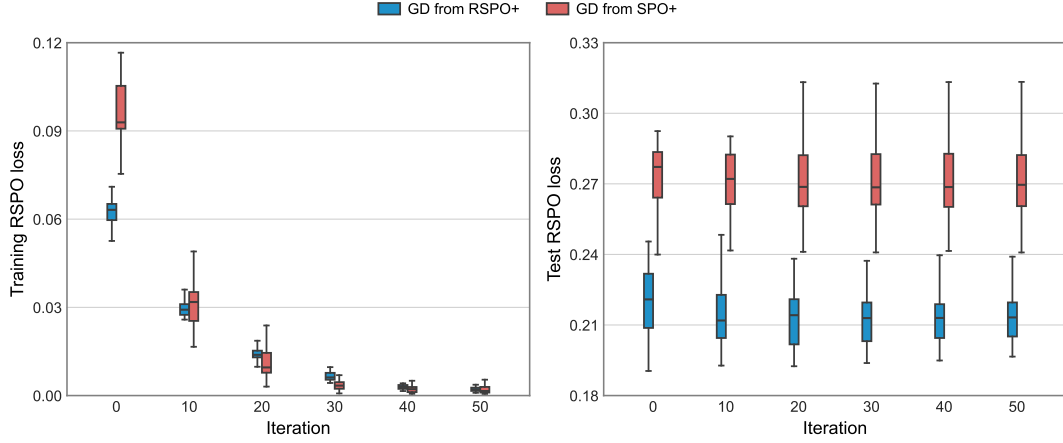


Figure 10 Training $RSPO$ loss (left) and test $RSPO$ loss (right) across gradient iterations on portfolio optimization instances, with $N = 60$, $p = 60$, $d = 40$, $\deg = 4$, and $\tau = 1$ fixed.

We apply the same refinement procedure to an SPO_+ predictor to compare the two starting points and the improvements produced from each. Appendix E.3 provides the complete tuning, validation, and implementation details.

Figure 9 shows that refinement from the $RSPO_+$ predictor lowers test decision loss at every reported N , with the largest gains at $N \in \{40, 60\}$ and the lowest final loss throughout the sweep. Refinement from the SPO_+ predictor also lowers decision loss, with a higher final loss at every N . In both cases, refinement improves decision quality while leaving relative prediction loss broadly unchanged, further separating the decision gain from prediction accuracy. Figure 10 shows that the training $RSPO$ losses decrease rapidly from both initializations and approach nearly the same low level. Most of the test improvement from the $RSPO_+$ initialization occurs within the first ten iterations and is maintained thereafter. Together, these results show that $RSPO_+$ provides

the stronger initialization and that direct optimization of the target $\mathcal{RSP}\mathcal{O}$ loss further improves out-of-sample decision quality in the early iterations.

7. Conclusion

This paper studies the contextual linear optimization problem with robust decisions. To simultaneously achieve both robustness and learning–decision alignment, we propose an integrated learning and robust optimization (ILRO) framework via $\mathcal{RSP}\mathcal{O}$ loss. The resulting learning problem is non-convex, as is the nominal counterpart ($\mathcal{SP}\mathcal{O}$). We address this in two complementary ways: (i) we construct a convex surrogate, the $\mathcal{RSP}\mathcal{O}_+$ loss, and we establish its Fisher consistency when the robustness parameter γ is below an explicitly computable threshold; (ii) unlike the $\mathcal{SP}\mathcal{O}$ loss, the $\mathcal{RSP}\mathcal{O}$ loss carries informative first-order information, and hence we develop a gradient descent algorithm to optimize the target loss directly. On the statistical side, we derive finite-sample excess risk bounds for both predictors: (i) a meta generalization bound for the target predictor that separates the effect of the robust decision mapping from the prediction class, and yields $N^{-1/2}$ rates (up to logarithmic factors) for four representative hypothesis classes; and (ii) within the Fisher-consistent regime, target excess risk bounds of order $N^{-1/4}$ for the surrogate predictor, improving to $N^{-1/2}$ under a quadratic growth condition. Numerical experiments on capacitated transportation and risk-constrained portfolio optimization problems compare the framework against predict-then-optimize, decision-focused, and post-hoc robust benchmarks and support these findings, with the clearest gains under limited samples, high-dimensional decisions, and model misspecification. Gradient-based refinement of the $\mathcal{RSP}\mathcal{O}_+$ solution improves out-of-sample decision quality further.

We outline several potential directions for future research. First, as Proposition 1 shows, under certain conditions, the empirical $\mathcal{RSP}\mathcal{O}$ objective is continuous and piecewise affine. This structural result invites approximation algorithms built from finitely many cuts, which may offer a globally convergent alternative to the local refinement discussed in Section 4.3. Second, the selection of the robustness parameter γ deserves further investigation. Unlike in the predict-then-robust-optimize framework where γ affects only the decision stage, under ILRO it enters the training problem, and hence the decision quality attained at a given γ reflects both the robustness it buys and the predictor it induces. Therefore, a judicious choice of γ should consider the interaction of the two. Third, we focused on an uncertain objective in this paper. Extending the ILRO framework to the case of both predictive uncertain objective and constraints requires novel modeling methods and analysis techniques, which is a natural next step.

References

- Agrawal A, Amos B, Barratt S, Boyd S, Diamond S, Kolter JZ (2019) Differentiable convex optimization layers. *Advances in Neural Information Processing Systems*, volume 32, 9562–9574.

- Amos B, Kolter JZ (2017) OptNet: Differentiable optimization as a layer in neural networks. *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 136–145.
- Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.
- Bartlett PL, Mendelson S (2002) Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.
- Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press).
- Ben-Tal A, Nemirovski A (1999) Robust solutions of uncertain linear programs. *Operations Research Letters* 25(1):1–13.
- Berthet Q, Blondel M, Teboul O, Cuturi M, Vert JP, Bach F (2020) Learning with differentiable perturbed optimizers. *Advances in Neural Information Processing Systems*, volume 33, 9508–9519.
- Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* 66(3):1025–1044.
- Bertsimas D, Sim M (2003) Robust discrete optimization and network flows. *Mathematical Programming* 98(1–3):49–71.
- Bertsimas D, Sim M (2004) The price of robustness. *Operations Research* 52(1):35–53.
- Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press).
- Chenreddy AR, Delage E (2024) End-to-end conditional robust optimization. *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244, 736–748.
- Costa G, Iyengar GN (2023) Distributionally robust end-to-end portfolio construction. *Quantitative Finance* 23(10):1465–1482.
- Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612.
- Donti PL, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*, volume 30, 5484–5494.
- El Balghiti O, Elmachtoub AN, Grigas P, Tewari A (2023) Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research* 48(4):2043–2065.
- Elmachtoub AN, Grigas P (2022) Smart “predict, then optimize”. *Management Science* 68(1):9–26.
- Elmachtoub AN, Lam H, Zhang H, Zhao Y (2023) Estimate-then-optimize versus integrated-estimation-optimization versus sample average approximation: a stochastic dominance perspective. arXiv preprint arXiv:2304.06833.
- Gao R, Kleywegt AJ (2023) Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research* 48(2):603–655.
- Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. *Mathematics of Operations Research* 28(1):1–38.
- Ho-Nguyen N, Kilinc-Karzan F (2022) Risk guarantees for end-to-end prediction and optimization processes. *Management Science* 68(12):8680–8698.
- Hu J, Chen Z, Wang S (2025) Budget-driven multiperiod hub location: A robust time-series approach. *Operations Research* 73(2):613–631.
- Hu Y, Kallus N, Mao X (2022) Fast rates for contextual linear optimization. *Management Science* 68(6):4236–4245.
- Im H, Benslimane W, Grigas P (2025) Smart surrogate losses for contextual stochastic linear optimization with robust constraints. *Advances in Neural Information Processing Systems*, volume 38, 75149–75174.
- Kakade SM, Sridharan K, Tewari A (2008) On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in Neural Information Processing Systems*, volume 21.

-
- Kannan R, Bayraksan G, Luedtke JR (2024) Residuals-based distributionally robust optimization with covariate information. *Mathematical Programming* 207(1):369–425.
- Kannan R, Bayraksan G, Luedtke JR (2025) Technical note—data-driven sample average approximation with covariate information. *Operations Research* 73(6):3245–3259.
- Kuhn D, Shafiee S, Wiesemann W (2025) Distributionally robust optimization. *Acta Numerica* 34:579–804, URL <http://dx.doi.org/10.1017/S0962492924000084>.
- Liu H, Grigas P (2021) Risk bounds and calibration for a smart predict-then-optimize method. *Advances in Neural Information Processing Systems*, volume 34, 22083–22094.
- Lu M, Shen ZJM (2021) A review of robust operations management under model uncertainty. *Production and Operations Management* 30(6):1927–1943.
- Mandi J, Kotary J, Berden S, Mulamba M, Bucarey V, Guns T, Fioretto F (2024) Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research* 80:1623–1701.
- Mao Y, Saldanha-da Gama F, Wang S, Wang S (2026) Predictive production-and-service planning: Ambiguity aversion with performance guarantees. *Management Science* Forthcoming.
- Massart P (2000) Some applications of concentration inequalities to statistics. *Annales de la Faculté des sciences de Toulouse: Mathématiques* 9(2):245–303.
- Maurer A (2006) The Rademacher complexity of linear transformation classes. *Learning Theory*, volume 4005, 65–78.
- Maurer A (2016) A vector-contraction inequality for Rademacher complexities. *Algorithmic Learning Theory*, volume 9925, 3–17.
- Micchelli CA, Pontil M (2005) On learning vector-valued functions. *Neural Computation* 17(1):177–204.
- Mohri M, Rostamizadeh A, Talwalkar A (2018) *Foundations of Machine Learning* (MIT Press), 2nd edition.
- Nocedal J, Wright SJ (2006) *Numerical Optimization* (Springer), 2nd edition.
- Rockafellar RT, Wets RJ (1998) *Variational Analysis* (Springer).
- Sadana U, Chenreddy A, Delage E, Forel A, Frejinger E, Vidal T (2025) A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research* 320(2):271–289.
- Schutte N, Postek K, Yorke-Smith N (2024) Robust losses for decision-focused learning. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 4868–4875.
- Sim M, Tang Q, Zhou M, Zhu T (2025) The analytics of robust satisficing: predict, optimize, satisfice, then fortify. *Operations Research* 73(5):2708–2728.
- Wainwright MJ (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press).
- Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Operations Research* 62(6):1358–1376.
- Wilder B, Dilkina B, Tambe M (2019) Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1):1658–1665.
- Yeh C, Christianson N, Wu A, Wierman A, Yue Y (2025) End-to-end conformal calibration for optimization under uncertainty. *Transactions on Machine Learning Research*.

Electronic Companion to “Integrated Learning and Robust Optimization”

Cheng Tan¹, Yuchen Mao², Shuming Wang¹, and Huan Xu³

¹ School of Economics & Management, University of Chinese Academy of Sciences, China

² Department of Statistics, Rutgers University, United States

³ Antai College of Economics and Management, Shanghai Jiao Tong University, China

tancheng241@mails.ucas.ac.cn, yuchen.mao@rutgers.edu, wangshuming@ucas.edu.cn, xuhuan_antai@sjtu.edu.cn

Appendix A Technical Lemmas

This appendix provides the details of five technical lemmas used throughout our analysis. In particular, Lemma 1 establishes the well-posedness and projection representation of the robust decision map and is invoked in the proofs of Proposition 5 and Lemma EC.2; Lemma 2 is required for Theorem 3; Lemma EC.1 proves the claims made in Remark 1; and Lemma EC.2 gives the closed-form pointwise target excess risk used in the proof of Theorem 4. Lemma EC.3 provides sufficient conditions, in terms of distributional primitives, for the local strong convexity of the surrogate risk invoked in Section 5.2. The proofs of the remaining results are provided in Appendix B.

Proof of Lemma 1. For every $\mathbf{y}_0 \in \mathbb{R}^d$, completing the square gives

$$\mathbf{y}_0^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 = \frac{\gamma}{2} \left\| \mathbf{z} + \frac{\mathbf{y}_0}{\gamma} \right\|_2^2 - \frac{1}{2\gamma} \|\mathbf{y}_0\|_2^2.$$

Because $\mathcal{Z}$ is nonempty and compact and the objective is continuous, a minimizer exists. Since $\gamma > 0$, the objective is strongly convex, so the minimizer is unique. The second term in the display is independent of $\mathbf{z}$; therefore,

$$\mathbf{z}_\gamma^*(\mathbf{y}_0) = \arg \min_{\mathbf{z} \in \mathcal{Z}} \left\| \mathbf{z} + \frac{\mathbf{y}_0}{\gamma} \right\|_2^2 = \Pi_{\mathcal{Z}} \left(-\frac{\mathbf{y}_0}{\gamma} \right),$$

which proves the projection representation. Thus, for any $\mathbf{y}_1, \mathbf{y}_2$, the nonexpansiveness of the Euclidean projection gives

$$\begin{aligned} \|\mathbf{z}_\gamma^*(\mathbf{y}_1) - \mathbf{z}_\gamma^*(\mathbf{y}_2)\|_2 &= \left\| \Pi_{\mathcal{Z}} \left(-\frac{\mathbf{y}_1}{\gamma} \right) - \Pi_{\mathcal{Z}} \left(-\frac{\mathbf{y}_2}{\gamma} \right) \right\|_2 \\ &\leq \left\| -\frac{\mathbf{y}_1}{\gamma} + \frac{\mathbf{y}_2}{\gamma} \right\|_2 \\ &= \frac{1}{\gamma} \|\mathbf{y}_1 - \mathbf{y}_2\|_2. \end{aligned}$$

Therefore, $\mathbf{z}_\gamma^*(\cdot)$ is Lipschitz continuous with constant $1/\gamma$. By Rademacher’s theorem, every Lipschitz continuous mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$ is differentiable almost everywhere. Hence, $\mathbf{z}_\gamma^*(\cdot)$ is differentiable almost everywhere. $\square$

Proof of Lemma 2. Condition on the sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [N]}$. For each $i \in [N]$, define $\psi_i : \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$\psi_i(\mathbf{u}) := \ell_{\mathcal{RSPD}}(\mathbf{u}, \mathbf{y}_i) = \mathbf{y}_i^\top \mathbf{z}_\gamma^*(\mathbf{u}) - v^*(\mathbf{y}_i).$$

By Lemma 1, each ψ_i is Lipschitz continuous with constant $\|\mathbf{y}_i\|_2/\gamma \leq r/\gamma$. By the vector contraction inequality (Maurer 2016, Corollary 4), which allows the Lipschitz functions to depend on the index i , the Rademacher complexity of a vector-valued function class composed with L -Lipschitz scalar functions is bounded, up to a multiplicative factor of $\sqrt{2}L$, by the vector-valued Rademacher complexity of this function class. Since each ψ_i is (r/γ) -Lipschitz, it follows that

$$\mathbb{E}_\sigma \left[\sup_{\mathbf{g} \in \mathcal{G}} \sum_{i \in [N]} \sigma_i \psi_i(\mathbf{g}(\mathbf{x}_i)) \right] \leq \sqrt{2} \frac{r}{\gamma} \mathbb{E}_\sigma \left[\sup_{\mathbf{g} \in \mathcal{G}} \sum_{i \in [N]} \sigma_i^\top \mathbf{g}(\mathbf{x}_i) \right],$$

where on the left-hand side σ_i are independent Rademacher random variables and on the right-hand side $\sigma_i \in \{+1, -1\}^d$ are independent Rademacher random vectors. Dividing both sides by N and taking the expectation with respect to the sample yields

$$\mathfrak{R}_1(\mathcal{H}_{\ell_{\mathcal{RSPD}}}) \leq \frac{\sqrt{2}r}{\gamma} \mathfrak{R}_d(\mathcal{G}).$$

This completes the proof. $\square$

LEMMA EC.1 (Solution-path Equivalence). Fix $\hat{\mathbf{y}} \in \mathbb{R}^d$ and let $\mathcal{Z} \subseteq \mathbb{R}^d$ be nonempty, compact, and convex. For $\vartheta \geq 0$, let $\mathcal{S}_{\text{ball}}(\vartheta) := \arg \min_{\mathbf{z} \in \mathcal{Z}} \{\hat{\mathbf{y}}^\top \mathbf{z} + \vartheta \|\mathbf{z}\|_2\}$. For $\gamma > 0$, let $\mathbf{z}_\gamma^*(\hat{\mathbf{y}})$ be the unique solution of *ILRO-DECISION*. By convention, define $\mathcal{Z}_0(\hat{\mathbf{y}}) = \arg \min_{\mathbf{z} \in \mathcal{Z}} \hat{\mathbf{y}}^\top \mathbf{z}$ and $\mathbf{z}_\infty^*(\hat{\mathbf{y}}) := \Pi_{\mathcal{Z}}(\mathbf{0}) = \lim_{\gamma \rightarrow \infty} \mathbf{z}_\gamma^*(\hat{\mathbf{y}})$. Then the following hold.

- (i) For every $\vartheta \geq 0$ and every $\bar{\mathbf{z}} \in \mathcal{S}_{\text{ball}}(\vartheta)$: if $\bar{\mathbf{z}} \neq \mathbf{0}$ and $\vartheta > 0$, then $\bar{\mathbf{z}} = \mathbf{z}_\gamma^*(\hat{\mathbf{y}})$ with $\gamma = \vartheta/\|\bar{\mathbf{z}}\|_2$; if $\vartheta = 0$, then $\bar{\mathbf{z}} \in \mathcal{Z}_0(\hat{\mathbf{y}})$; if $\bar{\mathbf{z}} = \mathbf{0}$, then $\bar{\mathbf{z}} = \mathbf{z}_\infty^*(\hat{\mathbf{y}})$.
- (ii) For every $\gamma > 0$, $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \mathcal{S}_{\text{ball}}(\vartheta(\gamma))$ with $\vartheta(\gamma) := \gamma \|\mathbf{z}_\gamma^*(\hat{\mathbf{y}})\|_2$.
- (iii) The map $\gamma \mapsto \vartheta(\gamma)$ is nondecreasing on $(0, \infty)$.

Proof. We first provide an elementary comparison fact. Let h be any function on $\mathcal{Z}$ and, for $c_1 < c_2$, let $\mathbf{x}_i$ minimize $\hat{\mathbf{y}}^\top \mathbf{x} + c_i h(\mathbf{x})$ over $\mathcal{Z}$. Summing the two optimality inequalities $\hat{\mathbf{y}}^\top \mathbf{x}_1 + c_1 h(\mathbf{x}_1) \leq \hat{\mathbf{y}}^\top \mathbf{x}_2 + c_1 h(\mathbf{x}_2)$ and $\hat{\mathbf{y}}^\top \mathbf{x}_2 + c_2 h(\mathbf{x}_2) \leq \hat{\mathbf{y}}^\top \mathbf{x}_1 + c_2 h(\mathbf{x}_1)$ yields $(c_2 - c_1)(h(\mathbf{x}_2) - h(\mathbf{x}_1)) \leq 0$, i.e., $h(\mathbf{x}_2) \leq h(\mathbf{x}_1)$.

Part (i). Let $\bar{\mathbf{z}} \in \mathcal{S}_{\text{ball}}(\vartheta)$ with $\bar{\mathbf{z}} \neq \mathbf{0}$ and $\vartheta > 0$. The objective $\mathbf{z} \mapsto \hat{\mathbf{y}}^\top \mathbf{z} + \vartheta \|\mathbf{z}\|_2$ is convex and differentiable at $\bar{\mathbf{z}}$, so optimality over the convex set $\mathcal{Z}$ is equivalent to $\mathbf{0} \in \hat{\mathbf{y}} + \vartheta \bar{\mathbf{z}}/\|\bar{\mathbf{z}}\|_2 + \mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}})$. Setting $\gamma = \vartheta/\|\bar{\mathbf{z}}\|_2$, this reads $\mathbf{0} \in \hat{\mathbf{y}} + \gamma \bar{\mathbf{z}} + \mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}})$, which is the necessary and sufficient optimality condition of *ILRO-DECISION*. Since the objective of *ILRO-DECISION* is strongly convex for $\gamma > 0$,

its minimizer is unique and $\bar{z} = z_\gamma^*(\hat{y})$. If $\vartheta = 0$, the ball problem is the nominal problem and $\bar{z} \in \mathcal{Z}_0(\hat{y})$ by definition. Finally, if $\bar{z} = \mathbf{0} \in \mathcal{S}_{\text{ball}}(\vartheta)$, then $\mathbf{0} \in \mathcal{Z}$ and hence $\Pi_{\mathcal{Z}}(\mathbf{0}) = \mathbf{0}$. Therefore, $\bar{z} = \mathbf{0} = \Pi_{\mathcal{Z}}(\mathbf{0}) = z_\infty^*(\hat{y})$.

Part (ii). Write $\bar{z} = z_\gamma^*(\hat{y})$, so that $\mathbf{0} \in \hat{y} + \gamma\bar{z} + \mathcal{N}_{\mathcal{Z}}(\bar{z})$. If $\bar{z} \neq \mathbf{0}$, then $\gamma\bar{z} = \vartheta(\gamma)\bar{z}/\|\bar{z}\|_2$, so the same inclusion is the (sufficient) optimality condition of the convex ball problem with radius $\vartheta(\gamma)$, and this gives $\bar{z} \in \mathcal{S}_{\text{ball}}(\vartheta(\gamma))$. If $\bar{z} = \mathbf{0}$, the inclusion reads $\mathbf{0} \in \hat{y} + \mathcal{N}_{\mathcal{Z}}(\mathbf{0})$, and this leads to $\bar{z} \in \mathcal{S}_{\text{ball}}(0)$.

Part (iii). Let $0 < \gamma_1 < \gamma_2$ and write $z_i = z_{\gamma_i}^*(\hat{y})$. Applying the comparison fact with $h = \frac{1}{2}\|\cdot\|_2^2$ gives $\|z_2\|_2 \leq \|z_1\|_2$. Suppose, for contradiction, that $\vartheta(\gamma_1) > \vartheta(\gamma_2)$. By part (ii), $z_i \in \mathcal{S}_{\text{ball}}(\vartheta(\gamma_i))$, and the comparison fact with $h = \|\cdot\|_2$ applied to $\vartheta(\gamma_2) < \vartheta(\gamma_1)$ gives $\|z_1\|_2 \leq \|z_2\|_2$. Hence $\|z_1\|_2 = \|z_2\|_2 =: r$, so $\vartheta(\gamma_i) = \gamma_i r$. If $r > 0$, then $\vartheta(\gamma_1) = \gamma_1 r < \gamma_2 r = \vartheta(\gamma_2)$, a contradiction; if $r = 0$, then $\vartheta(\gamma_1) = \vartheta(\gamma_2) = 0$, again a contradiction. Therefore, $\vartheta(\gamma_1) \leq \vartheta(\gamma_2)$. $\square$

LEMMA EC.2. *Suppose $\mathcal{Z}$ is a nonempty compact convex set, $\gamma > 0$, $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\|\mathbf{y}\|_2] < \infty$, and fix $\mathbf{x}$. Define $\Delta_\gamma(\mathbf{c}; \mathbf{x}) := R(\mathbf{c}; \mathbf{x}) - \inf_{\mathbf{c}' \in \mathbb{R}^d} R(\mathbf{c}'; \mathbf{x})$. Recall $\bar{\mathbf{y}} = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\mathbf{y}]$. Then*

$$\Delta_\gamma(\hat{\mathbf{y}}; \mathbf{x}) = \bar{\mathbf{y}}^\top z_\gamma^*(\hat{\mathbf{y}}) - v^*(\bar{\mathbf{y}}), \quad \text{with } v^*(\bar{\mathbf{y}}) = \min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z}.$$

In particular, $\Delta_\gamma(\hat{\mathbf{y}}; \mathbf{x}) \geq 0$, with equality if and only if $z_\gamma^(\hat{\mathbf{y}})$ is optimal for the nominal problem $\min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z}$, that is,*

$$\Delta_\gamma(\hat{\mathbf{y}}; \mathbf{x}) = 0 \Leftrightarrow \hat{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* \Leftrightarrow z_\gamma^*(\hat{\mathbf{y}}) \in \arg \min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z}.$$

Proof of Lemma EC.2. Fix $\mathbf{x}$ and recall $\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbf{y}^\top z_\gamma^*(\hat{\mathbf{y}}) - v^*(\mathbf{y})$. Taking conditional expectation gives

$$\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y})] = \bar{\mathbf{y}}^\top z_\gamma^*(\hat{\mathbf{y}}) - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[v^*(\mathbf{y})],$$

since $z_\gamma^*(\hat{\mathbf{y}})$ does not depend on $\mathbf{y}$. The second term is constant in $\hat{\mathbf{y}}$, so

$$\inf_{\mathbf{c} \in \mathbb{R}^d} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{c}, \mathbf{y})] = \inf_{\mathbf{c} \in \mathbb{R}^d} \bar{\mathbf{y}}^\top z_\gamma^*(\mathbf{c}) - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[v^*(\mathbf{y})].$$

By Lemma 1, $z_\gamma^*(\mathbf{c}) = \Pi_{\mathcal{Z}}(-\mathbf{c}/\gamma)$, and since the Euclidean projection maps $\mathbb{R}^d$ onto $\mathcal{Z}$ (every $\mathbf{z}_0 \in \mathcal{Z}$ satisfies $\Pi_{\mathcal{Z}}(\mathbf{z}_0) = \mathbf{z}_0$), the image $\{z_\gamma^*(\mathbf{c}) : \mathbf{c} \in \mathbb{R}^d\}$ equals $\mathcal{Z}$. Hence $\inf_{\mathbf{c}} \bar{\mathbf{y}}^\top z_\gamma^*(\mathbf{c}) = \min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z} = v^*(\bar{\mathbf{y}})$. Subtracting, the constant term cancels and

$$\Delta_\gamma(\hat{\mathbf{y}}; \mathbf{x}) = \bar{\mathbf{y}}^\top z_\gamma^*(\hat{\mathbf{y}}) - v^*(\bar{\mathbf{y}}).$$

This is nonnegative because $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \mathcal{Z}$, and equals zero if and only if $\mathbf{z}_\gamma^*(\hat{\mathbf{y}})$ attains $\min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z}$. By the definition, $\Delta_\gamma(\hat{\mathbf{y}}; \mathbf{x}) = 0$ holds exactly when $\hat{\mathbf{y}}$ minimizes the pointwise $\mathcal{RSP}\mathcal{O}$ risk, i.e., $\hat{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$. Therefore,

$$\Delta_\gamma(\hat{\mathbf{y}}; \mathbf{x}) = 0 \Leftrightarrow \hat{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* \Leftrightarrow \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \arg \min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z}.$$

This completes the proof. $\square$

LEMMA EC.3 (Sufficient Conditions for Local Strong Convexity of Surrogate Risk).

Suppose $\mathcal{Z}$ is a nonempty compact convex set, $\gamma > 0$, and $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\|\mathbf{y}\|_2] < \infty$. Set $a = 1$ and fix $\mathbf{x}$. Let $\rho > 0$ be such that $\mathcal{Z}$ contains a Euclidean ball of radius strictly larger than ρ/γ , and define

$$K_\rho := \{\mathbf{w} \in \mathbb{R}^d : \mathcal{B}(\mathbf{w}, \rho/\gamma) \subseteq \mathcal{Z}\}, \quad p_0 := \mathbb{P}_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \in 2\bar{\mathbf{y}} + \gamma K_\rho).$$

Then $R_+(\cdot; \mathbf{x})$ is (p_0/γ) -strongly convex on $\mathcal{B}(2\bar{\mathbf{y}}, \rho)$. In particular, $p_0 > 0$ whenever $\mathbb{P}_{\mathbf{y}|\mathbf{x}}$ admits a density that is positive on an open set containing $2\bar{\mathbf{y}} + \gamma K_\rho$.

Proof of Lemma EC.3. We prove that the gradient of $R_+(\cdot; \mathbf{x})$ is (p_0/γ) -strongly monotone on $\mathcal{B}(2\bar{\mathbf{y}}, \rho)$, which is equivalent to the claimed strong convexity.

Step 1: gradient formula. By the proof of Proposition 5 (with $a = 1$), $R_+(\cdot; \mathbf{x})$ is finite, convex, and differentiable on $\mathbb{R}^d$ with

$$\nabla_{\mathbf{y}} R_+(\hat{\mathbf{y}}; \mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\mathbf{z}_\gamma^*(\mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\mathbf{z}_\gamma^*(\hat{\mathbf{y}} - \mathbf{y})], \quad \mathbf{z}_\gamma^*(\mathbf{c}) = \Pi_{\mathcal{Z}}\left(-\frac{\mathbf{c}}{\gamma}\right).$$

Note that the first expectation does not depend on $\hat{\mathbf{y}}$ and will cancel in gradient differences.

Step 2: the monotonicity pairing as an expectation. Fix $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2 \in \mathcal{B}(2\bar{\mathbf{y}}, \rho)$ and set $\mathbf{w}_i(\mathbf{y}) := (\mathbf{y} - \hat{\mathbf{y}}_i)/\gamma$, so that $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}_i - \mathbf{y}) = \Pi_{\mathcal{Z}}(\mathbf{w}_i(\mathbf{y}))$ and $\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 = \gamma(\mathbf{w}_2(\mathbf{y}) - \mathbf{w}_1(\mathbf{y}))$ for every $\mathbf{y}$. By Step 1,

$$\langle \nabla_{\mathbf{y}} R_+(\hat{\mathbf{y}}_1; \mathbf{x}) - \nabla_{\mathbf{y}} R_+(\hat{\mathbf{y}}_2; \mathbf{x}), \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \rangle = \gamma \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\langle \Pi_{\mathcal{Z}}(\mathbf{w}_2) - \Pi_{\mathcal{Z}}(\mathbf{w}_1), \mathbf{w}_2 - \mathbf{w}_1 \rangle].$$

Step 3: lower bound on the expectation. Applying firm nonexpansiveness of the Euclidean projection, $\langle \Pi_{\mathcal{Z}}(\mathbf{u}) - \Pi_{\mathcal{Z}}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq \|\Pi_{\mathcal{Z}}(\mathbf{u}) - \Pi_{\mathcal{Z}}(\mathbf{v})\|_2^2 \geq 0$. Hence the expectation in Step 2 is non-negative. Moreover, on the event $E := \{\mathbf{y} : \mathbf{w}_1(\mathbf{y}) \in \mathcal{Z} \text{ and } \mathbf{w}_2(\mathbf{y}) \in \mathcal{Z}\}$, the projections act as the identity, so the integrand over E equals $\|\mathbf{w}_2 - \mathbf{w}_1\|_2^2 = \|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\|_2^2/\gamma^2$. Hence,

$$\langle \nabla R_+(\hat{\mathbf{y}}_1) - \nabla R_+(\hat{\mathbf{y}}_2), \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \rangle \geq \frac{\mathbb{P}(E)}{\gamma} \|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\|_2^2.$$

Step 4: the event E contains $\{\mathbf{y} \in 2\bar{\mathbf{y}} + \gamma K_\rho\}$. For any $\mathbf{y} \in 2\bar{\mathbf{y}} + \gamma K_\rho$, we have $(\mathbf{y} - 2\bar{\mathbf{y}})/\gamma \in K_\rho$, and since $\|\hat{\mathbf{y}}_i - 2\bar{\mathbf{y}}\|_2 \leq \rho$,

$$\mathbf{w}_i(\mathbf{y}) = \frac{\mathbf{y} - 2\bar{\mathbf{y}}}{\gamma} + \frac{2\bar{\mathbf{y}} - \hat{\mathbf{y}}_i}{\gamma} \in \mathcal{B}\left(\frac{\mathbf{y} - 2\bar{\mathbf{y}}}{\gamma}, \frac{\rho}{\gamma}\right) \subseteq \mathcal{Z}, \quad i = 1, 2$$

by the definition of K_ρ . Hence, $E \supseteq \{\mathbf{y} \in 2\bar{\mathbf{y}} + \gamma K_\rho\}$ and $\mathbb{P}(E) \geq p_0$, so that

$$\langle \nabla R_+(\hat{\mathbf{y}}_1) - \nabla R_+(\hat{\mathbf{y}}_2), \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \rangle \geq \frac{p_0}{\gamma} \|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\|_2^2 \quad \forall \hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2 \in \mathcal{B}(2\bar{\mathbf{y}}, \rho).$$

Step 5: strong convexity and positivity of p_0 . For a differentiable convex function, (p_0/γ) -strong monotonicity of the gradient on the convex set $\mathcal{B}(2\bar{\mathbf{y}}, \rho)$ is equivalent to (p_0/γ) -strong convexity of R_+ on that set, which proves the main claim. Finally, if $\mathcal{Z}$ contains a ball $\mathcal{B}(\mathbf{z}_0, \rho/\gamma + \eta)$ with $\eta > 0$, then $\mathcal{B}(\mathbf{z}_0, \eta) \subseteq K_\rho$, so $2\bar{\mathbf{y}} + \gamma K_\rho$ has nonempty interior. If in addition the conditional density of $\mathbf{y}$ is positive on an open set containing $2\bar{\mathbf{y}} + \gamma K_\rho$, then $p_0 > 0$. $\square$

Appendix B Technical Proofs

This appendix presents the proofs of the main results stated in the paper, following the same order as their appearance in the main text.

Proof of Proposition 1. By Lemma 1, $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \Pi_{\mathcal{Z}}(-\hat{\mathbf{y}}/\gamma)$ is single-valued and $(1/\gamma)$ -Lipschitz continuous, so the unregularized empirical $\mathcal{RSP}\mathcal{O}$ objective is continuous. It remains to establish the piecewise affine structure. Call $S \subseteq [m]$ *regular* if the rows $\{\mathbf{a}_j\}_{j \in S}$ of $\mathbf{A}$ are linearly independent (this includes $S = \emptyset$). For a regular S , define the affine maps

$$\mathbf{z}_S(\hat{\mathbf{y}}) = \frac{1}{\gamma} (\mathbf{A}_S^\top (\mathbf{A}_S \mathbf{A}_S^\top)^{-1} \mathbf{A}_S - \mathbf{I}) \hat{\mathbf{y}} + \mathbf{A}_S^\top (\mathbf{A}_S \mathbf{A}_S^\top)^{-1} \mathbf{b}_S, \quad \boldsymbol{\lambda}_S(\hat{\mathbf{y}}) = (\mathbf{A}_S \mathbf{A}_S^\top)^{-1} (\gamma \mathbf{b}_S + \mathbf{A}_S \hat{\mathbf{y}}),$$

with the convention $\mathbf{z}_\emptyset(\hat{\mathbf{y}}) = -\hat{\mathbf{y}}/\gamma$, and the closed polyhedron

$$\mathcal{Q}_S := \{\hat{\mathbf{y}} \in \mathbb{R}^d : \mathbf{a}_j^\top \mathbf{z}_S(\hat{\mathbf{y}}) \geq b_j \ (j \notin S), \ \boldsymbol{\lambda}_S(\hat{\mathbf{y}}) \geq \mathbf{0}\},$$

where all defining conditions are affine in $\hat{\mathbf{y}}$. A direct computation gives $\mathbf{A}_S \mathbf{z}_S(\hat{\mathbf{y}}) = \mathbf{b}_S$ and the stationarity identity $\gamma \mathbf{z}_S(\hat{\mathbf{y}}) + \hat{\mathbf{y}} = \mathbf{A}_S^\top \boldsymbol{\lambda}_S(\hat{\mathbf{y}})$ for every $\hat{\mathbf{y}}$. Hence, for $\hat{\mathbf{y}} \in \mathcal{Q}_S$, the pair $(\mathbf{z}_S(\hat{\mathbf{y}}), \boldsymbol{\lambda}_S(\hat{\mathbf{y}}))$ (with multipliers extended by zero off S) satisfies primal feasibility, dual feasibility, complementary slackness, and stationarity for the strongly convex problem $\min_{\mathbf{A}\mathbf{z} \geq \mathbf{b}} \hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2$. Since the constraints are affine, the KKT conditions are sufficient, and by strong convexity the minimizer is unique. Therefore, $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \mathbf{z}_S(\hat{\mathbf{y}})$ when $\hat{\mathbf{y}} \in \mathcal{Q}_S$.

We next show that the sets $\{\mathcal{Q}_S : S \text{ is regular}\}$ cover $\mathbb{R}^d$. Fix $\hat{\mathbf{y}}$, write $\mathbf{z}^* = \mathbf{z}_\gamma^*(\hat{\mathbf{y}})$, and let $S_{\hat{\mathbf{y}}}$ be its active set. By the KKT conditions (which are also necessary here, the constraints being affine), the multiplier set $\{\boldsymbol{\lambda} \in \mathbb{R}_+^m : \mathbf{A}^\top \boldsymbol{\lambda} = \gamma \mathbf{z}^* + \hat{\mathbf{y}}, \ \lambda_j = 0 \ (j \notin S_{\hat{\mathbf{y}}})\}$ is nonempty. It is closed, convex, and contains no line, and hence possesses an extreme point $\hat{\boldsymbol{\lambda}}$. By the standard characterization of extreme points of polyhedra in standard form, the rows $\{\mathbf{a}_j\}_{j \in S}$ with $S := \text{supp}(\hat{\boldsymbol{\lambda}}) \subseteq S_{\hat{\mathbf{y}}}$ are linearly independent, so S is regular. Since $\mathbf{A}_S \mathbf{z}^* = \mathbf{b}_S$ and $\mathbf{A}_S^\top \hat{\boldsymbol{\lambda}}_S = \gamma \mathbf{z}^* + \hat{\mathbf{y}}$, left-multiplying the latter by $(\mathbf{A}_S \mathbf{A}_S^\top)^{-1} \mathbf{A}_S$ yields $\hat{\boldsymbol{\lambda}}_S = \boldsymbol{\lambda}_S(\hat{\mathbf{y}}) \geq \mathbf{0}$, and substituting back gives $\mathbf{z}^* = \mathbf{z}_S(\hat{\mathbf{y}})$. Feasibility of $\mathbf{z}^*$ then shows $\hat{\mathbf{y}} \in \mathcal{Q}_S$. As S ranges over the finitely many regular subsets of $[m]$, the closed polyhedra $\{\mathcal{Q}_S\}$ therefore cover $\mathbb{R}^d$, and $\mathbf{z}_\gamma^*$ is continuous and piecewise affine.

Now fix $i \in [N]$. The map $\mathbf{B} \mapsto \mathbf{B}\mathbf{x}_i$ is linear, and the preimage of a polyhedron under a linear map is a polyhedron, so $\mathbf{B} \mapsto \mathbf{y}_i^\top \mathbf{z}_\gamma^*(\mathbf{B}\mathbf{x}_i)$ is continuous and is affine on each member of the finite polyhedral family $\{\mathbf{B} : \mathbf{B}\mathbf{x}_i \in \mathcal{Q}_S\}_S$ is regular. Taking the common refinement of these N families yields a finite polyhedral subdivision of $\mathbb{R}^{d \times p}$ on each cell of which every summand, and hence their average, is affine. Subtracting the constant $\frac{1}{N} \sum_{i \in [N]} v^*(\mathbf{y}_i)$ preserves this. Discarding the cells with empty interior leaves a finite family of full-dimensional polyhedra whose union is still $\mathbb{R}^{d \times p}$, because the discarded cells are contained in the union of the finitely many hyperplanes defining the subdivision. The gradient on the interior of each cell is the constant linear coefficient of the corresponding affine piece. $\square$

B.1 Proofs of Section 4

Proof of Proposition 2. For fixed $\mathbf{z} \in \mathcal{Z}$, the function

$$a \mapsto \mathbf{y}^\top \mathbf{z} - a \left(\hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 - v_\gamma^*(\hat{\mathbf{y}}) \right)$$

is affine. Hence $q(\cdot)$, as the pointwise maximum of affine functions, is convex on $(0, \infty)$. Let $\mathbf{z}_a$ be the unique maximizer in the definition of $q(a)$ for $a > 0$. By Danskin's theorem, the derivative of q at a is

$$q'(a) = v_\gamma^*(\hat{\mathbf{y}}) - \hat{\mathbf{y}}^\top \mathbf{z}_a - \frac{\gamma}{2} \|\mathbf{z}_a\|_2^2.$$

Since $v_\gamma^*(\hat{\mathbf{y}})$ is the minimum of $\hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2$ over $\mathcal{Z}$, we have $q'(a) \leq 0$. Thus, q is nonincreasing.

It remains to show the limiting representation. Write $h(\mathbf{z}) := \hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 - v_\gamma^*(\hat{\mathbf{y}}) \geq 0$, where equality holds if and only if $\mathbf{z} = \mathbf{z}_\gamma^*(\hat{\mathbf{y}})$, because $\gamma > 0$ makes the minimizer unique. Then, $q(a) = \max_{\mathbf{z} \in \mathcal{Z}} \{\mathbf{y}^\top \mathbf{z} - ah(\mathbf{z})\}$. Let the diameter $D_{\mathcal{Z}} < \infty$ and $M := \sup_{\mathbf{z} \in \mathcal{Z}} |\mathbf{y}^\top \mathbf{z}| < \infty$. For any $\varepsilon > 0$, compactness and uniqueness imply that

$$\eta_\varepsilon := \inf_{\mathbf{z} \in \mathcal{Z} : \|\mathbf{z} - \mathbf{z}_\gamma^*(\hat{\mathbf{y}})\|_2 \geq \varepsilon} h(\mathbf{z}) > 0.$$

For $a > 2M/\eta_\varepsilon$, any maximizer of $q(a)$ must lie within ε of $\mathbf{z}_\gamma^*(\hat{\mathbf{y}})$; otherwise its objective value is at most $M - a\eta_\varepsilon < -M$, whereas evaluating at $\mathbf{z}_\gamma^*(\hat{\mathbf{y}})$ gives $\mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \geq -M$. Therefore, every sequence of maximizers converges to $\mathbf{z}_\gamma^*(\hat{\mathbf{y}})$ as $a \rightarrow \infty$, i.e., $\mathbf{z}_a \rightarrow \mathbf{z}_\gamma^*(\hat{\mathbf{y}})$. Note that $\mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \leq q(a) \leq \mathbf{y}^\top \mathbf{z}_a$ and by continuity, we have

$$\lim_{a \rightarrow +\infty} q(a) = \mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}).$$

Subtracting $v^*(\mathbf{y})$ gives the stated representation of $\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y})$. $\square$

Proof of Proposition 3. For part (i), the inequality $\ell_{\mathcal{RSP}\mathcal{O}} \leq \ell_{\mathcal{RSP}\mathcal{O}_+}$ is the construction preceding the definition in the main text. Nonnegativity holds because $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \mathcal{Z}$, so that

$$\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) - v^*(\mathbf{y}) \geq \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{y}^\top \mathbf{z} - v^*(\mathbf{y}) = 0.$$

For part (ii), for each fixed $\mathbf{z} \in \mathcal{Z}$, the term $\mathbf{y}^\top \mathbf{z} - a\hat{\mathbf{y}}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2$ is affine in $\hat{\mathbf{y}}$. Taking the maximum over $\mathbf{z} \in \mathcal{Z}$ preserves convexity, and the remaining terms in $\ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y})$ are affine or constant in $\hat{\mathbf{y}}$. Hence $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ is convex.

For part (iii), define

$$\psi(\hat{\mathbf{y}}) := \max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - a\hat{\mathbf{y}}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\}.$$

The maximizer in $\psi(\hat{\mathbf{y}})$ is unique because the objective is strongly concave in $\mathbf{z}$, and it is

$$\arg \max_{\mathbf{z} \in \mathcal{Z}} \left\{ (\mathbf{y} - a\hat{\mathbf{y}})^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} = \mathbf{z}_\gamma^* \left(\hat{\mathbf{y}} - \frac{1}{a} \mathbf{y} \right).$$

By Danskin's theorem,

$$\nabla \psi(\hat{\mathbf{y}}) = -a\mathbf{z}_\gamma^* \left(\hat{\mathbf{y}} - \frac{1}{a}\mathbf{y} \right).$$

Adding the gradient of the affine term $a\mathbf{z}_\gamma^*(\mathbf{y})^\top \hat{\mathbf{y}}$ gives

$$\nabla_{\hat{\mathbf{y}}} \ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y}) = a \left(\mathbf{z}_\gamma^*(\mathbf{y}) - \mathbf{z}_\gamma^* \left(\hat{\mathbf{y}} - \frac{1}{a}\mathbf{y} \right) \right).$$

The uniqueness of the maximizer and Danskin's theorem establish differentiability at every prediction, so the displayed vector is the gradient everywhere.

For part (iv), by (ii)–(iii) the map $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ is convex and differentiable, and its gradient is a times the difference of two points of $\mathcal{Z}$, hence uniformly bounded in norm by $aD_{\mathcal{Z}}$. The mean value theorem then yields the claimed $aD_{\mathcal{Z}}$ -Lipschitz continuity. $\square$

Proof of Theorem 1. It suffices to reformulate the maximization term in each summand of $\ell_{\mathcal{RSP}\mathcal{O}_+}$. Fix a sample i and write $\mathbf{c}_i := \mathbf{y}_i - a\mathbf{g}_\theta(\mathbf{x}_i)$. The optimization term in the surrogate loss is

$$\max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{c}_i^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} \quad \text{with } \mathcal{Z} = \{\mathbf{z} : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}.$$

This is a concave quadratic maximization problem over a nonempty bounded polyhedron. Introduce a multiplier $\mathbf{p}_i \geq \mathbf{0}$ for the constraint $\mathbf{b} - \mathbf{A}\mathbf{z} \leq \mathbf{0}$. The Lagrangian for the maximization problem is

$$L_i(\mathbf{z}, \mathbf{p}_i) = \mathbf{c}_i^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 - \mathbf{p}_i^\top (\mathbf{b} - \mathbf{A}\mathbf{z}) = -\mathbf{b}^\top \mathbf{p}_i + (\mathbf{c}_i + \mathbf{A}^\top \mathbf{p}_i)^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2.$$

For fixed $\mathbf{p}_i$, maximizing over $\mathbf{z} \in \mathbb{R}^d$ gives

$$\sup_{\mathbf{z} \in \mathbb{R}^d} L_i(\mathbf{z}, \mathbf{p}_i) = -\mathbf{b}^\top \mathbf{p}_i + \frac{1}{2a\gamma} \|\mathbf{c}_i + \mathbf{A}^\top \mathbf{p}_i\|_2^2,$$

where the supremum is attained at $\mathbf{z} = \frac{1}{a\gamma}(\mathbf{c}_i + \mathbf{A}^\top \mathbf{p}_i)$. Strong duality and dual attainment hold because the primal is a concave quadratic maximization over a nonempty bounded polyhedron. Indeed, for convex quadratic programs with affine constraints, strong duality requires no Slater condition (see, e.g., [Boyd and Vandenberghe 2004](#), Section 5.2.3). Therefore,

$$\max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{c}_i^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} = \min_{\mathbf{p}_i \geq \mathbf{0}} \left\{ -\mathbf{b}^\top \mathbf{p}_i + \frac{1}{2a\gamma} \|\mathbf{y}_i - a\mathbf{g}_\theta(\mathbf{x}_i) + \mathbf{A}^\top \mathbf{p}_i\|_2^2 \right\}.$$

Substituting this identity into each $\mathcal{RSP}\mathcal{O}_+$ summand and retaining the remaining terms

$$a\mathbf{z}_\gamma^*(\mathbf{y}_i)^\top \mathbf{g}_\theta(\mathbf{x}_i) + \frac{a\gamma}{2} \|\mathbf{z}_\gamma^*(\mathbf{y}_i)\|_2^2 - v^*(\mathbf{y}_i)$$

yields the stated finite-dimensional reformulation. Since the inner minimum over $\mathbf{p}_i \in \mathbb{R}_+^m$ is attained for every θ , minimizing the reformulation jointly over $(\theta, \{\mathbf{p}_i\})$ is equivalent to minimizing the

partially minimized objective (the $\mathcal{RSP}\mathcal{O}_+$ objective) over θ alone. The two problems therefore share the same optimal value, and θ^* is optimal for Problem (8) if and only if $(\theta^*, \{\mathbf{p}_i^*\})$ is optimal for the reformulation, with $\{\mathbf{p}_i^*\}$ the attaining multipliers. The regularizer $\lambda\Omega(\mathbf{g}_\theta)$ and the constraint $\theta \in \Theta$ are unchanged. $\square$

Proof of Proposition 4. Throughout this proof, $\bar{\mathbf{z}} := \mathbf{z}^*(\bar{\mathbf{y}})$ denotes the fixed vector given by the unique nominal solution at the conditional mean $\bar{\mathbf{y}}$.

Step 1: pointwise risk and lower bound. For any $\hat{\mathbf{y}}$, the pointwise $\mathcal{RSP}\mathcal{O}$ risk at the given $\mathbf{x}$ satisfies

$$\begin{aligned} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y})] &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) - v^*(\mathbf{y})] \\ &= \bar{\mathbf{y}}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{y})] \\ &\geq \bar{\mathbf{y}}^\top \bar{\mathbf{z}} - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{y})], \end{aligned}$$

where the second equality holds because $\mathbf{z}_\gamma^*(\hat{\mathbf{y}})$ is independent of $\mathbf{y}$, and the inequality holds because $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \mathcal{Z}$ while $\bar{\mathbf{z}}$ minimizes $\bar{\mathbf{y}}^\top \mathbf{z}$ over $\mathcal{Z}$.

Step 2: the lower bound is attained. Consider the candidate prediction $\hat{\mathbf{y}}_0 := -\gamma\bar{\mathbf{z}}$. Completing the square, the robust problem at $\hat{\mathbf{y}}_0$ becomes

$$\min_{\mathbf{z} \in \mathcal{Z}} \left\{ \hat{\mathbf{y}}_0^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 \right\} = \min_{\mathbf{z} \in \mathcal{Z}} \frac{\gamma}{2} \|\mathbf{z} - \bar{\mathbf{z}}\|_2^2 - \frac{\gamma}{2} \|\bar{\mathbf{z}}\|_2^2,$$

i.e., it finds the point of $\mathcal{Z}$ closest to $\bar{\mathbf{z}}$. Since $\bar{\mathbf{z}} \in \mathcal{Z}$, the unique minimizer is $\bar{\mathbf{z}}$ itself, so $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}_0) = \bar{\mathbf{z}}$ and the risk at $\hat{\mathbf{y}}_0$ equals the lower bound of Step 1. Hence the infimum of the pointwise risk coincides with that bound, and because $\bar{\mathbf{z}}$ is the unique minimizer of $\bar{\mathbf{y}}^\top \mathbf{z}$ over $\mathcal{Z}$, equality in Step 1 holds precisely when $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \bar{\mathbf{z}}$. Therefore,

$$\hat{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* \iff \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \bar{\mathbf{z}}.$$

Step 3: normal-cone characterization. The condition $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \bar{\mathbf{z}}$ states that $\bar{\mathbf{z}}$ minimizes the convex differentiable function $\mathbf{z} \mapsto \hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2$ over the closed convex set $\mathcal{Z}$. By the first-order optimality condition for convex problems, this holds if and only if

$$-(\hat{\mathbf{y}} + \gamma\bar{\mathbf{z}}) \in \mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}}) \iff \hat{\mathbf{y}} \in -\mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}}) - \gamma\bar{\mathbf{z}}.$$

Combined with the equivalence of Step 2, this proves the general representation (11):

$$\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* = -\mathcal{N}_{\mathcal{Z}}(\mathbf{z}^*(\bar{\mathbf{y}})) - \gamma\mathbf{z}^*(\bar{\mathbf{y}}).$$

Step 4: polyhedral specialization. Suppose now that $\mathcal{Z} = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}$ is a bounded polyhedron. By Step 3, it suffices to verify $-\mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}}) = \{\mathbf{A}^\top \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, which we do via the KKT conditions. The condition $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \bar{\mathbf{z}}$ holds if and only if the KKT system

$$\begin{aligned} \hat{\mathbf{y}} + \gamma \bar{\mathbf{z}} - \mathbf{A}^\top \boldsymbol{\lambda} &= \mathbf{0}, \\ \mathbf{A} \bar{\mathbf{z}} &\geq \mathbf{b}, \quad \boldsymbol{\lambda} \geq \mathbf{0}, \\ \lambda_i (\mathbf{a}_i^\top \bar{\mathbf{z}} - b_i) &= 0, \quad i = 1, \dots, m, \end{aligned}$$

admits a multiplier $\boldsymbol{\lambda}$. Feasibility $\mathbf{A} \bar{\mathbf{z}} \geq \mathbf{b}$ holds automatically. Complementary slackness forces $\lambda_i = 0$ for every constraint inactive at $\bar{\mathbf{z}}$, i.e., $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$; conversely, for any $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ complementary slackness holds by the definition of $\boldsymbol{\Lambda}$. Hence the KKT system is solvable if and only if the stationarity equation can be met with some $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, that is,

$$\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) = \bar{\mathbf{z}} \iff \hat{\mathbf{y}} = \mathbf{A}^\top \boldsymbol{\lambda} - \gamma \bar{\mathbf{z}} \text{ for some } \boldsymbol{\lambda} \in \boldsymbol{\Lambda}.$$

(Note that $\hat{\mathbf{y}}_0$ of Step 2 corresponds to the choice $\boldsymbol{\lambda} = \mathbf{0} \in \boldsymbol{\Lambda}$.)

Combining the two displayed equivalences,

$$\mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^* = \{\mathbf{A}^\top \boldsymbol{\lambda} - \gamma \bar{\mathbf{z}} : \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\},$$

and comparing with Step 3 shows $-\mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}}) = \{\mathbf{A}^\top \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, which is precisely the polyhedral description asserted in the proposition with $\bar{\mathbf{z}} = \mathbf{z}^*(\bar{\mathbf{y}})$. $\square$

Proof of Proposition 5. Step 1: finiteness. Recall $R_{\mathcal{Z}} = \max_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z}\|_2 < \infty$. For any fixed $\hat{\mathbf{y}}$ and any $\mathbf{y}$, every inner product appearing in $\ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y})$ is bounded via Cauchy–Schwarz, giving

$$0 \leq \ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y}) \leq (2\|\mathbf{y}\|_2 + 2a\|\hat{\mathbf{y}}\|_2) R_{\mathcal{Z}} + \frac{a\gamma}{2} R_{\mathcal{Z}}^2,$$

where nonnegativity follows from $\ell_{\mathcal{RSP}\mathcal{O}_+} \geq \ell_{\mathcal{RSP}\mathcal{O}} \geq 0$ by Proposition 3(i). Since $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\|\mathbf{y}\|_2] < \infty$, the pointwise surrogate risk $R_+(\hat{\mathbf{y}}; \mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y})]$ is finite for every $\hat{\mathbf{y}}$.

Step 2: differentiation under the expectation. By Proposition 3, $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ is differentiable everywhere with $\nabla_{\hat{\mathbf{y}}} \ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y}) = a(\mathbf{z}_\gamma^*(\mathbf{y}) - \mathbf{z}_\gamma^*(\hat{\mathbf{y}} - \frac{1}{a}\mathbf{y}))$. The gradient is continuous in $\mathbf{y}$ because $\mathbf{z}_\gamma^*(\cdot) = \Pi_{\mathcal{Z}}(-\cdot/\gamma)$ is $1/\gamma$ -Lipschitz by Lemma 1, and uniformly bounded in norm by $aD_{\mathcal{Z}}$. Since the difference quotients of $\ell_{\mathcal{RSP}\mathcal{O}_+}(\cdot, \mathbf{y})$ are bounded by $aD_{\mathcal{Z}}$ uniformly in $\mathbf{y}$, the dominated convergence theorem yields that R_+ is differentiable with

$$\nabla_{\mathbf{y}} R_+(\hat{\mathbf{y}}; \mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\nabla_{\hat{\mathbf{y}}} \ell_{\mathcal{RSP}\mathcal{O}_+}(\hat{\mathbf{y}}, \mathbf{y})] = a \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\mathbf{z}_\gamma^*(\mathbf{y})] - a \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[\mathbf{z}_\gamma^* \left(\hat{\mathbf{y}} - \frac{1}{a} \mathbf{y} \right) \right].$$

Step 3: optimality condition. R_+ is convex (Proposition 3) and differentiable, so $\hat{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^*$ if and only if $\nabla_{\mathbf{y}} R_+(\hat{\mathbf{y}}; \mathbf{x}) = \mathbf{0}$, i.e.

$$\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\mathbf{z}_\gamma^*(\mathbf{y})] = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[\mathbf{z}_\gamma^* \left(\hat{\mathbf{y}} - \frac{1}{a} \mathbf{y} \right) \right].$$

The proof is complete. $\square$

Proof of Corollary 1. All expectations below are finite because $\mathcal{Z}$ is compact and $\mathbb{E}[\|\mathbf{y}\|_2] < \infty$; see Step 1 of the proof of Proposition 5. Consider any vector $\Delta \neq \mathbf{0}$. We compute the difference $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{RSP}\mathcal{O}_+}(2\bar{\mathbf{y}} + \Delta, \mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{RSP}\mathcal{O}_+}(2\bar{\mathbf{y}}, \mathbf{y})]$ as follows:

$$\begin{aligned} & \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{RSP}\mathcal{O}_+}(2\bar{\mathbf{y}} + \Delta, \mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{RSP}\mathcal{O}_+}(2\bar{\mathbf{y}}, \mathbf{y})] \\ &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[\max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - (2\bar{\mathbf{y}} + \Delta)^\top \mathbf{z} - \frac{\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + \{ \mathbf{z}_\gamma^*(\mathbf{y})^\top (2\bar{\mathbf{y}} + \Delta) + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(\mathbf{y})\|_2^2 \} - v^*(\mathbf{y}) \right] \\ & \quad - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[\max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - 2\bar{\mathbf{y}}^\top \mathbf{z} - \frac{\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + \{ 2\mathbf{z}_\gamma^*(\mathbf{y})^\top \bar{\mathbf{y}} + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(\mathbf{y})\|_2^2 \} - v^*(\mathbf{y}) \right] \\ &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[\max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - (2\bar{\mathbf{y}} + \Delta)^\top \mathbf{z} - \frac{\gamma}{2} \|\mathbf{z}\|_2^2 \right\} - \max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{y}^\top \mathbf{z} - 2\bar{\mathbf{y}}^\top \mathbf{z} - \frac{\gamma}{2} \|\mathbf{z}\|_2^2 \right\} + \mathbf{z}_\gamma^*(\mathbf{y})^\top \Delta \right] \\ &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[-v_\gamma^*(2\bar{\mathbf{y}} + \Delta - \mathbf{y}) - (\mathbf{y} - 2\bar{\mathbf{y}})^\top \mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y})\|_2^2 + \mathbf{z}_\gamma^*(\mathbf{y})^\top \Delta \right] \\ &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[-v_\gamma^*(2\bar{\mathbf{y}} + \Delta - \mathbf{y}) - (\mathbf{y} - 2\bar{\mathbf{y}} - \Delta)^\top \mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y})\|_2^2 \right] \\ &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[-v_\gamma^*(2\bar{\mathbf{y}} + \Delta - \mathbf{y}) + (2\bar{\mathbf{y}} + \Delta - \mathbf{y})^\top \mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y})\|_2^2 \right], \end{aligned}$$

where the fourth equality follows from the fact that the conditional distribution of $\mathbf{y}$ is symmetric about its conditional mean, which implies that $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y})] = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\mathbf{z}_\gamma^*(\mathbf{y})]$.

Since $v_\gamma^*(2\bar{\mathbf{y}} + \Delta - \mathbf{y})$ is the minimum value for the optimization problem $\min_{\mathbf{z} \in \mathcal{Z}} (2\bar{\mathbf{y}} + \Delta - \mathbf{y})^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2$, we have

$$-v_\gamma^*(2\bar{\mathbf{y}} + \Delta - \mathbf{y}) + (2\bar{\mathbf{y}} + \Delta - \mathbf{y})^\top \mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y})\|_2^2 \geq 0.$$

This proves that $2\bar{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^*$. Suppose now that $\text{int } \mathcal{Z} \neq \emptyset$. We now show that the expectation of the nonnegative integrand is strictly positive, by exhibiting an explicit open event on which the integrand is strictly positive. Define the open set

$$A := 2\bar{\mathbf{y}} + \gamma \text{int } \mathcal{Z} = \left\{ \mathbf{y} \in \mathbb{R}^d : -\frac{2\bar{\mathbf{y}} - \mathbf{y}}{\gamma} \in \text{int } \mathcal{Z} \right\}.$$

We claim that

$$\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y}) \neq \mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y} + \Delta), \quad \forall \mathbf{y} \in A.$$

Indeed, write $\mathbf{w} := -(2\bar{\mathbf{y}} - \mathbf{y})/\gamma$, so that $\mathbf{w} \in \text{int } \mathcal{Z}$ for $\mathbf{y} \in A$ and, by the projection representation, $\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y}) = \Pi_{\mathcal{Z}}(\mathbf{w}) = \mathbf{w}$, while $\mathbf{z}_\gamma^*(2\bar{\mathbf{y}} - \mathbf{y} + \Delta) = \Pi_{\mathcal{Z}}(\mathbf{w} - \Delta/\gamma)$. Suppose for contradiction that

$\Pi_{\mathcal{Z}}(\mathbf{w} - \Delta/\gamma) = \mathbf{w}$. If $\mathbf{w} - \Delta/\gamma \in \mathcal{Z}$, then $\Pi_{\mathcal{Z}}(\mathbf{w} - \Delta/\gamma) = \mathbf{w} - \Delta/\gamma \neq \mathbf{w}$ since $\Delta \neq \mathbf{0}$, a contradiction. If $\mathbf{w} - \Delta/\gamma \notin \mathcal{Z}$, then its projection lies on the boundary of $\mathcal{Z}$, whereas $\mathbf{w} \in \text{int } \mathcal{Z}$, again a contradiction. This proves the claim.

For $\mathbf{y} \in A$, the point $\mathbf{z}_{\gamma}^*(2\bar{\mathbf{y}} - \mathbf{y})$ is therefore not the unique minimizer (uniqueness holds since $\gamma > 0$) of $\min_{\mathbf{z} \in \mathcal{Z}} (2\bar{\mathbf{y}} + \Delta - \mathbf{y})^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2$, so the integrand is strictly positive on A . Since $\mathcal{Z}$ has nonempty interior, A is a nonempty open set, and since the density of $\mathbf{y}$ is positive on an open set containing $2\bar{\mathbf{y}} + \gamma\mathcal{Z} \supseteq A$, we have $\mathbb{P}_{\mathbf{y}|\mathbf{x}}(\mathbf{y} \in A) > 0$. Therefore,

$$\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[-v_{\gamma}^*(2\bar{\mathbf{y}} + \Delta - \mathbf{y}) + (2\bar{\mathbf{y}} + \Delta - \mathbf{y})^\top \mathbf{z}_{\gamma}^*(2\bar{\mathbf{y}} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{z}_{\gamma}^*(2\bar{\mathbf{y}} - \mathbf{y})\|_2^2 \right] > 0.$$

Thus $2\bar{\mathbf{y}} + \Delta$ is not a minimizer of the $\mathcal{RSP}\mathcal{O}_+$ risk, and $2\bar{\mathbf{y}}$ must be the unique minimizer. $\square$

Proof of Theorem 2. Throughout, write $\bar{\mathbf{z}} := \mathbf{z}^*(\bar{\mathbf{y}})$ and $\mathcal{C} := -\mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}})$, which is a closed convex cone because $\mathcal{Z}$ is closed and convex. We may suppress the dependence of $\mathbf{x}$ for concise exposition. Define the feasibility set

$$\Gamma := \{\gamma \geq 0 : 2\bar{\mathbf{y}} + \gamma\bar{\mathbf{z}} \in \mathcal{C}\},$$

so that, by the general representation (11) of Proposition 4, for every $\gamma > 0$, membership $\gamma \in \Gamma$ is equivalent to $2\bar{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$, and $\bar{\gamma}(\mathbf{x}) = \sup \Gamma$ coincides with (12).

Part (i). Since $\mathcal{C}$ is convex and the mapping $\gamma \mapsto 2\bar{\mathbf{y}} + \gamma\bar{\mathbf{z}}$ is affine, Γ is convex as the preimage of a convex set under an affine mapping, and therefore is an interval. Since $\mathcal{C}$ is closed and the mapping is continuous, Γ is closed. Moreover, by the first-order optimal condition, the optimality of $\bar{\mathbf{z}}$ for the nominal problem $\min_{\mathbf{z} \in \mathcal{Z}} \bar{\mathbf{y}}^\top \mathbf{z}$ means $-\bar{\mathbf{y}} \in \mathcal{N}_{\mathcal{Z}}(\bar{\mathbf{z}})$, i.e., $\bar{\mathbf{y}} \in \mathcal{C}$, and since $\mathcal{C}$ is a cone, $2\bar{\mathbf{y}} \in \mathcal{C}$. Thus, $0 \in \Gamma$. Consequently, $\bar{\gamma} = \sup \Gamma \in [0, +\infty]$ is attained whenever it is finite. If $\bar{\gamma} < \infty$, then $\Gamma = [0, \bar{\gamma}]$; otherwise, we have $\Gamma = [0, +\infty)$. It follows that for $0 < \gamma \leq \bar{\gamma}$, we have $2\bar{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$. Combining this with $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^* = \{2\bar{\mathbf{y}}\}$ from Corollary 1, whose uniqueness part uses $\text{int } \mathcal{Z} \neq \emptyset$, we have $\mathcal{Y}_{\mathcal{RSP}\mathcal{O}_+}^*(\mathbf{x}) \subseteq \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*(\mathbf{x})$ for fixed $\mathbf{x}$. If $0 < \gamma \leq \bar{\gamma}(\mathbf{x})$ for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$, the $\mathcal{RSP}\mathcal{O}_+$ loss is Fisher consistent with the $\mathcal{RSP}\mathcal{O}$ loss by definition. When $\mathcal{Z} = \{\mathbf{z} : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}$ is a bounded polyhedron, Proposition 4 gives $\mathcal{C} = \{\mathbf{A}^\top \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, so $\bar{\gamma}$ is the value of the linear program

$$\sup \left\{ \gamma \geq 0 : 2\bar{\mathbf{y}} = \mathbf{A}^\top \boldsymbol{\lambda} - \gamma \mathbf{z}^*(\bar{\mathbf{y}}), \boldsymbol{\lambda} \in \boldsymbol{\Lambda} \right\}.$$

Part (ii). Suppose that $\mathcal{Z}$ is a bounded polyhedron. Since $\text{int } \mathcal{Z} \neq \emptyset$ by assumption, $\mathcal{Z}$ is full-dimensional and hence has at least two vertices. If $\bar{\mathbf{z}} = \mathbf{0}$, then $2\bar{\mathbf{y}} + \gamma\bar{\mathbf{z}} = 2\bar{\mathbf{y}} \in \mathcal{C}$ for every $\gamma \geq 0$ by Part (i), so $\Gamma = [0, +\infty)$ and $\bar{\gamma} = +\infty$, as stipulated. Assume henceforth that $\bar{\mathbf{z}} \neq \mathbf{0}$.

Step 1: uniqueness places $\bar{\mathbf{y}}$ in the interior of $\mathcal{C}$. Let $V(\mathcal{Z})$ denote the finite vertex set of the bounded polyhedron $\mathcal{Z}$. Since a linear objective attains its minimum over $\mathcal{Z}$ at a vertex, uniqueness of $\bar{\mathbf{z}}$ implies $\bar{\mathbf{z}} \in V(\mathcal{Z})$ and

$$\epsilon_0 := \min_{\mathbf{v} \in V(\mathcal{Z}) \setminus \{\bar{\mathbf{z}}\}} \bar{\mathbf{y}}^\top (\mathbf{v} - \bar{\mathbf{z}}) > 0.$$

Recall that $D_{\mathcal{Z}}$ is the diameter of $\mathcal{Z}$ and set $\epsilon := \epsilon_0/D_{\mathcal{Z}}$. For any $\mathbf{c}$ with $\|\mathbf{c} - \bar{\mathbf{y}}\|_2 < \epsilon$ and any vertex $\mathbf{v} \neq \bar{\mathbf{z}}$, the Cauchy–Schwarz inequality gives

$$\mathbf{c}^\top (\mathbf{v} - \bar{\mathbf{z}}) = \bar{\mathbf{y}}^\top (\mathbf{v} - \bar{\mathbf{z}}) + (\mathbf{c} - \bar{\mathbf{y}})^\top (\mathbf{v} - \bar{\mathbf{z}}) \geq \epsilon_0 - \|\mathbf{c} - \bar{\mathbf{y}}\|_2 D_{\mathcal{Z}} > 0,$$

so $\bar{\mathbf{z}}$ is optimal (indeed uniquely optimal) for the cost vector $\mathbf{c}$, i.e., $\mathbf{c} \in \mathcal{C}$. Hence the open ball $\mathcal{B}_o(\bar{\mathbf{y}}, \epsilon) \subseteq \mathcal{C}$, so $\bar{\mathbf{y}} \in \text{int } \mathcal{C}$ and $\delta := \text{dist}(\bar{\mathbf{y}}, \partial \mathcal{C}) \geq \epsilon > 0$. Moreover, since $\mathcal{C}$ is closed and convex, the closed ball $\mathcal{B}(\bar{\mathbf{y}}, \delta) \subseteq \mathcal{C}$.

Step 2: membership under the translation. Let $0 \leq \gamma \leq 2\delta/\|\bar{\mathbf{z}}\|_2$. Then $\|(\gamma/2)\bar{\mathbf{z}}\|_2 \leq \delta$, so $\bar{\mathbf{y}} + (\gamma/2)\bar{\mathbf{z}} \in \mathcal{B}(\bar{\mathbf{y}}, \delta) \subseteq \mathcal{C}$, and since $\mathcal{C}$ is a cone,

$$2\bar{\mathbf{y}} + \gamma\bar{\mathbf{z}} = 2\left(\bar{\mathbf{y}} + \frac{\gamma}{2}\bar{\mathbf{z}}\right) \in \mathcal{C}.$$

Hence $\gamma \in \Gamma$, and therefore $\bar{\gamma}(\mathbf{x}) \geq 2\delta/\|\bar{\mathbf{z}}\|_2 = 2\text{dist}(\bar{\mathbf{y}}, \partial \mathcal{C})/\|\mathbf{z}^*(\bar{\mathbf{y}})\|_2 > 0$. In particular the regime $(0, \bar{\gamma}(\mathbf{x})]$ in Part (i) is nonempty. $\square$

Proof of Example 1. The nominal problem at $\bar{\mathbf{y}} = (1, 3/2)^\top$ has the unique solution $\mathbf{z}^*(\bar{\mathbf{y}}) = (-1, -2)^\top$. At this solution, the active inequalities in the representation $\mathbf{A}\mathbf{z} \geq \mathbf{b}$ are $-z_1 + z_2 \geq -1$ and $z_1 \geq -1$, whose normal vectors are $(-1, 1)^\top$ and $(1, 0)^\top$, respectively. Hence, by Proposition 4, the condition $2\bar{\mathbf{y}} \in \mathcal{Y}_{\mathcal{RSP}\mathcal{O}}^*$ is equivalent to the existence of $\lambda_1, \lambda_2 \geq 0$ such that

$$(2, 3)^\top = \lambda_1(-1, 1)^\top + \lambda_2(1, 0)^\top - \gamma(-1, -2)^\top.$$

Equating coordinates gives

$$\lambda_1 = 3 - 2\gamma, \quad \lambda_2 = 5 - 3\gamma.$$

Both multipliers are nonnegative if and only if $0 \leq \gamma \leq 3/2$, so the threshold in (12) is $\bar{\gamma} = 3/2$. In particular, $\gamma = 1$ lies in the Fisher-consistent regime, whereas $\gamma = 2$ does not, as illustrated in Figure 3. $\square$

Proof of Corollary 2. Fix $\mathbf{x}$ and write $\bar{\mathbf{z}} := \mathbf{z}^*(\bar{\mathbf{y}})$, $\mathbf{Z}^*(\mathbf{c}) := \arg \min_{\mathbf{z} \in \mathcal{Z}} \mathbf{c}^\top \mathbf{z}$, and let $V(\mathcal{Z})$ denote the finite vertex set of the bounded polyhedron $\mathcal{Z}$. All expectations below are finite because $\mathcal{Z}$ is compact and $\mathbb{E}[\|\mathbf{y}\|_2] < \infty$. Since $\mathbf{y}$ admits a density and the set of cost vectors $\mathbf{c}$ for which $\mathbf{Z}^*(\mathbf{c})$ is not a singleton is contained in the union of the boundaries of the finitely many cones

$\{-\mathcal{N}_{\mathcal{Z}}(\mathbf{v}) : \mathbf{v} \in V(\mathcal{Z})\}$, a Lebesgue-null set, the nominal decision map $\mathbf{z}^*(\mathbf{y})$ is single-valued almost surely; the same applies to $\mathbf{z}^*(2\bar{\mathbf{y}} - \mathbf{y})$, because $2\bar{\mathbf{y}} - \mathbf{y}$ also admits a density. If $\mathcal{Z}$ is a singleton the statement is trivial, so assume $|V(\mathcal{Z})| \geq 2$.

Step 1: the oracle predictor $2\bar{\mathbf{y}}$ is a target-risk minimizer. For any $\hat{\mathbf{y}}$ and any fixed $\mathbf{z}_0 \in \mathbf{Z}^*(\hat{\mathbf{y}})$,

$$\begin{aligned} \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}}[\ell_{\mathcal{SPO}}(\hat{\mathbf{y}}, \mathbf{y})] &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} \left[\max_{\mathbf{z} \in \mathbf{Z}^*(\hat{\mathbf{y}})} \mathbf{y}^\top \mathbf{z} \right] - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{y})] \\ &\geq \bar{\mathbf{y}}^\top \mathbf{z}_0 - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{y})] \geq v^*(\bar{\mathbf{y}}) - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{y})]. \end{aligned}$$

Since the nominal argmin is invariant under positive scaling of the cost vector, uniqueness of $\bar{\mathbf{z}}$ for $\bar{\mathbf{y}}$ gives $\mathbf{Z}^*(2\bar{\mathbf{y}}) = \{\bar{\mathbf{z}}\}$, so the pointwise $\mathcal{SPO}$ risk at $2\bar{\mathbf{y}}$ equals $\bar{\mathbf{y}}^\top \bar{\mathbf{z}} - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{y})] = v^*(\bar{\mathbf{y}}) - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{y})]$, which attains the lower bound. Hence $2\bar{\mathbf{y}} \in \mathcal{Y}_{\mathcal{SPO}}^*$.

Step 2: the oracle predictor minimizes the surrogate risk. With $a = 1$, using $\max_{\mathbf{z} \in \mathcal{Z}} (\mathbf{y} - \hat{\mathbf{y}})^\top \mathbf{z} = -v^*(\hat{\mathbf{y}} - \mathbf{y})$, the $\mathcal{SPO}_+$ loss reads $\ell_{\mathcal{SPO}_+}(\hat{\mathbf{y}}, \mathbf{y}) = -v^*(\hat{\mathbf{y}} - \mathbf{y}) + \hat{\mathbf{y}}^\top \mathbf{z}^*(\mathbf{y}) - v^*(\mathbf{y})$ almost surely. Therefore, for any $\Delta \in \mathbb{R}^d$,

$$\begin{aligned} &\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{SPO}_+}(2\bar{\mathbf{y}} + \Delta, \mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\ell_{\mathcal{SPO}_+}(2\bar{\mathbf{y}}, \mathbf{y})] \\ &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [-v^*(2\bar{\mathbf{y}} + \Delta - \mathbf{y}) + v^*(2\bar{\mathbf{y}} - \mathbf{y}) + \Delta^\top \mathbf{z}^*(\mathbf{y})]. \end{aligned}$$

By central symmetry, $\mathbf{w} := 2\bar{\mathbf{y}} - \mathbf{y}$ has the same distribution as $\mathbf{y}$, so $\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\Delta^\top \mathbf{z}^*(\mathbf{y})] = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [\Delta^\top \mathbf{z}^*(\mathbf{w})]$, and the difference becomes

$$\mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}|\mathbf{x}}} [v^*(\mathbf{w}) + \Delta^\top \mathbf{z}^*(\mathbf{w}) - v^*(\mathbf{w} + \Delta)]. \quad (\text{EC.1})$$

Since $v^*(\mathbf{w} + \Delta) \leq (\mathbf{w} + \Delta)^\top \mathbf{z}^*(\mathbf{w}) = v^*(\mathbf{w}) + \Delta^\top \mathbf{z}^*(\mathbf{w})$, the integrand is nonnegative pointwise, so the difference (EC.1) is nonnegative for every Δ , i.e., $2\bar{\mathbf{y}}$ minimizes the pointwise $\mathcal{SPO}_+$ risk with $a = 1$.

Step 3: uniqueness of the surrogate minimizer. Let $\Delta \neq \mathbf{0}$. Since $\text{int } \mathcal{Z} \neq \emptyset$, the linear function $\mathbf{z} \mapsto \Delta^\top \mathbf{z}$ is nonconstant on $\mathcal{Z}$, so there exist vertices $\mathbf{z}_1 \in \arg \max_{\mathbf{z} \in V(\mathcal{Z})} \Delta^\top \mathbf{z}$ and $\mathbf{z}_2 \in V(\mathcal{Z})$ with $\Delta^\top \mathbf{z}_2 < \Delta^\top \mathbf{z}_1$. Because $\mathcal{Z}$ is full-dimensional, the cone $\mathcal{C}_1 := \text{int}(-\mathcal{N}_{\mathcal{Z}}(\mathbf{z}_1))$ is a nonempty open cone with apex $\mathbf{0}$, on which $\mathbf{Z}^*(\mathbf{c}) = \{\mathbf{z}_1\}$ (by the same vertex-perturbation argument as in Step 1 of the proof of Theorem 2(ii)). Fix $\mathbf{c}_0 \in \mathcal{C}_1$ and $\varepsilon > 0$ with $\mathcal{B}(\mathbf{c}_0, \varepsilon) \subseteq \mathcal{C}_1$. Since $\mathcal{C}_1$ is a cone, $\mathcal{B}(t\mathbf{c}_0, t\varepsilon) \subseteq \mathcal{C}_1$ for every $t > 0$. For $\mathbf{w} \in \mathcal{B}(t\mathbf{c}_0, t\varepsilon)$ we have $\mathbf{z}^*(\mathbf{w}) = \mathbf{z}_1$ and $v^*(\mathbf{w}) = \mathbf{w}^\top \mathbf{z}_1$, so, bounding $v^*(\mathbf{w} + \Delta) \leq (\mathbf{w} + \Delta)^\top \mathbf{z}_2$,

$$v^*(\mathbf{w}) + \Delta^\top \mathbf{z}^*(\mathbf{w}) - v^*(\mathbf{w} + \Delta) \geq \mathbf{w}^\top (\mathbf{z}_1 - \mathbf{z}_2) + \Delta^\top (\mathbf{z}_1 - \mathbf{z}_2) \geq \Delta^\top (\mathbf{z}_1 - \mathbf{z}_2) - t(\|\mathbf{c}_0\|_2 + \varepsilon) D_{\mathcal{Z}},$$

where $D_{\mathcal{Z}}$ is the diameter of $\mathcal{Z}$. Since $\Delta^\top (\mathbf{z}_1 - \mathbf{z}_2) > 0$ is fixed, choosing $t > 0$ small enough makes $v^*(\mathbf{w}) + \Delta^\top \mathbf{z}^*(\mathbf{w}) - v^*(\mathbf{w} + \Delta) \geq \Delta^\top (\mathbf{z}_1 - \mathbf{z}_2)/2 > 0$ on $\mathcal{B}(t\mathbf{c}_0, t\varepsilon)$. Finally, $\mathbf{y}$ has a density positive on an open set $O \ni 2\bar{\mathbf{y}}$, so $\mathbf{w} = 2\bar{\mathbf{y}} - \mathbf{y}$ has a density positive on the open set $2\bar{\mathbf{y}} - O \ni \mathbf{0}$. For t

small enough, $\mathcal{B}(t\mathbf{c}_0, t\varepsilon) \subseteq 2\bar{\mathbf{y}} - O$, with $\mathbb{P}_{\mathbf{y}|\mathbf{x}}(\mathbf{w} \in \mathcal{B}(t\mathbf{c}_0, t\varepsilon)) > 0$ and the difference (EC.1) is strictly positive. Therefore, $2\bar{\mathbf{y}} + \Delta$ is not a surrogate minimizer, and hence $\mathcal{Y}_{\mathcal{SP}\mathcal{O}_+}^* = \{2\bar{\mathbf{y}}\}$ for $a = 1$.

Step 4: conclusion. By Steps 1-3, for every $\mathbf{x}$ the unique minimizer of the pointwise $\mathcal{SP}\mathcal{O}_+$ risk with $a = 1$ belongs to $\mathcal{Y}_{\mathcal{SP}\mathcal{O}}^*$. Integrating over $\mathbb{P}_{\mathbf{x}}$ via the interchangeability principle, as in Section 4.2, yields Fisher consistency of the $\mathcal{SP}\mathcal{O}_+$ loss with $a = 1$ with respect to the $\mathcal{SP}\mathcal{O}$ loss. $\square$

Proof of Proposition 6. Let $S = S_{\mathbf{y}_0}$ for notational simplicity and write $\mathbf{z}^0 = \mathbf{z}_\gamma^*(\mathbf{y}_0)$. Since the objective is strongly convex and the constraints are affine, the KKT conditions are necessary and sufficient: $\mathbf{z} \in \mathcal{Z}$ solves the problem at $\mathbf{y}$ if and only if there exists $\boldsymbol{\lambda} \geq \mathbf{0}$ such that

$$\mathbf{y} + \gamma\mathbf{z} - \mathbf{A}^\top \boldsymbol{\lambda} = \mathbf{0},$$

$$\mathbf{A}\mathbf{z} \geq \mathbf{b},$$

$$\lambda_i(\mathbf{a}_i^\top \mathbf{z} - b_i) = 0, \quad \forall i.$$

At $\mathbf{y}_0$, LICQ by assumption (i) guarantees that the multiplier $\boldsymbol{\lambda}^0$ is unique, and strict complementarity by assumption (ii) gives $\lambda_i^0 > 0$ for all $i \in S$ and, by definition of the active set, $\mathbf{a}_i^\top \mathbf{z}^0 > b_i$ for all $i \notin S$.

We now construct a candidate primal–dual pair for $\mathbf{y}$ near $\mathbf{y}_0$ and verify that it satisfies the KKT conditions with active set S . Guided by the stationarity condition restricted to S (i.e., $\boldsymbol{\lambda}_{S^c} = \mathbf{0}$ and $\mathbf{A}_S \mathbf{z} = \mathbf{b}_S$), consider the linear system

$$\gamma\mathbf{z} - \mathbf{A}_S^\top \boldsymbol{\lambda}_S = -\mathbf{y}, \quad \mathbf{A}_S \mathbf{z} = \mathbf{b}_S.$$

From the first equation, $\mathbf{z} = \frac{1}{\gamma}(\mathbf{A}_S^\top \boldsymbol{\lambda}_S - \mathbf{y})$. Substituting into $\mathbf{A}_S \mathbf{z} = \mathbf{b}_S$ and using the invertibility of $\mathbf{A}_S \mathbf{A}_S^\top$ (which follows from LICQ) yields the unique solution

$$\boldsymbol{\lambda}_S(\mathbf{y}) = (\mathbf{A}_S \mathbf{A}_S^\top)^{-1}(\gamma \mathbf{b}_S + \mathbf{A}_S \mathbf{y}), \quad \mathbf{z}(\mathbf{y}) = \frac{1}{\gamma} \left(\mathbf{A}_S^\top (\mathbf{A}_S \mathbf{A}_S^\top)^{-1} \mathbf{A}_S - \mathbf{I} \right) \mathbf{y} + \mathbf{A}_S^\top (\mathbf{A}_S \mathbf{A}_S^\top)^{-1} \mathbf{b}_S.$$

Both mappings are affine, hence continuous, in $\mathbf{y}$, and by construction $\mathbf{z}(\mathbf{y}_0) = \mathbf{z}^0$ and $\boldsymbol{\lambda}_S(\mathbf{y}_0) = \boldsymbol{\lambda}_S^0$ (uniqueness of the multiplier under LICQ). By continuity and strict complementarity, there exists a neighborhood $\mathcal{N}$ of $\mathbf{y}_0$ such that, for every $\mathbf{y} \in \mathcal{N}$,

$$\boldsymbol{\lambda}_S(\mathbf{y}) > \mathbf{0} \quad \text{and} \quad \mathbf{a}_i^\top \mathbf{z}(\mathbf{y}) > b_i, \quad \forall i \notin S.$$

Consequently, for every $\mathbf{y} \in \mathcal{N}$, the pair $(\mathbf{z}(\mathbf{y}), (\boldsymbol{\lambda}_S(\mathbf{y}), \mathbf{0}_{S^c}))$ satisfies stationarity, primal feasibility ($\mathbf{A}_S \mathbf{z}(\mathbf{y}) = \mathbf{b}_S$ and the inactive constraints hold strictly), dual feasibility, and complementary slackness. By sufficiency of the KKT conditions, $\mathbf{z}_\gamma^*(\mathbf{y}) = \mathbf{z}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{N}$, and the active set of $\mathbf{z}_\gamma^*(\mathbf{y})$ is exactly S on $\mathcal{N}$.

Thus, $\mathbf{z}_\gamma^*(\cdot)$ is affine on $\mathcal{N}$. In particular, it is differentiable at $\mathbf{y}_0$ with Jacobian

$$\nabla \mathbf{z}_\gamma^*(\mathbf{y}_0) = \frac{1}{\gamma} (\mathbf{A}_{S_{\mathbf{y}_0}}^\top (\mathbf{A}_{S_{\mathbf{y}_0}} \mathbf{A}_{S_{\mathbf{y}_0}}^\top)^{-1} \mathbf{A}_{S_{\mathbf{y}_0}} - \mathbf{I}).$$

This completes the proof. $\square$

B.2 Proofs of Section 5

Proof of Proposition 7. As noted in Subsection 5.1, since $\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \mathcal{Z}$, the definition of b implies

$$0 \leq \ell_{\mathcal{RSP}\mathcal{O}}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbf{y}^\top \mathbf{z}_\gamma^*(\hat{\mathbf{y}}) - v^*(\mathbf{y}) \leq b, \quad \forall \hat{\mathbf{y}} \in \mathbb{R}^d, \mathbf{y} \in \mathcal{Y}.$$

Let $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [N]}$ denote the sample and define

$$\Phi(S) := \sup_{\mathbf{g} \in \mathcal{G}} (R(\mathbf{g}) - \hat{R}(\mathbf{g})).$$

If one sample point $(\mathbf{x}_i, \mathbf{y}_i)$ is replaced by an independent copy $(\mathbf{x}'_i, \mathbf{y}'_i)$, then for every $\mathbf{g} \in \mathcal{G}$ the empirical risk changes by at most b/N , and hence Φ changes by at most b/N as well. By McDiarmid's bounded difference inequality, with probability at least $1 - \delta/2$,

$$\Phi(S) \leq \mathbb{E}[\Phi(S)] + b\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Next, by the standard symmetrization argument (Bartlett and Mendelson 2002), letting $S' = \{(\mathbf{x}'_i, \mathbf{y}'_i)\}_{i \in [N]}$ be an independent ghost sample and $\sigma_1, \dots, \sigma_N$ be independent Rademacher random variables,

$$\begin{aligned} \mathbb{E}[\Phi(S)] &= \mathbb{E}_S \left[\sup_{\mathbf{g} \in \mathcal{G}} \mathbb{E}_{S'} \left[\frac{1}{N} \sum_{i \in [N]} (\ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{g}(\mathbf{x}'_i), \mathbf{y}'_i) - \ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{g}(\mathbf{x}_i), \mathbf{y}_i)) \right] \right] \\ &\leq \mathbb{E}_{S, S', \sigma} \left[\sup_{\mathbf{g} \in \mathcal{G}} \frac{1}{N} \sum_{i \in [N]} \sigma_i (\ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{g}(\mathbf{x}'_i), \mathbf{y}'_i) - \ell_{\mathcal{RSP}\mathcal{O}}(\mathbf{g}(\mathbf{x}_i), \mathbf{y}_i)) \right] \\ &\leq 2\mathfrak{R}_1(\mathcal{H}_{\ell_{\mathcal{RSP}\mathcal{O}}}). \end{aligned}$$

The same two steps applied to $\Phi'(S) := \sup_{\mathbf{g} \in \mathcal{G}} (\hat{R}(\mathbf{g}) - R(\mathbf{g}))$ give the identical bound with probability at least $1 - \delta/2$. A union bound over the two events yields

$$\sup_{\mathbf{g} \in \mathcal{G}} |R(\mathbf{g}) - \hat{R}(\mathbf{g})| \leq 2\mathfrak{R}_1(\mathcal{H}_{\ell_{\mathcal{RSP}\mathcal{O}}}) + b\sqrt{\frac{\log(2/\delta)}{2N}}$$

with probability at least $1 - \delta$. □

Proof of Theorem 3. Let $\mathcal{E}$ denote the event

$$\sup_{\mathbf{g} \in \mathcal{G}} |R(\mathbf{g}) - \hat{R}(\mathbf{g})| \leq 2\mathfrak{R}_1(\mathcal{H}_{\ell_{\mathcal{RSP}\mathcal{O}}}) + b\sqrt{\frac{\log(2/\delta)}{2N}},$$

so that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ by Proposition 7. Fix an arbitrary $\epsilon > 0$ and choose $\mathbf{g}_\epsilon \in \mathcal{G}$ such that $R(\mathbf{g}_\epsilon) \leq R^*(\mathcal{G}) + \epsilon$ by the definition of the infimum. On the event $\mathcal{E}$, we have

$$\begin{aligned} R(\hat{\mathbf{g}}_{\text{RSPO}}) - R^*(\mathcal{G}) &= [R(\hat{\mathbf{g}}_{\text{RSPO}}) - \hat{R}(\hat{\mathbf{g}}_{\text{RSPO}})] + [\hat{R}(\hat{\mathbf{g}}_{\text{RSPO}}) - \hat{R}(\mathbf{g}_\epsilon)] \\ &\quad + [\hat{R}(\mathbf{g}_\epsilon) - R(\mathbf{g}_\epsilon)] + [R(\mathbf{g}_\epsilon) - R^*(\mathcal{G})] \\ &\leq 2 \sup_{\mathbf{g} \in \mathcal{G}} |R(\mathbf{g}) - \hat{R}(\mathbf{g})| + 0 + \epsilon \\ &\leq 4\mathfrak{R}_1(\mathcal{H}_{\ell_{\mathcal{RSP}\mathcal{O}}}) + b\sqrt{\frac{2\log(2/\delta)}{N}} + \epsilon, \end{aligned}$$

where the second inequality uses $\widehat{R}(\widehat{\mathbf{g}}_{\text{RSPO}}) \leq \widehat{R}(\mathbf{g}_\epsilon)$, which holds because $\widehat{\mathbf{g}}_{\text{RSPO}}$ minimizes the empirical risk, and the last inequality uses the inequality defining event $\mathcal{E}$. Since the event $\mathcal{E}$ does not depend on ϵ , letting $\epsilon \downarrow 0$ gives, on $\mathcal{E}$,

$$R(\widehat{\mathbf{g}}_{\text{RSPO}}) - R^*(\mathcal{G}) \leq 4\mathfrak{R}_1(\mathcal{H}_{\ell_{\text{RSPO}}}) + b\sqrt{\frac{2\log(2/\delta)}{N}}.$$

Finally, applying Lemma 2 to bound $4\mathfrak{R}_1(\mathcal{H}_{\ell_{\text{RSPO}}}) \leq \frac{4\sqrt{2}r}{\gamma}\mathfrak{R}_d(\mathcal{G})$ completes the proof. $\square$

Proof of Proposition 8. Following a similar argument to that in Proposition 7, for any $\delta > 0$, with probability at least $1 - \delta$, for all $\mathbf{g} \in \mathcal{G}$ it holds that

$$|R_+(\mathbf{g}) - \widehat{R}_+(\mathbf{g})| \leq 2\mathfrak{R}_1(\mathcal{H}_{\ell_{\text{RSPO}_+}}) + \mathfrak{L}(\gamma)\sqrt{\frac{\log(2/\delta)}{2N}}.$$

This result applies because the loss takes values in $[0, \mathfrak{L}(\gamma)]$ on $\mathcal{G} \times \mathcal{Y}$. Indeed, the upper bound holds by the definition of $\mathfrak{L}(\gamma)$, and nonnegativity follows from $\ell_{\text{RSPO}_+} \geq \ell_{\text{RSPO}} \geq 0$ (Proposition 3(i)).

Moreover, by Proposition 3(iv), $\ell_{\text{RSPO}_+}(\cdot, \mathbf{y})$ is $aD_{\mathcal{Z}}$ -Lipschitz continuous for every $\mathbf{y} \in \mathcal{Y}$. By the vector contraction inequality (Maurer 2016, Corollary 4), $\mathfrak{R}_1(\mathcal{H}_{\ell_{\text{RSPO}_+}}) \leq \sqrt{2}aD_{\mathcal{Z}}\mathfrak{R}_d(\mathcal{G})$. Thus,

$$|R_+(\mathbf{g}) - \widehat{R}_+(\mathbf{g})| \leq 2\sqrt{2}aD_{\mathcal{Z}}\mathfrak{R}_d(\mathcal{G}) + \mathfrak{L}(\gamma)\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Proceeding as in Theorem 3 and using $\mathfrak{R}_d(\mathcal{G}) \leq C_0/\sqrt{N}$, we obtain that, for all $\delta \in (0, 1)$, with probability $1 - \delta$,

$$R_+(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R_+(\mathcal{G}) \leq \frac{4\sqrt{2}aD_{\mathcal{Z}}C_0}{\sqrt{N}} + \mathfrak{L}(\gamma)\sqrt{\frac{2\log(2/\delta)}{N}}.$$

The proof is complete. $\square$

Proof of Theorem 4. We complete this proof in six steps. *Step 1* bounds the surrogate excess risk of $\widehat{\mathbf{g}}_{\text{RSPO}_+}$ relative to the oracle predictor $\mathbf{g}_0$ by $B_N + A_{\mathcal{G}}^+$. *Step 2* converts the local strong convexity of Condition 1(ii) into a growth bound valid on all of $\mathbb{R}^d$. *Step 3* integrates that bound to control $\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2$. *Step 4* transfers this control to the target risk through the Lipschitz continuity of the decision map. *Step 5* identifies the oracle predictor with the target optimum, which turns the resulting estimate into a bound on the target excess risk. *Step 6* reads off the rate. To simplify the exposition, we define

$$\Delta_\gamma(\mathbf{c}; \mathbf{x}) := R(\mathbf{c}; \mathbf{x}) - \inf_{\mathbf{c}' \in \mathbb{R}^d} R(\mathbf{c}'; \mathbf{x}), \quad \Delta_\gamma^+(\mathbf{c}; \mathbf{x}) := R_+(\mathbf{c}; \mathbf{x}) - \inf_{\mathbf{c}' \in \mathbb{R}^d} R_+(\mathbf{c}'; \mathbf{x}).$$

Step 1. Condition 1(i) and the tower property imply $R_+(\mathbf{g}_0) \leq R_+(\mathbf{g})$ for every measurable predictor $\mathbf{g}$ with finite surrogate risk. Hence

$$R_+(\mathbf{g}_0) = R_+^*(\mathcal{G}_{\text{all}}). \quad (\text{EC.2})$$

Proposition 8 implies that, with probability at least $1 - \delta$,

$$R_+(\hat{\mathbf{g}}_{\text{RSPO}_+}) - R_+^*(\mathcal{G}) \leq \frac{4\sqrt{2}D_{\mathcal{Z}}C_0}{\sqrt{N}} + \mathfrak{L}(\gamma)\sqrt{\frac{2\log(2/\delta)}{N}}.$$

Therefore, on this event,

$$\begin{aligned} R_+(\hat{\mathbf{g}}_{\text{RSPO}_+}) - R_+(\mathbf{g}_0) &= [R_+(\hat{\mathbf{g}}_{\text{RSPO}_+}) - R_+^*(\mathcal{G})] + [R_+^*(\mathcal{G}) - R_+^*(\mathcal{G}_{\text{all}})] \\ &\leq \frac{4\sqrt{2}D_{\mathcal{Z}}C_0}{\sqrt{N}} + \mathfrak{L}(\gamma)\sqrt{\frac{2\log(2/\delta)}{N}} + [R_+^*(\mathcal{G}) - R_+^*(\mathcal{G}_{\text{all}})] \\ &= B_N + A_{\mathcal{G}}^+. \end{aligned} \quad (\text{EC.3})$$

In particular, $B_N + A_{\mathcal{G}}^+ \geq 0$, so the square root appearing in (20) is well-defined.

Step 2. We claim that, for $\mathbb{P}_{\mathbf{x}}$ -almost every $\mathbf{x}$ and every $\mathbf{c} \in \mathbb{R}^d$,

$$\Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) \geq \frac{\mu}{2} \min \{ \|\mathbf{c} - \mathbf{g}_0(\mathbf{x})\|_2^2, \rho \|\mathbf{c} - \mathbf{g}_0(\mathbf{x})\|_2 \}. \quad (\text{EC.4})$$

The exceptional null set is the union of those in Conditions 1(i) and (ii) and does not depend on $\mathbf{c}$, so (EC.4) may be applied at a sample-dependent argument. Fix such an $\mathbf{x}$ and write $t := \|\mathbf{c} - \mathbf{g}_0(\mathbf{x})\|_2$. Since $\mathbf{g}_0(\mathbf{x})$ is a global minimizer of the convex function $R_+(\cdot; \mathbf{x})$, we have $\mathbf{0} \in \partial R_+(\mathbf{g}_0(\mathbf{x}); \mathbf{x})$. Under Condition 1(i) the pointwise surrogate excess risk takes the form

$$\Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) = R_+(\mathbf{c}; \mathbf{x}) - R_+(\mathbf{g}_0(\mathbf{x}); \mathbf{x}), \quad \mathbf{c} \in \mathbb{R}^d.$$

If $t \leq \rho$, strong convexity on the closed ball $\mathcal{B}(\mathbf{g}_0(\mathbf{x}), \rho)$ gives

$$\Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) \geq \frac{\mu}{2} t^2.$$

If $t > \rho$, set $\mathbf{c}_{\rho} := \mathbf{g}_0(\mathbf{x}) + \frac{\rho}{t}(\mathbf{c} - \mathbf{g}_0(\mathbf{x}))$, so that $\|\mathbf{c}_{\rho} - \mathbf{g}_0(\mathbf{x})\|_2 = \rho$. Convexity and $\mathbf{c}_{\rho} = (1 - \rho/t)\mathbf{g}_0(\mathbf{x}) + (\rho/t)\mathbf{c}$ give

$$\Delta_{\gamma}^+(\mathbf{c}_{\rho}; \mathbf{x}) \leq (1 - \frac{\rho}{t})\Delta_{\gamma}^+(\mathbf{g}_0(\mathbf{x}); \mathbf{x}) + \frac{\rho}{t}\Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) = \frac{\rho}{t}\Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}).$$

The first case applies to $\mathbf{c}_{\rho}$, so

$$\Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) \geq \frac{t}{\rho} \Delta_{\gamma}^+(\mathbf{c}_{\rho}; \mathbf{x}) \geq \frac{t}{\rho} \cdot \frac{\mu}{2} \rho^2 = \frac{\mu\rho}{2} t.$$

Since $\min\{t^2, \rho t\}$ equals t^2 when $t \leq \rho$ and ρt otherwise, the two cases prove (EC.4).

Step 3. Taking expectations in (EC.4) at $\mathbf{c} = \widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x})$, and using $\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}}[\Delta_{\gamma}^+(\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}); \mathbf{x})] = R_+(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R_+(\mathbf{g}_0)$ together with (EC.3), yields

$$\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} \left[\min \left\{ \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2^2, \rho \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2 \right\} \right] \leq \frac{2(B_N + A_{\mathcal{G}}^+)}{\mu}. \quad (\text{EC.5})$$

For every $t \geq 0$,

$$t \leq \sqrt{\min\{t^2, \rho t\}} + \frac{1}{\rho} \min\{t^2, \rho t\}. \quad (\text{EC.6})$$

Indeed, if $t \leq \rho$ the first term on the right equals t , and if $t > \rho$ the second term equals t ; in either case the remaining term is nonnegative. Applying (EC.6) with $t = \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2$, taking expectations, and bounding the first resulting term by Jensen's inequality (the square root is concave) gives

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} \left[\|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2 \right] \\ & \leq \left(\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} \left[\min \left\{ \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2^2, \rho \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2 \right\} \right] \right)^{1/2} \\ & \quad + \frac{1}{\rho} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} \left[\min \left\{ \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2^2, \rho \|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2 \right\} \right] \\ & \leq \sqrt{\frac{2(B_N + A_{\mathcal{G}}^+)}{\mu}} + \frac{2(B_N + A_{\mathcal{G}}^+)}{\mu\rho}. \end{aligned} \quad (\text{EC.7})$$

Step 4. By the definition of ℓ_{RSPO} and the tower property, every measurable $\mathbf{g}$ satisfies

$$R(\mathbf{g}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [\bar{\mathbf{y}}(\mathbf{x})^\top \mathbf{z}_{\gamma}^*(\mathbf{g}(\mathbf{x}))] - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_{\mathbf{y}}} [v^*(\mathbf{y})], \quad (\text{EC.8})$$

all terms being finite by (17). Subtracting (EC.8) at $\mathbf{g}$ and at $\mathbf{g}_0$ gives

$$R(\mathbf{g}) - R(\mathbf{g}_0) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [\bar{\mathbf{y}}(\mathbf{x})^\top (\mathbf{z}_{\gamma}^*(\mathbf{g}(\mathbf{x})) - \mathbf{z}_{\gamma}^*(\mathbf{g}_0(\mathbf{x})))]. \quad (\text{EC.9})$$

By Lemma 1, $\mathbf{z}_{\gamma}^*(\cdot)$ is $1/\gamma$ -Lipschitz on $\mathbb{R}^d$. Hence Condition 1(iv), Cauchy-Schwarz, and (EC.7) imply

$$\begin{aligned} |R(\widehat{\mathbf{g}}_{\text{RSPO}_+}) - R(\mathbf{g}_0)| & \leq \frac{\bar{\beta}}{2\gamma} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [\|\widehat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}) - \mathbf{g}_0(\mathbf{x})\|_2] \\ & \leq \frac{\bar{\beta}}{2\gamma} \left[\sqrt{\frac{2(B_N + A_{\mathcal{G}}^+)}{\mu}} + \frac{2(B_N + A_{\mathcal{G}}^+)}{\mu\rho} \right]. \end{aligned} \quad (\text{EC.10})$$

Step 5. By Condition 1(iii) and Lemma EC.2, $\Delta_{\gamma}(\mathbf{g}_0(\mathbf{x}); \mathbf{x}) = 0$ for almost every $\mathbf{x}$, that is, $\mathbf{g}_0(\mathbf{x})$ minimizes $R(\cdot; \mathbf{x})$ over $\mathbb{R}^d$. Integrating the pointwise inequality $R(\mathbf{g}_0(\mathbf{x}); \mathbf{x}) \leq R(\mathbf{g}(\mathbf{x}); \mathbf{x})$ and using the tower property gives $R(\mathbf{g}_0) \leq R(\mathbf{g})$ for every measurable $\mathbf{g}$. Since $R^* = R^*(\mathcal{G}_{\text{all}})$, this yields

$$R(\mathbf{g}_0) = R^*. \quad (\text{EC.11})$$

Combining (EC.11) with (EC.10) proves (20).

Step 6. Since $\mathbf{g}_0 \in \mathcal{G}$, we have $R_+^*(\mathcal{G}) = R_+^*(\mathcal{G}_{\text{all}})$ by Condition 1(i). The stated constants do not depend on N . Fix $\eta \in (0, 1)$ and take $\delta = \eta$ in (20). By (EC.3),

$$\bar{c}(\eta) := \sqrt{N} (B_N + A_{\mathcal{G}}^+) = 4\sqrt{2} D_{\mathcal{Z}} C_0 + \mathfrak{L}(\gamma) \sqrt{2 \log(2/\eta)} < \infty, \quad (\text{EC.12})$$

which does not depend on N . Hence, on the event of probability at least $1 - \eta$ on which (20) holds, and for every $N \geq 1$,

$$N^{1/4} [R(\widehat{\mathbf{g}}_{\text{RSPO}+}) - R^*] \leq \frac{\bar{\beta}}{2\gamma} \left[\sqrt{\frac{2\bar{c}(\eta)}{\mu}} + \frac{2\bar{c}(\eta)}{\mu\rho} N^{-1/4} \right] \leq \frac{\bar{\beta}}{2\gamma} \left[\sqrt{\frac{2\bar{c}(\eta)}{\mu}} + \frac{2\bar{c}(\eta)}{\mu\rho} \right] =: M(\eta).$$

Since $M(\eta)$ depends only on η and the structural constants,

$$\mathbb{P} \left(N^{1/4} [R(\widehat{\mathbf{g}}_{\text{RSPO}+}) - R^*] > M(\eta) \right) \leq \eta, \quad N \geq 1,$$

which is precisely the statement that $R(\widehat{\mathbf{g}}_{\text{RSPO}+}) - R^* = \mathcal{O}_p(N^{-1/4})$. $\square$

Proof of Corollary 3.

Fix $\mathbf{x}$ outside the exceptional null sets and $\mathbf{c} \in \mathbb{R}^d$. For concise exposition, write $t := \|\mathbf{c} - \mathbf{g}_0(\mathbf{x})\|_2$, $\tilde{r} := \min\{r_0, \rho\}$, and $\kappa_{\text{dir}} := \min\{\mu/L, \mu\gamma\tilde{r}/\bar{\beta}\}$.

If $0 \leq t \leq \tilde{r}$, then $t \leq r_0$, so (22) applies, and $t \leq \rho$, so (EC.4) is in its quadratic regime. The two give, respectively,

$$\Delta_{\gamma}(\mathbf{c}; \mathbf{x}) \leq \frac{L}{2} t^2, \quad \Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) \geq \frac{\mu}{2} t^2.$$

Consequently,

$$\Delta_{\gamma}(\mathbf{c}; \mathbf{x}) \leq \frac{L}{\mu} \Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) \leq \frac{1}{\kappa_{\text{dir}}} \Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}),$$

where the second inequality uses $\kappa_{\text{dir}} \leq \mu/L$.

If $t > \tilde{r}$, then Lemma EC.2, Condition 1(iii)–(iv) and the $1/\gamma$ -Lipschitz property of $\mathbf{z}_{\gamma}^*(\cdot)$ give

$$\Delta_{\gamma}(\mathbf{c}; \mathbf{x}) = \bar{\mathbf{y}}(\mathbf{x})^{\top} (\mathbf{z}_{\gamma}^*(\mathbf{c}) - \mathbf{z}_{\gamma}^*(\mathbf{g}_0(\mathbf{x}))) \leq \frac{\bar{\beta}}{2\gamma} t.$$

Since $\tilde{r} = \min\{r_0, \rho\}$, the condition $t > \tilde{r}$ implies

$$\min\{t^2, \rho t\} \geq \tilde{r} t.$$

Indeed, if $t \leq \rho$ then $t^2 > \tilde{r} t$, and if $t > \rho$ then $\rho t \geq \tilde{r} t$. Thus (EC.4) gives

$$\Delta_{\gamma}^+(\mathbf{c}; \mathbf{x}) \geq \frac{\mu\tilde{r}}{2} t,$$

and it follows that

$$\Delta_\gamma(\mathbf{c}; \mathbf{x}) \leq \frac{\bar{\beta}}{\mu\gamma\tilde{r}} \Delta_\gamma^+(\mathbf{c}; \mathbf{x}) \leq \frac{1}{\kappa_{\text{dir}}} \Delta_\gamma^+(\mathbf{c}; \mathbf{x}),$$

where the last inequality uses $\kappa_{\text{dir}} \leq \mu\gamma\tilde{r}/\bar{\beta}$. The two cases establish the global comparison

$$\Delta_\gamma(\mathbf{c}; \mathbf{x}) \leq \frac{1}{\kappa_{\text{dir}}} \Delta_\gamma^+(\mathbf{c}; \mathbf{x}), \quad \mathbf{c} \in \mathbb{R}^d. \quad (\text{EC.13})$$

Applying (EC.13) at $\mathbf{c} = \hat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x})$ and integrating, we have

$$\begin{aligned} R(\hat{\mathbf{g}}_{\text{RSPO}_+}) - R^* &= \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [R(\hat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}); \mathbf{x}) - R(\mathbf{g}_0(\mathbf{x}); \mathbf{x})] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [\Delta_\gamma(\hat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}); \mathbf{x})] \\ &\leq \frac{1}{\kappa_{\text{dir}}} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [\Delta_\gamma^+(\hat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}); \mathbf{x})] \\ &= \frac{1}{\kappa_{\text{dir}}} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [R_+(\hat{\mathbf{g}}_{\text{RSPO}_+}(\mathbf{x}); \mathbf{x}) - R_+(\mathbf{g}_0(\mathbf{x}); \mathbf{x})] \\ &= \frac{1}{\kappa_{\text{dir}}} [R_+(\hat{\mathbf{g}}_{\text{RSPO}_+}) - R_+(\mathbf{g}_0)] \leq \frac{B_N + A_{\mathcal{G}}^+}{\kappa_{\text{dir}}}, \end{aligned}$$

which proves (23).

For the rate, fix $\eta \in (0, 1)$ and take $\delta = \eta$. Since $\mathbf{g}_0 \in \mathcal{G}$, we have $R_+(\mathcal{G}) = R_+(\mathcal{G}_{\text{all}})$ by Condition 1(i). Thus, $\bar{c}(\eta) = \sqrt{N} B_N$ of (EC.12) does not depend on N , so on an event of probability at least $1 - \eta$,

$$N^{1/2} [R(\hat{\mathbf{g}}_{\text{RSPO}_+}) - R^*] \leq \frac{\bar{c}(\eta)}{\kappa_{\text{dir}}}, \quad N \geq 1,$$

and $\kappa_{\text{dir}} > 0$ does not depend on N , which gives the stated $\mathcal{O}_p(N^{-1/2})$ rate. $\square$

Appendix C The Stochastic Gradient-Descent Algorithm

The procedure of the stochastic gradient descent scheme with Armijo Backtracking for Problem (14) is summarized in Algorithm 1, in which

$$\hat{L}_{B_t}(\theta) = \frac{1}{B_s} \sum_{i \in B_t} [\mathbf{y}_i^\top \mathbf{z}_\gamma^*(\mathbf{g}_\theta(\mathbf{x}_i)) - v^*(\mathbf{y}_i)] + \lambda \Omega(\mathbf{g}_\theta)$$

denotes the batch counterpart of (14). Taking $B_s = N$ recovers the full-batch method. Per iteration, the gradient phase solves the decision problem **ILRO-DECISION** once per sample and forms one $|S_i^{(t)}| \times |S_i^{(t)}|$ linear system per sample for $\mathbf{q}_i^{(t)}$; each line-search trial $k = 0, 1, \dots$ costs another B_s decision map evaluations. The total is $(k^* + 2)B_s$ decision map evaluations when the step is accepted at trial k^* , and at most $(K + 2)B_s$ otherwise. The cost is thus dominated by the decision map, which favors a small number of refinement steps over running the scheme to convergence.

For the full-batch version the method is a genuine descent method, and the line search terminates after at most $K + 1$ trials by construction. The proposition below shows that a strictly decreasing step is accepted whenever $\hat{L}_N$ is differentiable at the current iterate with nonzero gradient, provided K is large enough.

PROPOSITION EC.1 (Monotone Descent). *Let $B_s = N$, let $\mathcal{Z}$ be a nonempty bounded polyhedron, and let $\hat{L}_N$ be given by (14) with $\theta \rightarrow \Omega(\mathbf{g}_\theta)$ differentiable and bounded below. Then $\{\hat{L}_N(\theta^{(t)})\}_{t \geq 0}$ is nonincreasing and bounded below, hence convergent. Moreover, if $\hat{L}_N$ is differentiable at $\theta^{(t)}$ and $\mathbf{h}^{(t)} = \nabla \hat{L}_N(\theta^{(t)}) \neq \mathbf{0}$, then the Armijo test is satisfied for all sufficiently small step-sizes, so for K large enough the step is accepted and the resulting decrease is at least $c_{\text{ls}} \rho_{\text{ls}}^{k^*} \eta_t \|\mathbf{h}^{(t)}\|_2^2 > 0$.*

Proof of Proposition EC.1. Monotonicity is immediate from the acceptance rule: either the test succeeds, in which case

$$\hat{L}_N(\theta^{(t+1)}) \leq \hat{L}_N(\theta^{(t)}) - c_{\text{ls}} \rho_{\text{ls}}^{k^*} \eta_t \|\mathbf{h}^{(t)}\|_2^2 \leq \hat{L}_N(\theta^{(t)}),$$

or it fails and $\theta^{(t+1)} = \theta^{(t)}$. For the lower bound, $\ell_{\mathcal{RSPD}} \geq 0$ by Proposition 3(i) and Ω is bounded below by assumption, so $\hat{L}_N \geq \lambda \inf_{\theta \in \Theta} \Omega(\mathbf{g}_\theta) > -\infty$. A nonincreasing sequence that is bounded below converges.

For the second claim, suppose $\hat{L}_N$ is differentiable at $\theta^{(t)}$ with $\mathbf{h}^{(t)} = \nabla \hat{L}_N(\theta^{(t)}) \neq \mathbf{0}$. Differentiability gives

$$\hat{L}_N(\theta^{(t)} - s\mathbf{h}^{(t)}) = \hat{L}_N(\theta^{(t)}) - s\|\mathbf{h}^{(t)}\|_2^2 + o(s), \quad s \downarrow 0,$$

so that

$$\hat{L}_N(\theta^{(t)} - s\mathbf{h}^{(t)}) - \hat{L}_N(\theta^{(t)}) + c_{\text{ls}} s \|\mathbf{h}^{(t)}\|_2^2 = -(1 - c_{\text{ls}}) s \|\mathbf{h}^{(t)}\|_2^2 + o(s) < 0$$

Algorithm 1: Stochastic Gradient Descent with Armijo Backtracking for Problem (14)**Input:** sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$; polyhedral set $\mathcal{Z} = \{\mathbf{z} : \mathbf{A}\mathbf{z} \geq \mathbf{b}\}$; $\gamma > 0$; $\Omega(\cdot)$, $\lambda \geq 0$; stepsize $\eta_0 > 0$; batch size B_s ($B_s = N$ gives the full-batch method); iteration budget T ;backtracking parameters $\rho_{\text{ls}} \in (0, 1)$, $K \in \mathbb{N}$, $c_{\text{ls}} \in (0, 1)$.

1 **Initialize:** $\theta^{(0)} \leftarrow$ solution of the $\mathcal{RSP}\mathcal{O}_+$ problem (8);

2 **for** $t = 0, 1, \dots, T - 1$ **do**

3 Sample $\mathcal{B}_t \subseteq [N]$, $|\mathcal{B}_t| = B_s$, uniformly without replacement;

 // batch gradient

4 **for** $i \in \mathcal{B}_t$ **do**

5 $\hat{\mathbf{y}}_i^{(t)} \leftarrow \mathbf{g}_{\theta^{(t)}}(\mathbf{x}_i)$, $\mathbf{z}_i^{(t)} \leftarrow \mathbf{z}_\gamma^*(\hat{\mathbf{y}}_i^{(t)})$, $\mathbf{q}_i^{(t)} \leftarrow \nabla \mathbf{z}_\gamma^*(\hat{\mathbf{y}}_i^{(t)})^\top \mathbf{y}_i$ via (13) with

$S_i^{(t)} = \{j : \mathbf{a}_j^\top \mathbf{z}_i^{(t)} = b_j\}$;

6 $\mathbf{h}^{(t)} \leftarrow \frac{1}{B_s} \sum_{i \in \mathcal{B}_t} [\nabla_{\theta} \mathbf{g}_{\theta^{(t)}}(\mathbf{x}_i)]^\top \mathbf{q}_i^{(t)} + \lambda \nabla_{\theta} \Omega(\mathbf{g}_{\theta^{(t)}})$;

7 $\eta_t \leftarrow \eta_0$ if $B_s = N$, and $\eta_t \leftarrow \eta_0 / \sqrt{t + 1}$ otherwise;

 // Armijo backtracking

8 **if** $\exists k \in \{0, \dots, K\} : \hat{L}_{\mathcal{B}_t}(\theta^{(t)} - \rho_{\text{ls}}^k \eta_t \mathbf{h}^{(t)}) \leq \hat{L}_{\mathcal{B}_t}(\theta^{(t)}) - c_{\text{ls}} \rho_{\text{ls}}^k \eta_t \|\mathbf{h}^{(t)}\|_2^2$ **then**

9 $\theta^{(t+1)} \leftarrow \theta^{(t)} - \rho_{\text{ls}}^{k^*} \eta_t \mathbf{h}^{(t)}$, where k^* is the smallest such k ;

10 **else**

11 $\theta^{(t+1)} \leftarrow \theta^{(t)}$;

Output: $\theta^{(T)}$.

for all sufficiently small $s > 0$, because $c_{\text{ls}} < 1$ and $\|\mathbf{h}^{(t)}\|_2 > 0$. Since $\rho_{\text{ls}} \in (0, 1)$, the trial stepsizes $\rho_{\text{ls}}^k \eta_t$ decrease to zero as k grows, so the Armijo test is satisfied for all k large enough. If K is at least the smallest such index, the step is accepted, and the stated decrease is the acceptance inequality itself. $\square$

Appendix D Concrete Excess Risk Bounds of the $\mathcal{RSP}\mathcal{O}$ Predictor

In this appendix, we instantiate the meta generalization bound of Theorem 3 for four representative classes of predictors: (i) the bounded affine class; (ii) the polynomial discrimination class; (iii) the Dudley-entropy-integral bounded class; and (iv) the vector-valued RKHS class. Throughout, we assume $\gamma > 0$ and $\|\mathbf{y}\|_2 \leq r$ for all $\mathbf{y} \in \mathcal{Y}$, as stated in Theorem 3. The scalar antecedents of the complexity bounds below are classical; the content of this appendix is their adaptation to the multivariate Rademacher complexity $\mathfrak{R}_d$ of Maurer (2016), with the explicit constants required by Table 2. In each example, the excess risk bound follows from the complexity bound by a direct application of Theorem 3, so only the complexity bounds require proof.

D.1 Bounded Affine Class

We begin with affine predictors, which cover the linear models used in predict-then-optimize pipelines and yield an explicit dimension-dependent bound. The scalar analogue of the bound below is classical (Bartlett and Mendelson 2002, Kakade et al. 2008); see also Maurer (2006) for complexity bounds on classes of linear transformations. We include a short proof because the multivariate form with the explicit factor $\sqrt{d}$ does not appear verbatim in these references.

ASSUMPTION EC.1. *There exists a context map $\varphi : \mathcal{X} \rightarrow \mathbb{R}^m$ with $\|\varphi(\mathbf{x})\|_2 \leq \kappa$ for all $\mathbf{x} \in \mathcal{X}$, and the hypothesis class is*

$$\mathcal{G}_{\text{aff}} = \{\mathbf{g}_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}\varphi(\mathbf{x}) : \mathbf{W} \in \mathbb{R}^{d \times m}, \|\mathbf{W}\|_F \leq B\}.$$

EXAMPLE EC.1 (BOUNDED AFFINE CLASS). Under Assumption EC.1, the multivariate Rademacher complexity satisfies

$$\mathfrak{R}_d(\mathcal{G}_{\text{aff}}) \leq B\kappa\sqrt{\frac{d}{N}}, \quad (\text{EC.14})$$

and Theorem 3 therefore yields that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R(\hat{\mathbf{g}}_{\text{RSPO}}) - R^*(\mathcal{G}_{\text{aff}}) \leq \frac{4\sqrt{2}rB\kappa}{\gamma} \sqrt{\frac{d}{N}} + b\sqrt{\frac{2\log(2/\delta)}{N}}.$$

The bound is explicit in the decision dimension: the two radii B and κ enter multiplicatively, and the factor $\sqrt{d}$ reflects that a d -dimensional cost vector must be predicted from the same sample.

Proof of Example EC.1. By the definition of multivariate Rademacher complexity,

$$\mathfrak{R}_d(\mathcal{G}_{\text{aff}}) = \mathbb{E}_{\boldsymbol{\sigma}, \mathbf{X}} \left[\sup_{\|\mathbf{W}\|_F \leq B} \frac{1}{N} \sum_{i=1}^N \boldsymbol{\sigma}_i^\top \mathbf{W} \varphi(\mathbf{x}_i) \right],$$

where $\boldsymbol{\sigma}_i \in \{+1, -1\}^d$ are independent Rademacher random vectors. Using the Frobenius inner product, we have

$$\sum_{i=1}^N \boldsymbol{\sigma}_i^\top \mathbf{W} \varphi(\mathbf{x}_i) = \left\langle \mathbf{W}, \sum_{i=1}^N \boldsymbol{\sigma}_i \varphi(\mathbf{x}_i)^\top \right\rangle_F.$$

Therefore, by the Cauchy–Schwarz inequality,

$$\mathfrak{R}_d(\mathcal{G}_{\text{aff}}) \leq \frac{B}{N} \mathbb{E}_{\sigma, \mathbf{X}} \left[\left\| \sum_{i=1}^N \sigma_i \boldsymbol{\varphi}(\mathbf{x}_i)^\top \right\|_F \right].$$

By Jensen’s inequality,

$$\mathbb{E}_{\sigma, \mathbf{X}} \left[\left\| \sum_{i=1}^N \sigma_i \boldsymbol{\varphi}(\mathbf{x}_i)^\top \right\|_F \right] \leq \left(\mathbb{E}_{\sigma, \mathbf{X}} \left[\left\| \sum_{i=1}^N \sigma_i \boldsymbol{\varphi}(\mathbf{x}_i)^\top \right\|_F^2 \right] \right)^{1/2}.$$

Since the Rademacher vectors are independent and centered, the cross terms vanish, and thus

$$\mathbb{E}_{\sigma, \mathbf{X}} \left[\left\| \sum_{i=1}^N \sigma_i \boldsymbol{\varphi}(\mathbf{x}_i)^\top \right\|_F^2 \right] = \sum_{i=1}^N \mathbb{E}_{\sigma, \mathbf{X}} [\|\sigma_i\|_2^2 \|\boldsymbol{\varphi}(\mathbf{x}_i)\|_2^2] \leq Nd\kappa^2,$$

where we used $\|\sigma_i\|_2^2 = d$ and $\|\boldsymbol{\varphi}(\mathbf{x}_i)\|_2 \leq \kappa$. Hence,

$$\mathfrak{R}_d(\mathcal{G}_{\text{aff}}) \leq \frac{B}{N} \sqrt{Nd\kappa^2} = B\kappa \sqrt{\frac{d}{N}}.$$

The proof is complete. $\square$

D.2 Polynomial Discrimination Class

Affine classes do not cover discontinuous or finitely generated prediction rules. We next consider classes whose number of distinct prediction patterns on any finite sample grows only polynomially in the sample size. The scalar version of the resulting bound is standard (see, e.g., [Wainwright 2019](#), Chapter 4); the vector-valued case reduces to it by identifying $(\mathbb{R}^d)^N$ with $\mathbb{R}^{Nd}$.

DEFINITION EC.1. Let $\mathcal{G} \subseteq \{\mathbf{g} : \mathcal{X} \rightarrow \mathbb{R}^d\}$ and, for any collection $\mathbf{x}_1^N = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, define $\mathcal{G}(\mathbf{x}_1^N) := \{(\mathbf{g}(\mathbf{x}_1), \dots, \mathbf{g}(\mathbf{x}_N)) : \mathbf{g} \in \mathcal{G}\} \subseteq (\mathbb{R}^d)^N$. We say that $\mathcal{G}$ has polynomial discrimination of order $\nu \geq 1$ if $\text{card}(\mathcal{G}(\mathbf{x}_1^N)) \leq (N+1)^\nu$ for every positive integer N and every collection $\mathbf{x}_1^N$. The exponent ν may depend on the output dimension d .

ASSUMPTION EC.2. *The hypothesis class $\mathcal{G}$ has polynomial discrimination of order $\nu \geq 1$ in the sense of Definition EC.1, and there is a constant $B_1 > 0$ such that*

$$D_{\mathcal{G}}(\mathbf{x}^N) := \sup_{\mathbf{g} \in \mathcal{G}} \left(\frac{1}{N} \sum_{i=1}^N \|\mathbf{g}(\mathbf{x}_i)\|_2^2 \right)^{1/2} \leq B_1$$

for every collection $\mathbf{x}_1^N$.

EXAMPLE EC.2 (POLYNOMIAL DISCRIMINATION CLASS). Under Assumption EC.2, the multivariate Rademacher complexity satisfies

$$\mathfrak{R}_d(\mathcal{G}) \leq B_1 \sqrt{\frac{2\nu \log(N+1)}{N}}, \tag{EC.15}$$

and Theorem 3 therefore yields that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R(\widehat{\mathcal{G}}_{\text{RSPO}}) - R^*(\mathcal{G}) \leq \frac{8rB_2}{\gamma} \sqrt{\frac{\nu \log(N+1)}{N}} + b \sqrt{\frac{2\log(2/\delta)}{N}}.$$

Here the dimension d does not appear explicitly; its effect is absorbed into the discrimination order ν , which counts prediction patterns rather than parameters. The price of this generality is the factor $\sqrt{\log(N+1)}$, since the argument controls a finite but growing set of sample patterns instead of exploiting a parameterization.

Proof of Example EC.2. Fix a collection $\mathbf{x}_1^N$ and identify each sample evaluation $(\mathbf{g}(\mathbf{x}_1), \dots, \mathbf{g}(\mathbf{x}_N))$ with a vector in $\mathbb{R}^{Nd}$, so that the empirical multivariate Rademacher complexity equals $\frac{1}{N} \mathbb{E}_{\boldsymbol{\sigma}} [\sup_{\mathbf{v} \in \mathcal{G}(\mathbf{x}_1^N)} \langle \boldsymbol{\sigma}, \mathbf{v} \rangle]$ with $\boldsymbol{\sigma} \in \{\pm 1\}^{Nd}$. Since $\text{card}(\mathcal{G}(\mathbf{x}_1^N)) \leq (N+1)^\nu$ by Definition EC.1 and $\sup_{\mathbf{v} \in \mathcal{G}(\mathbf{x}_1^N)} \|\mathbf{v}\|_2 \leq \sqrt{N}B_1$ by Assumption EC.2, Massart's finite class lemma (Massart 2000) yields

$$\frac{1}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{v} \in \mathcal{G}(\mathbf{x}_1^N)} \langle \boldsymbol{\sigma}, \mathbf{v} \rangle \right] \leq \frac{1}{N} \sqrt{2\nu \log(N+1)} \cdot \sqrt{N}B_1 = B_1 \sqrt{\frac{2\nu \log(N+1)}{N}}.$$

Taking expectation over the sample establishes (EC.15). $\square$

REMARK EC.1 (THE TWO CLASSES ARE NOT NESTED). Neither class contains the other. A bounded affine class generates infinitely many prediction patterns already on even a single sample point: for the identity context map and any $\mathbf{x}_0 \neq \mathbf{0}$, the choice $\mathbf{W}_a = \frac{a}{\|\mathbf{x}_0\|_2^2} \mathbf{e}_1 \mathbf{x}_0^\top$ satisfies $\|\mathbf{W}_a\|_F \leq B$ for every $a \in [-B\|\mathbf{x}_0\|_2, B\|\mathbf{x}_0\|_2]$ while producing the distinct values $\mathbf{W}_a \mathbf{x}_0 = a \mathbf{e}_1$, so $\text{card}(\mathcal{G}_{\text{aff}}(\mathbf{x}_0)) = \infty$ and polynomial discrimination fails. Conversely, the scalar threshold class $\mathcal{G}_{\text{th}} = \{g_t(x) = \mathbf{1}\{x \leq t\} : t \in \mathbb{R}\}$ induces at most $N+1$ patterns on any sample and hence has polynomial discrimination of order $\nu = 1$, yet the family $\{\mathbf{1}\{x \leq t\} : t \in \mathbb{R}\}$ contains infinitely many linearly independent functions, whereas any class of the form $\{\mathbf{w}^\top \boldsymbol{\varphi}(\cdot)\}$ spans a function space of dimension at most m . Therefore, $\mathcal{G}_{\text{th}}$ is not contained in any bounded affine class.

D.3 Dudley-entropy-integral Bounded Class

The preceding arguments rely either on a linear parameterization or on a finite number of sample prediction patterns. For more general classes we control the complexity through a uniform entropy integral. The bound below is Dudley's entropy integral theorem (Wainwright 2019, Theorem 5.22); the only step specific to our setting is verifying that the vector-valued Rademacher process is sub-Gaussian with respect to the scaled empirical metric $d_N/\sqrt{N}$, which we record in the proof.

DEFINITION EC.2 (COVERING NUMBER). Fix a sample $\mathbf{x}^N$ and equip $\mathcal{G}$ with the empirical metric d_N . For $\epsilon > 0$, a finite collection $\mathbf{g}_1, \dots, \mathbf{g}_M \in \mathcal{G}$ is an ϵ -cover of $\mathcal{G}$ if every $\mathbf{g} \in \mathcal{G}$ satisfies $d_N(\mathbf{g}, \mathbf{g}_m) \leq \epsilon$ for some $m \in [M]$. The covering number $\mathcal{N}(\epsilon, \mathcal{G}, d_N)$ is the smallest cardinality of such a cover, set to $+\infty$ when no finite cover exists.

The covering number measures how many representative predictors are needed to approximate every member of $\mathcal{G}$ to accuracy ϵ on the sample, and its logarithm $\log \mathcal{N}(\epsilon, \mathcal{G}, d_N)$ is referred to as the metric entropy of $\mathcal{G}$.

ASSUMPTION EC.3. Let $\mathcal{G} \subseteq \{\mathbf{g} : \mathcal{X} \rightarrow \mathbb{R}^d\}$ and, for any sample $\mathbf{x}^N$, define the empirical metric $d_N(\mathbf{g}, \tilde{\mathbf{g}}) := \left(\frac{1}{N} \sum_{i=1}^N \|\mathbf{g}(\mathbf{x}_i) - \tilde{\mathbf{g}}(\mathbf{x}_i)\|_2^2\right)^{1/2}$. There exist constants $\Delta_{\mathcal{G}} > 0$ and $J_{\mathcal{G}} < \infty$, independent of N and of the sample, such that $\text{diam}_N(\mathcal{G}) := \sup_{\mathbf{g}, \tilde{\mathbf{g}} \in \mathcal{G}} d_N(\mathbf{g}, \tilde{\mathbf{g}}) \leq \Delta_{\mathcal{G}}$ and $\int_0^{\Delta_{\mathcal{G}}} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{G}, d_N)} d\epsilon \leq J_{\mathcal{G}}$, where $\mathcal{N}(\epsilon, \mathcal{G}, d_N)$ is the covering number of $\mathcal{G}$ under d_N .

EXAMPLE EC.3 (DUDLEY-ENTROPY-INTEGRAL BOUNDED CLASS). Under Assumption EC.3, there exists a universal constant $C > 0$ such that

$$\mathfrak{R}_d(\mathcal{G}) \leq C \frac{J_{\mathcal{G}}}{\sqrt{N}}, \quad (\text{EC.16})$$

and Theorem 3 therefore yields that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R(\hat{\mathbf{g}}_{\text{RSPO}}) - R^*(\mathcal{G}) \leq \frac{4\sqrt{2}Cr}{\gamma} \frac{J_{\mathcal{G}}}{\sqrt{N}} + b\sqrt{\frac{2\log(2/\delta)}{N}}.$$

The entire geometry of the class is summarized by the single quantity $J_{\mathcal{G}}$, and neither the dimension d nor any parameterization appears. This is what allows the guarantee to cover infinite-dimensional classes, at the cost of requiring covering numbers that are uniform over samples.

Proof of Example EC.3. Fix a sample $\mathbf{x}^N = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ and for each $\mathbf{g} \in \mathcal{G}$, define the Rademacher process

$$Z_{\mathbf{g}} := \frac{1}{N} \sum_{i=1}^N \boldsymbol{\sigma}_i^\top \mathbf{g}(\mathbf{x}_i),$$

where $\boldsymbol{\sigma}_i = (\sigma_{i1}, \dots, \sigma_{id})$ and the coordinates $\{\sigma_{ij}\}$ are independent Rademacher random variables. Then, for any $\mathbf{g}, \tilde{\mathbf{g}} \in \mathcal{G}$ and any $\lambda \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\sigma}} [\exp \{\lambda (Z_{\mathbf{g}} - Z_{\tilde{\mathbf{g}}})\}] &= \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp \left\{ \frac{\lambda}{N} \sum_{i=1}^N \boldsymbol{\sigma}_i^\top (\mathbf{g}(\mathbf{x}_i) - \tilde{\mathbf{g}}(\mathbf{x}_i)) \right\} \right] \\ &\leq \exp \left\{ \frac{\lambda^2}{2N^2} \sum_{i=1}^N \|\mathbf{g}(\mathbf{x}_i) - \tilde{\mathbf{g}}(\mathbf{x}_i)\|_2^2 \right\} \\ &= \exp \left\{ \frac{\lambda^2}{2N} d_N^2(\mathbf{g}, \tilde{\mathbf{g}}) \right\}. \end{aligned}$$

Hence $\{Z_{\mathbf{g}} : \mathbf{g} \in \mathcal{G}\}$ is a centered sub-Gaussian process with respect to the metric $d_N/\sqrt{N}$.

Fix any $\mathbf{g}_0 \in \mathcal{G}$. Since $\mathbb{E}_{\boldsymbol{\sigma}}[Z_{\mathbf{g}_0}] = 0$, we have $\mathbb{E}_{\boldsymbol{\sigma}}[\sup_{\mathbf{g} \in \mathcal{G}} Z_{\mathbf{g}}] \leq \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{\mathbf{g} \in \mathcal{G}} (Z_{\mathbf{g}} - Z_{\mathbf{g}_0})]$. By Dudley's entropy integral bound (Wainwright 2019, Theorem 5.22), there exists a universal constant $C > 0$ such that

$$\mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{g} \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^N \boldsymbol{\sigma}_i^\top \mathbf{g}(\mathbf{x}_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{g} \in \mathcal{G}} Z_{\mathbf{g}} \right] \leq \frac{C}{\sqrt{N}} \int_0^{\text{diam}_N(\mathcal{G})} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{G}, d_N)} d\epsilon.$$

Since $\mathcal{N}(\epsilon, \mathcal{G}, d_N/\sqrt{N}) = \mathcal{N}(\sqrt{N}\epsilon, \mathcal{G}, d_N)$, the substitution $u = \sqrt{N}\epsilon$ pulls out the factor $1/\sqrt{N}$. By Assumption EC.3, the entropy integral is uniformly bounded by $J_{\mathcal{G}}$, and therefore

$$\mathbb{E}_{\sigma} \left[\sup_{\mathbf{g} \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^N \sigma_i^{\top} \mathbf{g}(\mathbf{x}_i) \right] \leq C \frac{J_{\mathcal{G}}}{\sqrt{N}}.$$

Since this bound holds for every fixed sample $\mathbf{x}^N$, taking expectation with respect to the sample gives

$$\mathfrak{R}_d(\mathcal{G}) \leq C \frac{J_{\mathcal{G}}}{\sqrt{N}}.$$

This establishes (EC.16), and the proof is complete. $\square$

Assumption EC.3 covers smooth infinite-dimensional classes that the previous two conditions do not. Let $\mathcal{X} = [0, 1]^p$ and consider the vector-valued Hölder class $\mathcal{G}_{s,L} = \{\mathbf{g} = (g_1, \dots, g_d) : g_j \in C^s([0, 1]^p), \|g_j\|_{C^s} \leq L\}$ with $s > p/2$. Since $d_N(\mathbf{g}, \tilde{\mathbf{g}}) \leq \sqrt{d} \max_j \|g_j - \tilde{g}_j\|_{\infty}$ and the C^s norm dominates the uniform norm, one may take $\Delta_{\mathcal{G}} = 2L\sqrt{d}$; applying the standard entropy bound for scalar Hölder balls coordinatewise and taking product covers gives $\log \mathcal{N}(\epsilon, \mathcal{G}_{s,L}, d_N) \leq C_{p,s} d(L\sqrt{d}/\epsilon)^{p/s}$, whose entropy integral is finite because $s > p/2$. The class contains a continuum of values at any fixed $\mathbf{x}_0$ and therefore lies outside the finite-pattern regime of Definition EC.1.

D.4 Vector-valued RKHS class

Finally, we consider norm balls in a vector-valued reproducing kernel Hilbert space (Micchelli and Pontil 2005), which fit the meta theorem in the same way as the previous three classes. The scalar kernel-class bound is classical (Bartlett and Mendelson 2002); see also Mohri et al. (2018). We give a short proof of the vector-valued form with the coordinate-sum norm constraint of Assumption EC.4, since this exact statement does not appear in these references.

ASSUMPTION EC.4. *Let $\mathcal{H}$ be an RKHS on $\mathcal{X}$ with reproducing kernel K , and assume that K is bounded, in the sense that $\sup_{\mathbf{x} \in \mathcal{X}} K(\mathbf{x}, \mathbf{x}) \leq \kappa^2$. For some $R_{\mathcal{H}} > 0$, the prediction class is*

$$\mathcal{G}_{\mathcal{H}} := \left\{ \mathbf{g} = (g_1, \dots, g_d) : g_j \in \mathcal{H}, \sum_{j=1}^d \|g_j\|_{\mathcal{H}}^2 \leq R_{\mathcal{H}}^2 \right\}.$$

EXAMPLE EC.4 (VECTOR-VALUED RKHS BALL). Under Assumption EC.4, the multivariate Rademacher complexity satisfies

$$\mathfrak{R}_d(\mathcal{G}_{\mathcal{H}}) \leq \kappa R_{\mathcal{H}} \sqrt{\frac{d}{N}}, \quad (\text{EC.17})$$

and Theorem 3 therefore yields that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$R(\hat{\mathbf{g}}_{\text{RSPO}}) - R^*(\mathcal{G}_{\mathcal{H}}) \leq \frac{4\sqrt{2}r\kappa R_{\mathcal{H}}}{\gamma} \sqrt{\frac{d}{N}} + b \sqrt{\frac{2\log(2/\delta)}{N}}.$$

The guarantee has the same form as the affine case, with the Frobenius radius B replaced by the RKHS radius $R_{\mathcal{H}}$ and the context bound κ by the kernel bound. The class is infinite-dimensional,

yet the $\sqrt{d/N}$ dependence persists: it is the dimension of the cost vector, not of the function space, that the complexity tracks.

Proof of Example EC.4. Fix a sample $\mathbf{x}^N = (\mathbf{x}_1, \dots, \mathbf{x}_N)$. By the reproducing property, $g_j(\mathbf{x}_i) = \langle g_j, K(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}}$ for every $j \in [d]$ and $i \in [N]$. Hence, for any $\mathbf{g} = (g_1, \dots, g_d) \in \mathcal{G}_{\mathcal{H}}$,

$$\frac{1}{N} \sum_{i=1}^N \boldsymbol{\sigma}_i^\top \mathbf{g}(\mathbf{x}_i) = \sum_{j=1}^d \langle g_j, h_j \rangle_{\mathcal{H}}, \quad h_j := \frac{1}{N} \sum_{i=1}^N \sigma_{ij} K(\cdot, \mathbf{x}_i).$$

By the Cauchy–Schwarz inequality, applied first in $\mathcal{H}$ and then in $\mathbb{R}^d$,

$$\sum_{j=1}^d \langle g_j, h_j \rangle_{\mathcal{H}} \leq \left(\sum_{j=1}^d \|g_j\|_{\mathcal{H}}^2 \right)^{1/2} \left(\sum_{j=1}^d \|h_j\|_{\mathcal{H}}^2 \right)^{1/2} \leq R_{\mathcal{H}} \left(\sum_{j=1}^d \|h_j\|_{\mathcal{H}}^2 \right)^{1/2}.$$

Moreover, since the Rademacher variables are independent and centered,

$$\mathbb{E}_{\boldsymbol{\sigma}} [\|h_j\|_{\mathcal{H}}^2] = \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N \mathbb{E}[\sigma_{ij} \sigma_{i'j}] K(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{1}{N^2} \sum_{i=1}^N K(\mathbf{x}_i, \mathbf{x}_i) \leq \frac{\kappa^2}{N}.$$

Therefore, by Jensen’s inequality,

$$\mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{g} \in \mathcal{G}_{\mathcal{H}}} \frac{1}{N} \sum_{i=1}^N \boldsymbol{\sigma}_i^\top \mathbf{g}(\mathbf{x}_i) \right] \leq R_{\mathcal{H}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\left(\sum_{j=1}^d \|h_j\|_{\mathcal{H}}^2 \right)^{1/2} \right] \leq R_{\mathcal{H}} \left(\sum_{j=1}^d \mathbb{E}_{\boldsymbol{\sigma}} [\|h_j\|_{\mathcal{H}}^2] \right)^{1/2} \leq \kappa R_{\mathcal{H}} \sqrt{\frac{d}{N}}.$$

Taking the expectation with respect to the sample establishes (EC.17). $\square$

Appendix E Experimental Details and Additional Results

This appendix records the portfolio learning reformulations, quadratic-risk results, and case-specific calibration and implementation details supporting Section 6. The common data-generating, validation, testing, and evaluation procedures are stated in that section.

E.1 Portfolio Learning Reformulations

For the main portfolio experiment, the constraints in Section 6.2 define a bounded polyhedron after the standard epigraph reformulation of $\|\Sigma z\|_1$. Theorem 1 therefore yields finite-dimensional learning formulations for linear predictors.

ERM formulations. For linear predictors, the $\mathcal{RSP}\mathcal{O}_+$ empirical risk minimization problem admits the following equivalent reformulation:

$$\begin{aligned}
\min \quad & \frac{1}{N} \sum_{i \in [N]} \left[\beta \zeta_i + \mu_i + \frac{1}{2a\gamma} \|\mathbf{s}_i\|_2^2 + a \mathbf{z}_\gamma^\star(\mathbf{y}_i)^\top \mathbf{B} \mathbf{x}_i + \frac{a\gamma}{2} \|\mathbf{z}_\gamma^\star(\mathbf{y}_i)\|_2^2 - v^\star(\mathbf{y}_i) \right] + \lambda \|\mathbf{B}\|_F^2 \\
\text{s.t.} \quad & -\zeta_i \mathbf{1} \leq \mathbf{p}_i \leq \zeta_i \mathbf{1} & i \in [N] \\
& \mathbf{s}_i \geq \mathbf{y}_i - a \mathbf{B} \mathbf{x}_i - \Sigma \mathbf{p}_i - \mu_i \mathbf{1} & i \in [N] \\
& \mathbf{p}_i \in \mathbb{R}^d, \mathbf{s}_i \geq \mathbf{0}, \mu_i \geq 0, \zeta_i \geq 0 & i \in [N] \\
& \mathbf{B} \in \mathbb{R}^{d \times p}.
\end{aligned}$$

For comparison, the ILO benchmark based on the nominal $\mathcal{SP}\mathcal{O}_+$ loss can be reformulated as

$$\begin{aligned}
\min \quad & \frac{1}{N} \sum_{i \in [N]} \left[\beta \zeta_i + \mu_i + a \mathbf{z}^\star(\mathbf{y}_i)^\top \mathbf{B} \mathbf{x}_i - v^\star(\mathbf{y}_i) \right] + \lambda \|\mathbf{B}\|_F^2 \\
\text{s.t.} \quad & \mathbf{y}_i - a \mathbf{B} \mathbf{x}_i - \Sigma \mathbf{p}_i - \mu_i \mathbf{1} \leq \mathbf{0} & i \in [N] \\
& -\zeta_i \mathbf{1} \leq \mathbf{p}_i \leq \zeta_i \mathbf{1} & i \in [N] \\
& \mathbf{p}_i \in \mathbb{R}^d, \mu_i \geq 0, \zeta_i \geq 0 & i \in [N] \\
& \mathbf{B} \in \mathbb{R}^{d \times p}.
\end{aligned}$$

Dual representation. To obtain the first formulation, let $\mathbf{c} = \mathbf{y} - a \hat{\mathbf{y}}$ and consider

$$\max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{c}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\}, \quad \mathcal{Z} = \{\mathbf{z} \geq \mathbf{0} : \mathbf{1}^\top \mathbf{z} \leq 1, \|\Sigma \mathbf{z}\|_1 \leq \beta\}.$$

Introducing $\mathbf{w} = \Sigma \mathbf{z}$, multiplier $\mu \geq 0$ for $\mathbf{1}^\top \mathbf{z} \leq 1$, and multiplier $\mathbf{p}$ for $\mathbf{w} = \Sigma \mathbf{z}$ gives the Lagrangian of the equivalent minimization problem,

$$L(\mathbf{z}, \mathbf{w}, \mu, \mathbf{p}) = \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 - \mathbf{c}^\top \mathbf{z} + \mu(\mathbf{1}^\top \mathbf{z} - 1) + \mathbf{p}^\top (\Sigma \mathbf{z} - \mathbf{w}).$$

Because $\beta > 0$, a sufficiently small $\mathbf{z} > \mathbf{0}$ satisfies both inequalities strictly, so Slater's condition yields strong duality. Minimizing the Lagrangian over $\mathbf{z} \geq \mathbf{0}$ and $\|\mathbf{w}\|_1 \leq \beta$ gives

$$\max_{\mathbf{z} \in \mathcal{Z}} \left\{ \mathbf{c}^\top \mathbf{z} - \frac{a\gamma}{2} \|\mathbf{z}\|_2^2 \right\} = \min_{\mu \geq 0, \mathbf{p} \in \mathbb{R}^d} \left\{ \frac{1}{2a\gamma} \|(\mathbf{c} - \mu \mathbf{1} - \Sigma \mathbf{p})_+\|_2^2 + \beta \|\mathbf{p}\|_\infty + \mu \right\}.$$

Introducing $\mathbf{s}_i \geq \mathbf{y}_i - a \mathbf{B} \mathbf{x}_i - \mu_i \mathbf{1} - \Sigma \mathbf{p}_i$ with $\mathbf{s}_i \geq \mathbf{0}$ and $-\zeta_i \mathbf{1} \leq \mathbf{p}_i \leq \zeta_i \mathbf{1}$ gives the stated $\mathcal{RSP}\mathcal{O}_+$ ERM formulation; the $\mathcal{SP}\mathcal{O}_+$ formulation follows from the corresponding linear-program dual.

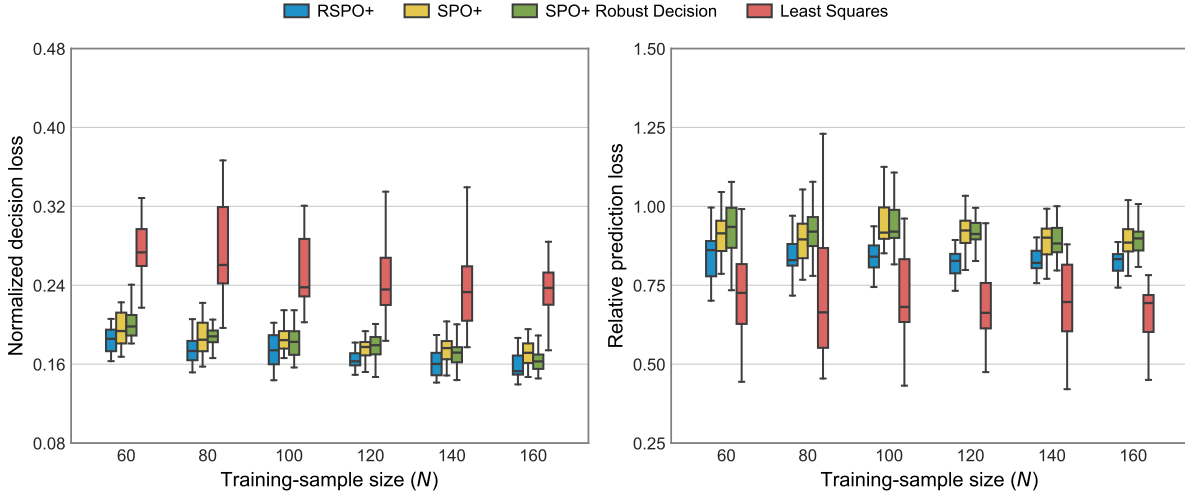


Figure EC.1 Normalized decision loss (left) and relative prediction loss (right) as N varies under nonlinear returns on quadratic-risk portfolio optimization instances, with $\text{deg} = 4$, $p = 80$, $d = 60$, and $\tau = 1$ fixed.

E.2 Quadratic-Risk Portfolio Experiment

We complement the main portfolio study by considering portfolio instances with the quadratic variance constraint used in the classical Markowitz formulation and in the portfolio experiment of [Elmachtoub and Grigas \(2022\)](#). Given a predicted cost vector $\hat{\mathbf{y}}$, the deployed portfolio solves

$$\mathbf{z}_\gamma^*(\hat{\mathbf{y}}) \in \arg \min_{\mathbf{z}} \left\{ \hat{\mathbf{y}}^\top \mathbf{z} + \frac{\gamma}{2} \|\mathbf{z}\|_2^2 : \mathbf{z} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{z} \leq 1, \mathbf{z}^\top \Sigma \mathbf{z} \leq \beta \right\},$$

where $\mathbf{z}^\top \Sigma \mathbf{z}$ is the portfolio variance and β is the variance budget; setting $\gamma = 0$ recovers the nominal decision problem. We use the factor-model return process from Section 6.2, augment each context vector with an unpenalized intercept, and set $\beta = 2.25 \mathbf{z}_{\text{unif}}^\top \Sigma \mathbf{z}_{\text{unif}}$ for $\mathbf{z}_{\text{unif}} = d^{-1} \mathbf{1}$. Because the quadratic constraint defines a nonpolyhedral feasible set, we train the four predictors using stochastic first-order updates based on their respective loss gradients or subgradients. For each configuration, we return the stepsize-weighted average of up to 1,000 iterates, using minibatches of size 10, stepsize $0.1/\sqrt{t}$, and gradient-norm tolerance 10^{-5} . The validation, testing, and replication procedures otherwise follow Section 6.

Figures EC.1 and EC.2 examine the effect of the training-sample size under nonlinear and linear return specifications. In both cases, $\mathcal{RSPO}_+$ attains the lowest normalized decision loss at every reported N , and its decision loss declines overall as the training sample grows. Its advantage over $\mathcal{SPO}_+$ and $\mathcal{SPO}_+$ with robust decisions persists under both specifications, while least squares has the largest decision loss. The prediction panels show that least squares attains lower relative prediction loss, whereas $\mathcal{RSPO}_+$ converts its predictions into better portfolio decisions.

Figure EC.3 examines the effect of the number of assets. Across $d \in \{20, 40, 60, 80, 100\}$, the decision loss of $\mathcal{RSPO}_+$ remains approximately stable and is the lowest among the four methods, while

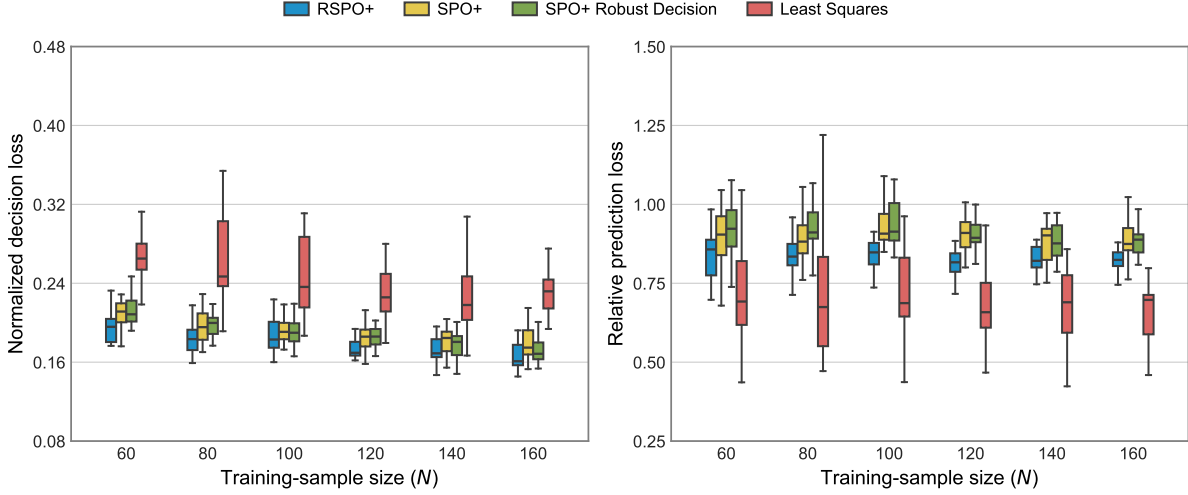


Figure EC.2 Normalized decision loss (left) and relative prediction loss (right) as N varies under linear returns on quadratic-risk portfolio optimization instances, with $\text{deg} = 1$, $p = 80$, $d = 60$, and $\tau = 1$ fixed.

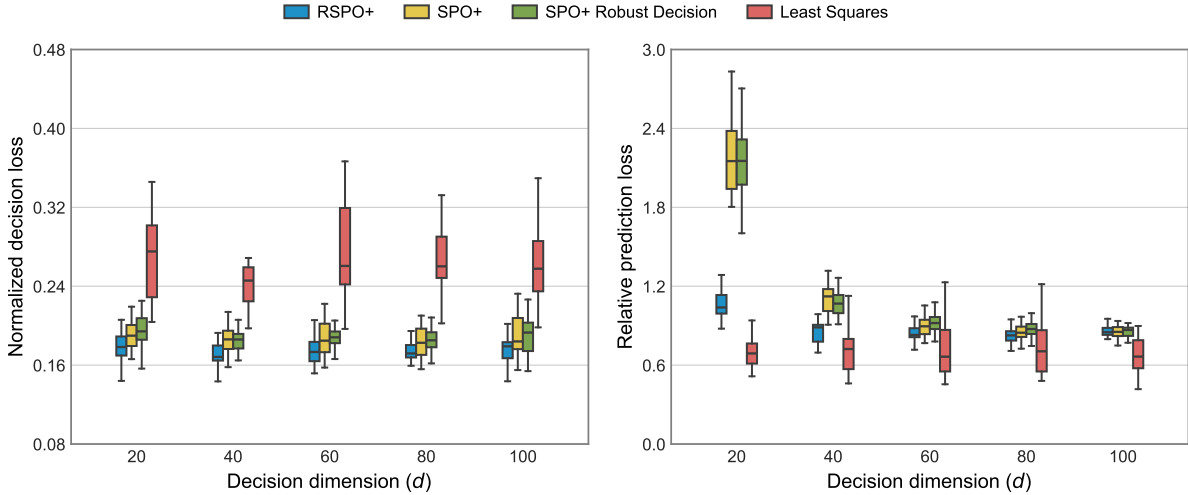


Figure EC.3 Normalized decision loss (left) and relative prediction loss (right) as d varies on quadratic-risk portfolio optimization instances, with $N = 80$, $p = 80$, $\text{deg} = 4$, and $\tau = 1$ fixed.

least squares has the largest decision loss. This advantage persists even as the relative prediction losses of $RSPO_+$, SPO_+ , and SPO_+ with robust decisions converge at the larger dimensions. Together, these results show that using the same robust decision map to define the training loss and generate deployed decisions continues to improve downstream decision quality under a smooth quadratic risk constraint, extending the empirical evidence beyond the polyhedral risk set used in the main experiment.

E.3 Implementation and Calibration Details

This subsection gives the risk-budget calibration for the main ℓ_1 -risk study and the implementation settings for gradient-based refinement.

Risk-budget calibration. For the main ℓ_1 -risk experiment, we calibrate the budget against the uniform portfolio $\mathbf{z}_{\text{unif}} = d^{-1}\mathbf{1}$. Defining its aggregate risk exposure as $\rho_{\text{base}} = \|\Sigma \mathbf{z}_{\text{unif}}\|_1$, we set $\beta = 2\rho_{\text{base}}$.

Gradient-based refinement. At each iteration, Algorithm 1 predicts the cost vector, solves the robust portfolio problem, identifies the active constraints, and updates the predictor using the Jacobian in Proposition 6. The ℓ_1 risk constraint is represented through its lifted epigraph, and the required Jacobian–vector products are obtained from the corresponding KKT system. We set the batch size to $B_s = 5$ and run $T = 50$ iterations, with backtracking factor $\rho_{\text{ls}} = 0.5$ and at most $K = 10$ trial steps per iteration. In place of the Armijo sufficient-decrease test of Algorithm 1, our implementation accepts a trial step as soon as it does not increase the batch objective by more than a numerical tolerance $\epsilon = 10^{-10}$, i.e., the acceptance rule with $c_{\text{ls}} = 0$ relaxed by ϵ to guard against floating-point error. The validation procedure jointly selects (γ, λ) and the base stepsize $\eta_0 \in \{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$ by minimizing validation $\mathcal{RSP}\mathcal{O}$ loss.