WHEN DOES UNCERTAINTY QUANTIFICATION MATTER IN PREDICT-THEN-OPTIMIZE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Predict-then-optimize (PtO) systems estimate uncertain quantities and feed the resulting predictive distribution into a downstream optimization problem. Better uncertainty quantification (UQ), however, improves decisions only when a distributional feature is both imperfectly recoverable and visible to the optimizer. We formalize this coupling for expected-cost stochastic optimization. A directed decision discrepancy controls excess risk exactly up to task normalization; its symmetric envelope is an oscillation distance that removes common objective shifts, and both are bounded by an integral probability semimetric. Uniform local continuity with tail control then connects these task-induced quantities to Wasserstein error without convexity, strong convexity, uniqueness, or continuity of the solution map. Our main theorem separates finite-sample estimation error and population projection misspecification from the downstream sensitivity to those errors. Moment couplings connect mean and covariance estimation to the risk bound, while invariances identify distributional directions that need not be estimated. Experiments on retail inventory, industry portfolios, and day-ahead electricity procurement vary the statistical and decision sides separately. The results show that UQ matters through their coupling, not as an intrinsic property of the prediction task or through distributional fidelity alone.

1 INTRODUCTION

Operational decisions are often made before demand, travel times, renewable generation, or asset returns are observed. A predict-then-optimize (PtO) pipeline first estimates their conditional distribution and then solves a downstream stochastic program (Donti et al., 2017; Kotary et al., 2021). When the optimizer uses expected costs, risk measures, or chance constraints, the relevant input is a predictive law rather than a single value. Dispersion, dependence, and tails may therefore alter resource allocation, feasibility, and risk exposure.

The central question is not whether one predictive distribution is statistically closer to the truth, but whether the improvement changes the relative ranking of candidate decisions. Likelihood, calibration, and proper scores evaluate a predictive law globally, whereas deployment is assessed through cost, regret, reliability, or risk-adjusted utility. A large distributional error can lie in a direction that the optimizer cancels, while a smaller error near a critical quantile or covariance direction can change the decision substantially. UQ importance therefore has two inseparable axes: the statistical cost of recovering a feature and the downstream response to errors in that feature.

We study a true conditional law p^*, an estimated law $\hat{p}$, and the decision $x^*(q)$ induced by any candidate law q. Our main insight is that the decision value of UQ is determined jointly by a statistical side and a decision side. The former becomes finite-sample error E_T plus population projection misspecification B under structured UQ; the latter is measured sharply by a directed discrepancy δ_f between normalized objectives and more coarsely by a local modulus ω_R. The symmetric envelope of δ_f is an oscillation semimetric, which ignores objective shifts shared by every decision and is tighter than the usual uniform objective distance. Thus neither a difficult estimation problem nor a sensitive optimizer alone makes UQ valuable: the estimation error must occur in a direction that changes relative objective values.

The analysis does not require convexity, strong convexity, a unique optimizer, or a continuous solution map. Additional growth is needed only to control the location of a solution rather than

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

its risk. For high-dimensional UQ, we introduce a structured projection and separate finite-sample estimation from population misspecification. A moment coupling translates mean and covariance errors into Wasserstein error; scale invariance and problem-specific sufficient statistics further remove irrelevant directions. A matching finite construction shows that the factor two incurred when replacing oscillation by a uniform objective metric cannot be improved.

Our contributions are: (i) a task-induced hierarchy from directed decision discrepancy to oscillation, integral probability, and locally controlled Wasserstein distances; (ii) an end-to-end theorem combining estimability, structural misspecification, and downstream sensitivity; (iii) explicit consequences for location–scatter models, precision weighting, critical quantiles, and dependence; and (iv) three temporally held-out real-data studies in inventory, portfolio allocation, and electricity procurement. Proofs, complete experimental specifications, and a controlled bound check appear in the appendix.

2 Related Work

Decision-focused learning differentiates or otherwise trains through an optimizer so that predictive errors are weighted by their downstream consequences (Donti et al., 2017; Wilder et al., 2019; Berthet et al., 2020; Shah et al., 2022; Schutte et al., 2024); see Kotary et al. (2021); Mandi et al. (2024) for surveys. Operational statistics and integrated estimation–optimization similarly couple statistical fitting and decisions (Liyanage & Shanthikumar, 2005; Chu et al., 2008; Ban & Rudin, 2019; Qi et al., 2025; Sadana et al., 2025). Decision-aware UQ studies calibration or uncertainty representations relative to downstream losses and utilities (Zhao et al., 2021; Kiyani et al., 2025; Cortes-Gomez et al., 2025; Schneider et al., 2026). We do not propose a new training algorithm. Instead, we characterize which distribution errors can affect deployment risk and provide quantities that can guide losses, evaluation metrics, and model complexity.

Quantitative stability in stochastic programming bounds changes in values and solution sets through probability metrics (Rachev & Römisch, 2002; Römisch, 2003; Pflug & Pichler, 2014). Wasserstein distributionally robust optimization changes the deployment objective to protect against an ambiguity set (Gao et al., 2024); here the predicted law is used directly, and we quantify the resulting risk. Statistically, empirical Wasserstein convergence suffers from ambient or intrinsic dimension (Fournier & Guillin, 2015), whereas means and structured scatter can be estimated parametrically. Dependence makes the tradeoff concrete: marginal forecasts may be insufficient when products of uncertain quantities enter the objective (Cameron et al., 2022), and the price of correlation quantifies the cost of ignoring dependence (Agrawal et al., 2012). Our framework combines these statistical and optimization effects in one risk decomposition.

The object controlled here differs from those in the three closest lines of work. Decision-focused learning changes the estimator or training loss, while our results apply after any probabilistic estimator has been chosen and diagnose which of its errors matter. Classical stochastic-program stability often controls optimal values or solution sets symmetrically; deployment excess risk is directed and can be zero even when both a probability metric and a solution distance are positive. Distributionally robust optimization optimizes against a neighborhood of laws, whereas our deployed optimizer minimizes under one estimated law. These distinctions motivate the directed normalization in Section 4: it retains exactly the comparison required for evaluating a fitted PtO decision without imposing a robust counterpart or differentiability of the optimization layer.

3 Problem Setup

Let $\theta \in \Theta \subseteq \mathbb{R}^d$ be the uncertain parameter, $\mathcal{X} \subseteq \mathbb{R}^m$ the decision set, and $p^* = \mathrm{Law}(\theta \mid \mathcal{I})$ the true conditional law given information $\mathcal{I}$. The scenario loss $f(x, \theta)$ is the cost of decision x after θ is realized. For $q \in \mathcal{P}(\Theta)$ define

$$v(q) = \min_{x \in \mathcal{X}} F(x, q), \qquad F(x, q) = \mathbb{E}_q[f(x, \theta)], \qquad X^*(q) = \arg\min_{x \in \mathcal{X}} F(x, q). \tag{1}$$

We fix an arbitrary selection $x^*(q) \in X^*(q)$. Its true risk and optimal risk are

$$\mathcal{R}(x) = \mathbb{E}_{p^*}[f(x, \theta)], \qquad \mathcal{R}^* = \min_{x \in \mathcal{X}} \mathcal{R}(x) = v(p^*). \tag{2}$$

The target is the deployment excess risk $\mathcal{R}(x^*(\hat{p})) - \mathcal{R}^*$. Risk-averse extensions require a discrepancy matched to the corresponding law functional; Proposition 4.4 gives one such distinction for CVaR.

Assumption 3.1 (Existence, uniform local continuity, and tails)**.** $\mathcal{X}$ is nonempty and $F(\cdot, q)$ is finite and attains its minimum for every law used below. There are increasing compact sets $K_R \uparrow \Theta$ and 1-Lipschitz truncations $T_R : \Theta \to K_R$ that fix K_R. For every R, a nondecreasing concave modulus ω_R, with $\omega_R(0) = 0$, satisfies

$$\sup_{x\in\mathcal{X}} |f(x,\theta) - f(x,\theta')| \le \omega_R(\|\theta - \theta'\|) \quad (\theta, \theta' \in K_R).$$

Every relevant law has a finite first moment, and $\tau_R(q) := \sup_x \mathbb{E}_q|f(x,\theta) - f(x, T_R\theta)| < \infty$ with $\tau_R(q) \to 0$ as $R \to \infty$.

This assumption permits unbounded Θ and $\mathcal{X}$ and imposes no regularity in x. Compact support removes τ_R; local uniform Hölder continuity gives $\omega_R(t) = L_R t^\alpha$, with the Lipschitz case at $\alpha = 1$.

4 METHODOLOGY

Our construction separates three questions that are often conflated in PtO. First, does a distributional perturbation change the relative objective values of feasible decisions? Second, can that decision-relevant perturbation be controlled by a probability metric available from statistical analysis? Third, does a structural model reduce estimation error without discarding a sensitive component of the true law? Together these questions implement the paper's central coupling: E_T and B describe what the predictive model fails to recover, while δ_f and ω_R describe how the downstream problem filters that failure. The discrepancies below answer the first question exactly, the local-modulus argument connects them to W_1, and the structured decomposition makes the statistical tradeoff explicit. Starting directly from a global probability distance can overstate decision error because common shifts of every objective value have no effect on optimization. Figure 1 summarizes the argument.

4.1 TASK-INDUCED DISTRIBUTION DISCREPANCIES

For $p, q \in \mathcal{P}(\Theta)$ let $\Delta_{p,q}(x) = F(x,p) - F(x,q)$ and define

$$\begin{aligned} d_f^{\uparrow}(p,q) &= \sup_x \Delta_{p,q}(x), & d_f(p,q) &= \sup_x |\Delta_{p,q}(x)|, \\ \mathrm{osc}_f(p,q) &= \sup_x \Delta_{p,q}(x) - \inf_x \Delta_{p,q}(x), & \delta_f(p\|q) &= \sup_x\{F(x,p) - v(p) - F(x,q) + v(q)\}. \end{aligned} \tag{3}$$

$\delta_f(p\|q)$ is the directed decision discrepancy from deployment law q to evaluation law p; osc_f is invariant to common objective shifts. Directly from the definitions,

$$\delta_f(p\|q) + \delta_f(q\|p) = \mathrm{osc}_f(p,q) = d_f^{\uparrow}(p,q) + d_f^{\uparrow}(q,p) \le 2d_f(p,q). \tag{4}$$

Both δ_f and osc_f satisfy the appropriate directed or symmetric triangle inequality. The four quantities retain progressively less task information. The directed discrepancy is aligned with deployment because only the orientation from the fitted law to the true law is needed. Symmetrization gives osc_f, which removes every decision-independent objective shift. The uniform objective metric d_f is easier to bound but restores such shifts and can therefore lose a factor of two. Finally, W_1 is task-agnostic but statistically interpretable. Consequently, W_1 should be viewed as an endpoint certificate rather than the definition of UQ quality for a downstream decision.

Lemma 4.1 (Local modulus and tail control)**.** *Under Assumption 3.1, for every $R > 0$,*

$$\begin{aligned} d_f(p,q) \le D_R(p,q) &:= \omega_R(W_1(p,q)) + \tau_R(p) + \tau_R(q), \\ 0 \le \delta_f(p\|q) \le \mathrm{osc}_f(p,q) &\le 2D_R(p,q). \end{aligned} \tag{5}$$

Consequently $d_f(p,q) \le D(p,q) := \inf_R D_R(p,q)$.

The proof couples p and q, truncates both variables to K_R, and applies Jensen's inequality to the concave modulus. Unlike a global Lipschitz premise, the bound records explicitly what happens outside the locally regular region.

Theorem 4.2 (Decision-risk stability)**.** *Under Assumption 3.1, for every $x^*(\hat{p}) \in X^*(\hat{p})$,*

$$0 \le \mathcal{R}(x^*(\hat{p})) - \mathcal{R}^* \le \delta_f(p^*\|\hat{p}) \le \mathrm{osc}_f(p^*, \hat{p}) \le 2d_f(\hat{p}, p^*) \le 2D(\hat{p}, p^*). \tag{6}$$

If $F(x_\varepsilon(\hat{p}), \hat{p}) \le v(\hat{p}) + \varepsilon$, the right side gains only $+\varepsilon$. The factor 2 in a bound expressed solely through d_f is unimprovable in general.

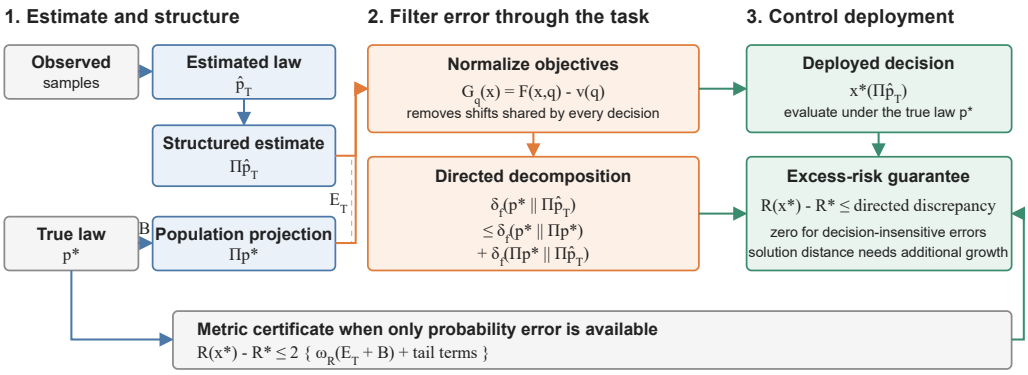

Figure 1: Methodology overview. The statistical side produces finite-sample error E_T and population misspecification B; the decision side filters them through δ_f or its local-modulus certificate ω_R. Deployment risk is governed by their coupling.

The first inequality evaluates the normalized-objective difference at the deployed optimizer; it therefore remains valid with nonunique and discontinuous solutions. If decision locations are needed, an additional growth condition $F(x,p) - v(p) \geq c_p \operatorname{dist}(x, X^*(p))^r$ yields

$$\operatorname{dist}(x^*(q), X^*(p)) \leq \{\delta_f(p\|q)/c_p\}^{1/r}. \tag{7}$$

Strong convexity is only one sufficient route to such growth and is not assumed for the risk theory.

4.2 Estimability and structured UQ

For location–scatter laws $p = \mathrm{Law}(\bar{\theta} + \Sigma^{1/2}Z)$ and $\hat{p} = \mathrm{Law}(\hat{\theta} + \hat{\Sigma}^{1/2}Z)$ with $\mathbb{E}Z = 0$ and $\mathrm{Cov}(Z) = I$,

$$W_1(\hat{p}, p) \leq \left\|\hat{\theta} - \bar{\theta}\right\| + m_Z \left\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\right\|_{\mathrm{op}}, \quad m_Z = \mathbb{E}\,\|Z\| \leq \sqrt{d}. \tag{8}$$

For positive definite scatter matrices, the second factor is at most $\left\|\hat{\Sigma} - \Sigma\right\|_{\mathrm{op}} / (\sqrt{\lambda_{\min}(\hat{\Sigma})} + \sqrt{\lambda_{\min}(\Sigma)})$; for positive semidefinite matrices it is at most $\left\|\hat{\Sigma} - \Sigma\right\|_{\mathrm{op}}^{1/2}$. Thus parameter estimation enters Theorem 4.2 through an explicit coupling.

The empirical law without structural restrictions satisfies the familiar dimension-dependent rate $\mathbb{E}W_1(\hat{p}_T, p^*) = O(T^{-1/d'})$, $d' = \max\{2, d\}$, for compactly supported classes (Fournier & Guillin, 2015). Because compactly supported laws are themselves sub-Gaussian, a tail assumption alone cannot yield a uniform parametric W_1 rate over an otherwise unrestricted distribution class; it instead controls the unbounded-support remainder and concentration. Rates such as $O(\sqrt{d/T})$ apply to means or suitably regular finite-dimensional models, not to arbitrary empirical laws in W_1. By contrast, for coordinatewise sub-Gaussian observations, diagonal sample variances obey, with probability at least $1 - \delta$,

$$\max_i |\hat{\sigma}_i^2 - \sigma_i^2| \leq C\sigma_{\max}^2 \left\{ \sqrt{\frac{\log(2d/\delta)}{T}} + \frac{\log(2d/\delta)}{T} \right\}. \tag{9}$$

For a diagonal location–scatter family, (8) then gives

$$W_1(\hat{p}_d, p_d) = O_p\left(a_T + m_Z \sigma_{\max} \left\{ \frac{\log(d/\delta)}{T} \right\}^{1/4} \right),$$

where $a_T = \left\|\hat{\theta} - \bar{\theta}\right\|$. The possible misspecification $W_1(p_d, p^*)$ must be handled separately.

Let $\mathcal{Q}$ be a structured family and let $\Pi : \mathcal{P}(\Theta) \to \mathcal{Q}$ be a projection that fixes $\mathcal{Q}$. The population projection Πp^* separates the law that could be recovered with infinite data under the structural restriction from the finite-sample estimate $\Pi \hat{p}_T$.

Theorem 4.3 (Estimability, misspecification, and sensitivity)**.** *Let $E_T = W_1(\Pi \hat{p}_T, \Pi p^*)$ and $B = W_1(\Pi p^*, p^*)$. Under Assumption 3.1, for every $R > 0$,*

$$0 \le \mathcal{R}(x^*(\Pi \hat{p}_T)) - \mathcal{R}^* \le \delta_f(p^* \| \Pi p^*) + \delta_f(\Pi p^* \| \Pi \hat{p}_T), \tag{10}$$

$$\mathcal{R}(x^*(\Pi \hat{p}_T)) - \mathcal{R}^* \le 2\{\omega_R(E_T + B) + \tau_R(\Pi \hat{p}_T) + \tau_R(p^*)\}. \tag{11}$$

Moreover: (i) excess risk is zero exactly when the selected deployed decision lies in $X^*(p^*)$, regardless of E_T; (ii) $p^* \in \mathcal{Q}$ implies $B = 0$, while population structural decision bias vanishes exactly when $x^*(\Pi p^*) \in X^*(p^*)$; and (iii) substituting (9) *gives the end-to-end diagonal location–scatter rate stated above inside ω_R.*

Equation (10) is the sharper characterization: the first term is structural and the second is estimative, but both are filtered by decision sensitivity. Equation (11) exposes the two axes in computable form: $E_T + B$ measures statistical approximation, while ω_R converts that error into objective perturbation. A two-decision construction in Appendix C attains $\mathrm{osc}_f = 2d_f$ and excess risk arbitrarily close to $2d_f$, matching the worst-case constant.

The theorem also provides a criterion for choosing model complexity. Consider candidate projections Π_k indexed by covariance rank, graphical sparsity, mixture size, or dependence complexity. Increasing k will typically decrease the population term $\delta_f(p^* \| \Pi_k p^*)$ but increase the finite-sample term $\delta_f(\Pi_k p^* \| \Pi_k \hat{p}_T)$. A likelihood criterion replaces both by global fit and may therefore select complexity in directions that the optimizer cancels. In contrast, validation can evaluate normalized objectives on a representative decision set, estimating the two directed terms without reconstructing the entire law. The metric bound remains useful when only parameter or probability-distance guarantees are available, but the directed form explains why the statistically closest model need not have the lowest deployed risk.

This coupling is not a literal product: sensitivity acts on the direction of statistical error through the task-induced discrepancy and the local modulus. It nevertheless yields two exact zero mechanisms. A model can be statistically imperfect yet decision-exact because its error is invariant to the optimizer; conversely, a highly sensitive feature causes no loss when it is recovered exactly. Away from these endpoints, E_T identifies what more data can improve, B what requires changing the structural family, and δ_f which of those improvements can affect deployment.

Optimization invariances further reduce what must be estimated. If x^* is unchanged under $\hat{\Sigma} \mapsto c\hat{\Sigma}$, risk factors through the quotient $\mathbb{S}^d_+/\mathbb{R}_+$ and depends on relative scatter, for example centered log-variance errors. In precision-weighted least squares, the common log-scale direction has zero directional derivative. Appendix A.3 gives the explicit risk decomposition and correlation-dependent bound.

4.3 Concrete implications for moment-based decisions

The abstract decomposition becomes especially transparent when the expected objective has a finite-dimensional sufficient statistic. Suppose $f(x; y_1, y_2) = \sum_{i=1}^d y_{i1} y_{i2} f_i(x) + C$. Then

$$F(x, p) = \sum_{i=1}^{d} \{\mathbb{E}_p[y_{i1}]\mathbb{E}_p[y_{i2}] + \mathrm{Cov}_p(y_{i1}, y_{i2})\} f_i(x) + C. \tag{12}$$

Thus two laws with the same displayed moments induce the same optimizer even if their W_1 distance is positive. Conversely, a product-law projection preserves the marginal means but deletes the covariance contribution. Its population structural error is therefore decision-relevant exactly through the variation across x of $\sum_i \mathrm{Cov}_{p^*}(y_{i1}, y_{i2}) f_i(x)$. This example distinguishes a statistically large but irrelevant distribution error from a low-dimensional error that changes objective rankings.

A second implication concerns precision-weighted quadratic decisions. For a diagonal variance estimate M and a signal in $\mathrm{col}(S)$, let

$$P_M = S(S^\top M^{-1} S)^{-1} S^\top M^{-1}, \qquad \mathcal{R}(M) = \mathrm{tr}(P_M W P_M^\top).$$

Since $P_{cM} = P_M$ for every $c > 0$, common variance scale is exactly irrelevant. Writing $\eta_i = \log \hat{\sigma}_i^2 - \log \sigma_i^2$ and $\tilde{\eta}_i = \eta_i - d^{-1} \sum_j \eta_j$, the estimation contribution satisfies

$$|\mathcal{R}(\hat{W}_d) - \mathcal{R}(W_d)| \leq C_1' \max_i |\tilde{\eta}_i|, \tag{13}$$

where the explicit pathwise constant C_1' is given in Appendix A.3. The remaining gap $\mathcal{R}(W_d) - \mathcal{R}(W)$ is a structural correlation term. Hence absolute variance error is not the appropriate diagnostic: relative scale estimation controls the finite-sample term, while omitted correlation controls misspecification. This is the mechanism examined by the portfolio experiment.

4.4 Decision-relevant distributional information

Suppose $x^*(q) = h(\mathcal{T}(q))$ for a statistic $\mathcal{T}$. Any perturbation preserving $\mathcal{T}$ is decision-irrelevant even when W_1 is nonzero. In the newsvendor problem, $f(x, \theta) = c_u(\theta - x)^+ + c_o(x - \theta)^+$ and $x^*(q) = F_q^{-1}(\alpha)$ with $\alpha = c_u/(c_u + c_o)$. If the true density is bounded below by $\underline{f}$ between the true and predicted critical quantiles,

$$|x^*(\hat{p}) - x^*(p^*)| \leq \underline{f}^{-1} \left\| F_{\hat{p}} - F_{p^*} \right\|_\infty . \tag{14}$$

Thus local CDF error is more informative than global fidelity, while Theorem 4.2 still supplies the distribution-free bound $2 \max\{c_u, c_o\} W_1(\hat{p}, p^*)$.

Proposition 4.4 (Copula sensitivity)**.** *If $C(x, D) = \sum_i C_i(x, D_i)$ and the feasible set does not depend on D, then $\mathbb{E}_q C(x, D)$ depends only on the univariate marginals of q. Laws with the same marginals induce the same expected objective and optimizer set. In contrast, CVaR of total cost can change with the copula, including both its value and its optimizer.*

This proposition explains why the risk-neutral curves in Figures 2(a) and 4(a) coincide, whereas total-cost CVaR exposes cross-item and cross-hour dependence.

5 Experiments

All three studies estimate a predictive law from past observations, solve the induced optimization problem, and evaluate only on subsequent observations. They are organized as two-axis tests of Theorem 4.3: distributional structure changes the statistical terms E_T and B, while objectives, constraints, and risk aversion change sensitivity through δ_f and ω_R. Table 1 summarizes the feature varied in each study. Models share the same temporal split and downstream objective within each comparison, so differences cannot be attributed to test-data reuse. Full specifications are in Appendix B.

Retail inventory. Using roughly 1.07 million UCI Online Retail II transactions (Chen, 2012), we select 30 active SKUs, estimate joint weekly demand from 53 training weeks, and evaluate on the next 53 weeks. Joint empirical, independently shuffled, and comonotone scenarios have identical empirical marginals and differ only in dependence. Their risk-neutral objectives and decisions coincide. With total-cost $\mathrm{CVaR}_{0.8}$, allocations separate: the maximum ℓ_1 shifts relative to the joint law are 12.3% and 8.4% for shuffled and comonotone scenarios. No dependence model dominates at every capacity.

This comparison holds the statistical perturbation fixed and switches its decision sensitivity. The three laws can be separated by a global probability metric because their copulas differ, yet Proposition 4.4 makes $\delta_f = 0$ for expected separable costs. The zero-gap annotation in Figure 2(a) verifies that a statistically visible shift need not matter. Under aggregate-cost CVaR, the same copula error changes which joint shortages enter the upper tail, making δ_f nonzero. The heatmap shows that its risk consequence also depends on capacity, while the lollipop plot in panel (b) confirms that the induced allocations have moved. The experiment therefore isolates the decision side of the coupling while holding all marginal estimates fixed; its capacity-dependent ordering supports no unconditional ranking of dependence models.

Industry portfolios. We use daily returns of 30 Fama–French industry portfolios (French, 2026), estimate moments on a rolling 252-day window, rebalance every 21 days, and compare full, diagonal, diagonal-shrinkage, and five-factor covariance estimators. At low risk aversion, realized utility is similar. At $\gamma = 30$, annualized volatility is 18.8% under the diagonal model versus approximately

Table 1: Real-data protocols and the distributional information tested downstream. All evaluation observations occur strictly after their corresponding estimation windows.

Data and split	Decision	Distributional variation	Held-out criterion
Retail, 53 train / 53 test weeks	30-SKU capacity allocation	Copula with fixed SKU marginals	Expected and $\mathrm{CVaR}_{0.8}$ inventory cost
Fama–French, rolling 252 / 21 days	Long-only industry weights	Full, diagonal, shrinkage, and rank-five covariance	Risk–return position and certainty equivalent
OPSD, rolling 60 days / next day	24-hour ramp-constrained purchase	Temporal copula with fixed hourly marginals	Cost–reliability position and day-level regret

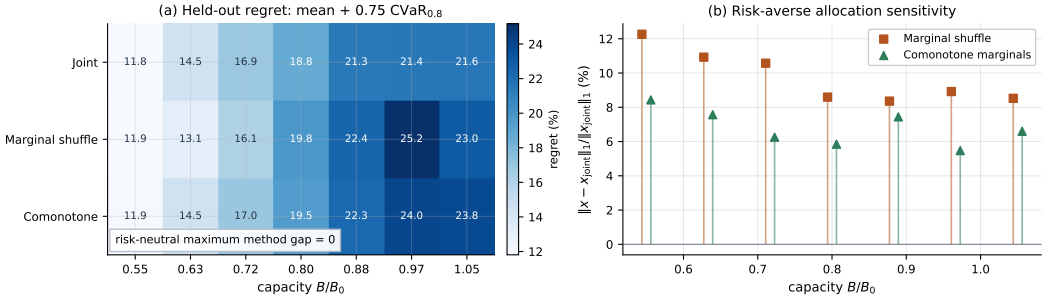

Figure 2: Capacitated multi-item inventory on UCI Online Retail II. The predictive laws share SKU marginals and differ only in dependence. (a) Annotated heatmap of held-out risk-averse regret across dependence models and capacities; the inset records exact risk-neutral invariance. (b) Allocation shifts relative to the joint empirical law.

16.0% under full and five-factor covariance; the diagonal model's realized certainty equivalent is about 12 percentage points lower. Higher downstream curvature amplifies covariance misspecification, while low rank preserves dominant shared-risk directions.

This experiment varies both axes of Theorem 4.3. Diagonal and low-rank structures reduce finite-sample error E_T, but only the diagonal projection creates a large B in dominant shared-risk directions. Varying γ changes downstream sensitivity without changing the data or estimators. In Figure 3(a), increasing marker size traces this sensitivity change directly in the realized risk–return plane. Panel (b) shows small low-γ utility gaps but a pronounced diagonal-model penalty at high γ: the same E_T–B tradeoff becomes consequential only when portfolio variance receives sufficient decision weight. The result supports the coupling, not a task-independent ranking of covariance estimators.

Day-ahead electricity procurement. Using OPSD DK1 day-ahead forecasts, realized load, and prices (Open Power System Data, 2020), we form 24-hour error scenarios from the previous 60 valid days and solve ramp-constrained procurement. Joint daily errors, hourwise shuffling, and comonotone errors share hourly marginals. Risk-neutral plans coincide; adding daily total-cost $\mathrm{CVaR}_{0.9}$ changes procurement. The comonotone model buys more and reduces unmet load, but incurs higher realized cost. Dependence therefore changes an operational cost–reliability tradeoff rather than only a forecast score.

The risk-neutral comparison again holds a copula error fixed while testing whether the optimizer senses it. Ramp constraints couple adjacent decisions, but expected scenario cost remains additive across hours, so $\delta_f = 0$ for laws with common hourly marginals; the inset in Figure 4(a) records this zero-sensitivity result. Daily CVaR makes co-occurring hourly errors decision-relevant. Panel (a) shows that comonotone scenarios move the plan toward lower unmet load at higher cost, while panel (b) reveals the distribution of this dependence effect over test days rather than only its mean. Thus identical marginal estimability can lead to identical or different plans solely because the downstream functional changes sensitivity to temporal dependence.

Controlled bound diagnostic.

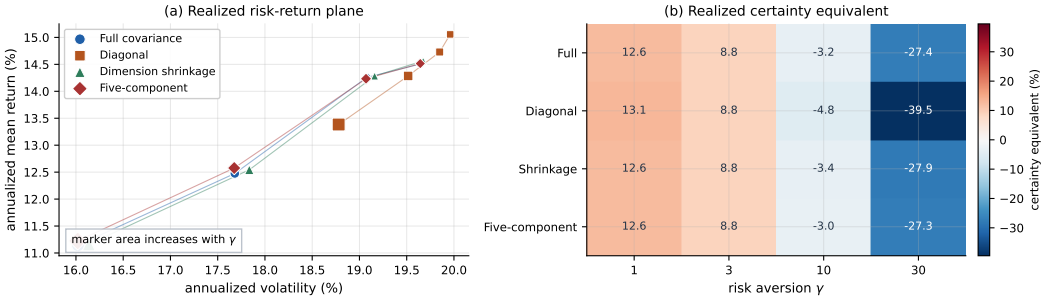

Figure 3: Rolling allocation across 30 Fama–French industries. (a) Held-out risk–return positions, with marker area increasing in risk aversion. (b) Certainty-equivalent heatmap across covariance structures and risk-aversion levels. Full and low-rank covariance outperform the diagonal model at high risk aversion.

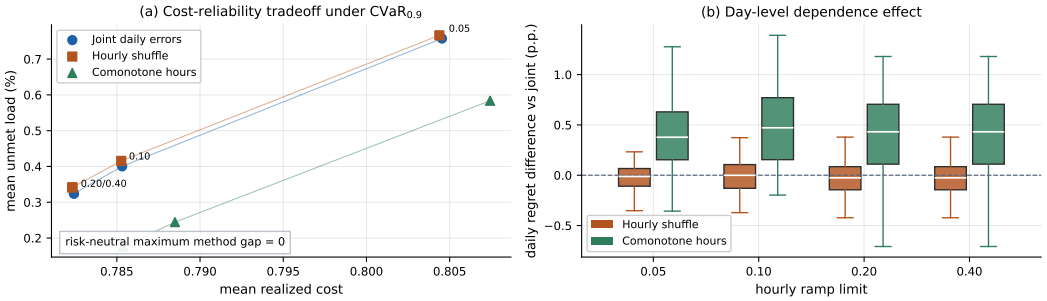

Figure 4: Rolling day-ahead procurement on OPSD DK1. (a) Risk-averse cost–reliability positions across ramp limits; the inset records exact risk-neutral invariance. (b) Test-day distributions of regret differences relative to the joint-error model.

The real-data studies identify sensitive distributional directions, but they do not reveal how close a universal certificate is to realized excess risk. We therefore complement them with a model in which every quantity is known. For $f_\mu(x,\theta) = \mu x^2/2 - \theta x$ on $[-1,1]$, let $p^* = \mathrm{Unif}[-h,h]$ and $\hat{p} = \mathrm{Unif}[\delta - h, \delta + h]$. Before clipping becomes active,

$$W_1(\hat{p},p^*) = \delta, \qquad |x^*(\hat{p}) - x^*(p^*)| = \frac{\delta}{\mu}, \qquad \mathcal{R}(x^*(\hat{p})) - \mathcal{R}^* = \frac{\delta^2}{2\mu}. \tag{15}$$

Figure 5 compares these exact quantities with the distribution-free risk certificate $2W_1$ and the growth-based solution certificate $2\sqrt{W_1/\mu}$. The excess risk is much smaller than the linear upper bound near the truth because this smooth instance has a second-order risk response. Smaller μ nevertheless increases both displacement and risk, confirming that an identical distribution error can have different consequences under different objective geometry.

Unlike the real-data studies, this construction directly calibrates the decision side of Theorem 4.2. Panel (a) holds W_1 fixed across curvature levels and shows that smaller μ increases displacement and risk: an identical statistical error has different consequences under different sensitivity. The universal linear certificate is conservative locally because realized excess risk is quadratic in δ, although the theorem must cover nonsmooth worst cases. Panel (b) holds the probability-error scale comparable across directions: a location shift changes the optimizer, whereas a mean-preserving scale change has positive W_1 but zero task-induced discrepancy. Together the panels show why neither probability distance nor objective geometry alone determines UQ value. Appendix B gives the full construction.

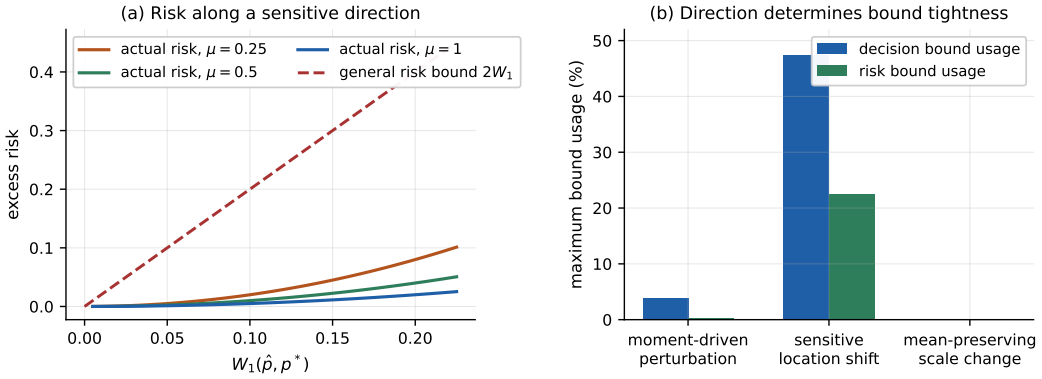

Figure 5: Controlled stability diagnostic. (a) Exact excess risk and the universal certificate across three objective curvatures on a common W_1 range. (b) Maximum utilization of risk and solution bounds for moment-driven, location, and decision-irrelevant scale perturbations.

6 Discussion and Conclusion

The theory has three boundaries. First, Theorem 4.3 is stated for expected objectives; chance constraints and nonlinear law-invariant risks require discrepancies defined directly for their law functionals. Proposition 4.4 identifies this distinction for CVaR, but a complete risk-measure extension must also account for estimation of the tail functional. Second, local continuity must be uniform over decisions and paired with tail control. Heavy-tailed settings may therefore require robust law estimators or problem-specific truncation. Third, risk stability does not imply solution stability without growth such as (7). Flat objectives can admit distant decisions with nearly identical risk, which is benign for cost but potentially important when decision interpretability or operational switching costs matter.

The practical prescription follows the two axes. First diagnose δ_f or ω_R to identify which distributional directions alter relative decision quality. Then choose the least complex predictive family that controls E_T without introducing B in those directions, and evaluate it through deployment risk or a justified task-induced surrogate. UQ matters when a feature is imperfectly recovered and the optimizer responds to that particular error; either statistical exactness or decision invariance makes additional UQ immaterial.

For model evaluation, fit each candidate law on training data and compare its induced decisions on an independent validation set after subtracting a common objective reference. This finite-decision analogue of δ_f tests whether reductions in E_T or B improve deployment rather than global fit alone. A zero gap does not establish equality of predictive laws; it identifies operational equivalence for the tested objective. Because constraints, penalties, and risk aversion change the sensitive subspace, model complexity must be selected jointly with the intended operating regime.

Several extensions follow naturally. Conditional PtO requires controlling the task-induced discrepancy uniformly or on average over covariates, rather than for one fixed conditional law. Data-driven selection of the structured family $\mathcal{Q}$ should balance an estimable sensitivity-weighted bias against variance, not select covariance rank or dependence complexity solely by likelihood. Finally, the directed discrepancy suggests validation losses that compare normalized objectives on a finite decision set. Developing statistically efficient estimators of this discrepancy would turn the present characterization into a model-selection procedure.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX restructuring, and code-level checks of figures and experiments. They were not used as an autonomous source of empirical data or as a substitute for mathematical verification. The authors reviewed all AI-assisted text, checked the stated derivations and proofs, reran the experimental code, and inspected the generated figures and compiled manuscript. The authors take responsibility for all claims, results, and artifacts in the submission.

REPRODUCIBILITY STATEMENT

Appendix B specifies data filtering, temporal splits, optimization models, hyperparameters, and evaluation metrics. Appendix C provides proofs of all theoretical claims. The supplementary code regenerates every figure with fixed random seeds where random shuffling is used.

REFERENCES

Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Price of correlations in stochastic optimization. *Operations Research*, 60(1):150–162, 2012.

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

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9508–9519, 2020.

Peter Bloomfield and Geoffrey S. Watson. The inefficiency of least squares. *Biometrika*, 62(1): 121–128, 1975.

Chris Cameron, Jason Hartford, Taylor Lundy, and Kevin Leyton-Brown. The perils of learning before optimizing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3708–3715, 2022.

Daqing Chen. Online retail ii. UCI Machine Learning Repository, 2012.

Leon Yang Chu, J. George Shanthikumar, and Zuo-Jun Max Shen. Solving operational statistics via a bayesian analysis. *Operations Research Letters*, 36(1):110–116, 2008.

Sebastian Cortes-Gomez, Camilo Patino, Youngseog Byun, Shizhe Wu, et al. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *International Conference on Learning Representations*, 2025.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015.

Kenneth R. French. Data library: 30 industry portfolios, daily value-weighted returns. Tuck School of Business, Dartmouth College, 2026.

Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

Shahab Kiyani, George Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 30943–30965, 2025.

Martin Knott. On the minimum efficiency of least squares. *Biometrika*, 62(1):129–132, 1975.

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

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

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

Open Power System Data. Time series: Load, wind and solar, prices in hourly resolution, 2020. Version 2020-10-06.

Georg Ch. Pflug and Alois Pichler. *Multistage Stochastic Optimization*. Springer, Cham, 2014.

Meng Qi, Paul Grigas, and Zuo-Jun Max Shen. Integrated conditional estimation-optimization. *Operations Research*, 74(3):1604–1625, 2025.

Svetlozar T. Rachev and Werner Römisch. Quantitative stability in stochastic programming: The method of probability metrics. *Mathematics of Operations Research*, 27(4):792–818, 2002.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Werner Römisch. Stability of stochastic programming problems. In Andrzej Ruszczyński and Alexander Shapiro (eds.), *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pp. 483–554. Elsevier, 2003.

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

Alexander Schneider, Thomas Rochussen, Jonathan Stiller, et al. Decision-aligned evaluation of uncertainty quantification, 2026.

Niels Schutte, Krzysztof Postek, and Neil Yorke-Smith. Robust losses for decision-focused learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4868–4875, 2024.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-focused learning without decision-making: Learning locally optimized decision losses. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1320–1332, 2022.

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

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

A Additional Mathematical Details

A.1 Optimal values and solution sets

The one-sided optimal-value bounds are

$$v(q) - v(p) \le d_f^{\uparrow}(q,p), \qquad v(p) - v(q) \le d_f^{\uparrow}(p,q),$$

and therefore $|v(p) - v(q)| \le d_f(p,q) \le D(p,q)$. A solution-set refinement of Theorem 4.2 is

$$F(x^*(q),p) - v(p) \le \sup_{x \in X^*(q)} \Delta_{p,q}(x) - \inf_{x \in X^*(p)} \Delta_{p,q}(x) \le \mathrm{osc}_f(p,q).$$

It can replace δ_f in (7) when both optimizer sets are available.

A.2 Cross moments as sufficient statistics

Suppose

$$f(x; y_1, y_2) = \sum_{i=1}^{d} y_{i1} y_{i2} f_i(x) + C.$$

Then

$$F(x, p) = \sum_{i=1}^{d} \{\mathbb{E}_p[y_{i1}]\mathbb{E}_p[y_{i2}] + \mathrm{Cov}_p(y_{i1}, y_{i2})\} f_i(x) + C.$$

Hence the first moments and cross-covariances are sufficient for the objective. A product-law projection that preserves marginals removes exactly $\Delta(x) = \sum_i \mathrm{Cov}_{p^*}(y_{i1}, y_{i2}) f_i(x)$, making the structural bias the image of this covariance term through the argmin map.

A.3 Precision-weighted quadratic decisions

Consider

$$x^*(q) = \operatorname*{arg\,min}_{x \in \{Sb : b \in \mathbb{R}^m\}} \sum_{i=1}^{n} \frac{(x_i - \hat{\theta}_i)^2}{\hat{\sigma}_i^2},$$

where S has full column rank, $\theta = Sb_0$, $\hat{\theta} = \theta + \epsilon$, $\mathbb{E}\epsilon = 0$, and $\mathrm{Cov}(\epsilon) = W$. Let $W_d = \mathrm{diag}(W)$, $\hat{W}_d = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2)$, $P_M = S(S^\top M^{-1} S)^{-1} S^\top M^{-1}$, and $\mathcal{R}(M) = \mathrm{tr}(P_M W P_M^\top)$. Then

$$\mathcal{R}(\hat{W}_d) - \mathcal{R}(W) = [\mathcal{R}(\hat{W}_d) - \mathcal{R}(W_d)] + [\mathcal{R}(W_d) - \mathcal{R}(W)].$$

Set $\eta_i = \log \hat{\sigma}_i^2 - \log \sigma_i^2$, $\tilde{\eta}_i = \eta_i - n^{-1} \sum_j \eta_j$, and $M(t) = \mathrm{diag}(\sigma_i^2 e^{t\tilde{\eta}_i})$. The scale-consistent estimation bound is

$$|\mathcal{R}(\hat{W}_d) - \mathcal{R}(W_d)| \le C_1' \max_i |\tilde{\eta}_i|, \tag{16}$$

where

$$C_1' = \sup_{0 \le t \le 1} \sum_i \left| 2\,\mathrm{tr}\{P_{M(t)} E_i (P_{M(t)} - I) W P_{M(t)}^\top\} \right|, \quad E_i = e_i e_i^\top.$$

Writing $\Phi = W_d^{-1/2} W W_d^{-1/2}$ and $\kappa = \lambda_{\max}(\Phi)/\lambda_{\min}(\Phi)$, the correlation-driven structural term satisfies

$$0 \le \mathcal{R}(W_d) - \mathcal{R}(W) \le \frac{(\kappa - 1)^2}{4\kappa} \mathcal{R}(W). \tag{17}$$

B Experimental Details

B.1 Retail inventory

We remove cancellations, missing descriptions, nonpositive quantities and prices, and nonstandard stock codes from the two Online Retail II worksheets. Transactions are aggregated by week and SKU. Selection uses only the training period: products must be active in at least 40 weeks, after which the 30 most active and highest-volume SKUs are retained. The split date is December 6, 2010, giving 53 training and 53 test weeks; missing SKU–week pairs are zero-filled and each series is normalized by its training mean.

For weekly demand D and replenishment x, we solve

$$\min_{x \ge 0,\, \sum_i x_i \le B} \mathbb{E}_q C(x, D) + \lambda\, \mathrm{CVaR}_{0.8,q}(C(x, D)), \quad C(x, D) = \sum_i \{3(D_i - x_i)^+ + (x_i - D_i)^+\},$$

with $\lambda \in \{0, 0.75\}$ and $B/B_0 \in [0.55, 1.05]$, where B_0 is total mean normalized demand. The finite-scenario CVaR problem uses the Rockafellar–Uryasev linear formulation (Rockafellar & Uryasev, 2000). Joint scenarios preserve weekly vectors; shuffled scenarios independently permute each SKU; comonotone scenarios sort and rank-match every column. Every solution is reevaluated on the 53 joint test weeks. Since inventory states are unavailable, sales proxy latent demand and the study is semi-synthetic.

B.2 FAMA–FRENCH PORTFOLIOS

The sample contains 6,539 complete trading days from January 3, 2000 through December 31, 2025. At each rebalance date, annualized mean $\hat{\mu}$ and covariance $\hat{\Sigma}$ are estimated from the previous 252 observations and

$$\min_{w\geq 0,\, \mathbf{1}^\top w=1,\, w_i\leq 0.1} -\hat{\mu}^\top w + \frac{\gamma}{2} w^\top \hat{\Sigma} w, \quad \gamma \in \{1, 3, 10, 30\},$$

is solved. Weights are held for the next 21 days. Models use the full sample covariance, its diagonal, $(1-a)\hat{\Sigma} + a\,\mathrm{diag}(\hat{\Sigma})$ with $a = \min\{1, 30/252\}$, or the top five principal components plus a diagonal residual. A $10^{-8}I$ ridge is added only for numerical stability. The held-out sequence has 6,279 daily observations. Metrics are annualized mean and volatility, certainty equivalent $\bar{r} - \gamma s^2/2$, and half-ℓ_1 turnover.

B.3 OPSD PROCUREMENT

We retain DK1 days with 24 complete hours and hourly relative forecast errors below 25%, normalizing load by the day's mean forecast. Test dates run from January 2018 through June 2020 at 14-day spacing, producing 65 tests. Each test uses only the preceding 60 valid daily error trajectories. Prices are truncated at zero and divided by 50 to place cost components on comparable scales.

For purchase x_t, load D_t, normalized price p_t, and ramp r, we solve

$$\min_{0\leq x_t\leq 2} \mathbb{E}_q C(x, D; p) + \lambda\, \mathrm{CVaR}_{0.9,q}(C(x, D; p)) \quad \text{s.t. } |x_t - x_{t-1}| \leq r,$$

where $r \in \{0.05, 0.10, 0.20, 0.40\}$, $\lambda \in \{0, 0.75\}$, and

$$C(x, D; p) = \frac{1}{24}\sum_{t=1}^{24} [p_t x_t + 3(D_t - x_t)^+ + 0.5(x_t - D_t)^+].$$

Joint scenarios preserve daily trajectories; shuffled scenarios permute errors separately by hour; comonotone scenarios sort each hour and match ranks. The public data do not identify true shortage and overpurchase penalties, so these coefficients define a sensitivity study rather than an estimate of actual DK1 dispatch cost.

B.4 CONTROLLED STABILITY CHECK

For $f_\mu(x, \theta) = \mu x^2/2 - \theta x$, $x \in [-1, 1]$, let $p^* = \mathrm{Unif}[-h, h]$ and $\hat{p} = \mathrm{Unif}[\delta - h, \delta + h]$ with $h = 0.02$, $0 < \delta \leq 0.225$, and $\mu \in \{0.25, 0.5, 1\}$. Then

$$W_1 = \delta, \quad |x^*(\hat{p}) - x^*(p^*)| = \delta/\mu, \quad \mathcal{R}(x^*(\hat{p})) - \mathcal{R}^* = \delta^2/(2\mu).$$

The universal risk bound is $2W_1$ and the quadratic-growth solution bound is $2\sqrt{W_1/\mu}$. A decision-irrelevant control changes only the scale of a centered uniform law, giving positive W_1 but unchanged decisions. Figure 5 reports the corresponding exact curves and bound utilization in the main paper.

C PROOFS

Proposition C.1 (Tightness of the d_f constant)**.** *For every $\delta > 0$ and $0 < \varepsilon < \delta$, there is a finite decision problem and laws p, q such that $d_f(p, q) = \delta$, $\mathrm{osc}_f(p, q) = 2\delta$, and $F(x^*(q), p) - v(p) = 2\delta - \varepsilon$.*

Proof. Take decisions $\{a, b\}$ and two scenarios, with p and q concentrated on different scenarios. Define $(F(a, p), F(b, p)) = (0, 2\delta - \varepsilon)$ and $(F(a, q), F(b, q)) = (\delta, \delta - \varepsilon)$. Then $x^*(p) = a$, $x^*(q) = b$, and $\Delta_{p,q}(a) = -\delta$, $\Delta_{p,q}(b) = \delta$. Hence $d_f = \delta$, $\mathrm{osc}_f = 2\delta$, and deployment under q incurs excess risk $2\delta - \varepsilon$. Letting $\varepsilon \downarrow 0$ proves tightness. □

C.1 Proof of Lemma 4.1

Fix x and write $f_R(x,\theta) = f(x, T_R\theta)$. By the truncation definition,

$$|F(x,p) - F(x,q)| \le \tau_R(p) + \tau_R(q) + |\mathbb{E}_p f_R(x,\theta) - \mathbb{E}_q f_R(x,\theta)|.$$

For a W_1-optimal coupling (U,V), the 1-Lipschitz property of T_R, the common modulus on K_R, and concavity give

$$|\mathbb{E}f(x,T_RU) - \mathbb{E}f(x,T_RV)| \le \mathbb{E}\omega_R(\|T_RU - T_RV\|) \le \omega_R(\mathbb{E}\,\|U-V\|) = \omega_R(W_1(p,q)).$$

Taking the supremum over x proves the bound on d_f. With $\bar{F}_p = F(\cdot,p) - v(p)$,

$$\delta_f(p\|q) + \delta_f(q\|p) = \sup_x(\bar{F}_p - \bar{F}_q) + \sup_x(\bar{F}_q - \bar{F}_p) = \mathrm{osc}_f(p,q).$$

The remaining inequalities follow from (4); taking the infimum over R proves the final claim.

C.2 Proof of Theorem 4.2

Let $\hat{x} = x^*(\hat{p})$ and $\bar{F}_p = F(\cdot,p) - v(p)$. Since $\bar{F}_{\hat{p}}(\hat{x}) = 0$,

$$\mathcal{R}(\hat{x}) - \mathcal{R}^* = \bar{F}_{p^*}(\hat{x}) = \bar{F}_{p^*}(\hat{x}) - \bar{F}_{\hat{p}}(\hat{x}) \le \delta_f(p^*\|\hat{p}).$$

Nonnegativity follows from optimality under p^*, and Lemma 4.1 plus (4) yields the chain. For an ε-optimal solution, $\bar{F}_{\hat{p}}(x_\varepsilon) \le \varepsilon$, which adds ε. Proposition C.1 establishes the last statement.

C.3 Proof of the moment coupling

Couple the laws using the same Z: $U = \hat{\theta} + \hat{\Sigma}^{1/2}Z$ and $V = \bar{\theta} + \Sigma^{1/2}Z$. Then

$$W_1(\hat{p},p) \le \mathbb{E}\,\|U-V\| \le \left\|\hat{\theta} - \bar{\theta}\right\| + \left\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\right\|_{\mathrm{op}} \mathbb{E}\,\|Z\|\,.$$

Jensen gives $m_Z \le \sqrt{\mathbb{E}\,\|Z\|^2} = \sqrt{d}$. For positive definite matrices, set $A = \hat{\Sigma}^{1/2}$, $B = \Sigma^{1/2}$, and $X = A - B$. The Sylvester equation $AX + XB = \hat{\Sigma} - \Sigma$ and its integral representation yield

$$\|X\|_{\mathrm{op}} \le \frac{\left\|\hat{\Sigma} - \Sigma\right\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}(\hat{\Sigma})} + \sqrt{\lambda_{\min}(\Sigma)}}.$$

The positive semidefinite case follows from the standard operator Hölder bound $\left\|A^{1/2} - B^{1/2}\right\|_{\mathrm{op}} \le \|A - B\|_{\mathrm{op}}^{1/2}$.

C.4 Proof of Theorem 4.3

Apply Theorem 4.2 with deployment law $\Pi\hat{p}_T$ and then the directed triangle inequality:

$$\mathcal{R}(x^*(\Pi\hat{p}_T)) - \mathcal{R}^* \le \delta_f(p^*\|\Pi\hat{p}_T) \le \delta_f(p^*\|\Pi p^*) + \delta_f(\Pi p^*\|\Pi\hat{p}_T).$$

The Wasserstein triangle inequality gives $W_1(\Pi\hat{p}_T, p^*) \le E_T + B$. Applying Lemma 4.1 proves (11). Excess risk vanishes exactly when the selected decision belongs to $X^*(p^*)$. Since Π fixes $\mathcal{Q}$, $p^* \in \mathcal{Q}$ implies $\Pi p^* = p^*$ and $B = 0$; the more general zero-bias criterion follows from the same optimizer-set condition. The stated rate follows by substituting (9) into the moment coupling.

C.5 Proof of Proposition 4.4

Linearity gives $\mathbb{E}_q C(x,D) = \sum_i \mathbb{E}_{q_i} C_i(x,D_i)$, where q_i is the ith marginal. Thus equal marginals yield equal objectives for every feasible x. For CVaR, let D_1, D_2 have Bernoulli$(1/2)$ marginals. Under q^+, (D_1,D_2) is equally likely to be $(0,0)$ or $(1,1)$; under q^- it is equally likely to be $(0,1)$ or $(1,0)$. For decisions A, B, let $C_i(A,D_i) = D_i$ and $C_i(B,D_i) = 3/4$. Then $\mathrm{CVaR}_{1/2,q^+}(C(A,D)) = 2 > 3/2$, whereas $\mathrm{CVaR}_{1/2,q^-}(C(A,D)) = 1 < 3/2$. The optimal decision is B under q^+ and A under q^-.

C.6 PROOFS FOR PRECISION WEIGHTING AND THE NEWSVENDOR

Because $\theta \in \mathrm{col}(S)$, $P_M\theta = \theta$ and $\mathbb{E}\left\|P_M\hat{\theta} - \theta\right\|^2 = \mathrm{tr}(P_M W P_M^\top)$. For $M = \mathrm{diag}(e^{\zeta_1}, \ldots, e^{\zeta_n})$,

$$\frac{\partial P_M}{\partial \zeta_i} = P_M E_i (P_M - I), \qquad \frac{\partial \mathcal{R}(M)}{\partial \zeta_i} = 2\,\mathrm{tr}\{P_M E_i (P_M - I) W P_M^\top\}.$$

Since $P_M^2 = P_M$, the gradient components sum to zero, proving common-scale invariance. Integrating along $M(t)$ proves (16). The lower bound in (17) is Gauss–Markov optimality; whitening by $W_d^{-1/2}$ and applying the Bloomfield–Watson–Knott inequality (Bloomfield & Watson, 1975; Knott, 1975) gives the upper bound.

For the newsvendor, a subgradient of $F(x, q)$ is a positive multiple of $F_q(x) - \alpha$, hence $x^*(q) = F_q^{-1}(\alpha)$. If $x^*(\hat{p}) \geq x^*(p^*)$, the density lower bound and the quantile definition give

$$\underline{f}|x^*(\hat{p}) - x^*(p^*)| \leq F_{p^*}(x^*(\hat{p})) - F_{p^*}(x^*(p^*)) \leq \|F_{\hat{p}} - F_{p^*}\|_\infty .$$

The reverse ordering is identical, proving (14).