JACOBIAN RANK COLLAPSE IN DECISION-FOCUSED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decision-focused learning (DFL) trains predictors through downstream objectives, yet changing the loss does not necessarily create a new direction for updating the model. We study this limitation through the predictor Jacobian. At rank one, all nonzero backpropagated loss gradients are collinear; a spectral bound characterizes near-collinearity when a leading singular direction dominates and the gradients have nonzero components along it. This local restriction does not imply that different losses share an optimum.

We test its empirical relevance in sparse index tracking across six equity markets, and examine transfer to weather, air quality, and shortest paths. Across 38 one-parameter equity configurations spanning a $54\times$ range of sparsity, DFL gains over MSE remain below 1.8% (median 0.25%), with no detected monotone association with sparsity ($\rho = -0.16$, $p = 0.35$). Structured predictors yield larger gains in several markets and up to 9.4% on air quality. A calibrated effective-dimension proxy organizes these outcomes: the archived blind-evaluation record contains 27 correct predictions among 31 non-abstaining cases, with two additional abstentions. Transfer results also expose the limits of fixed thresholds. For higher-capacity models, covariance regularization mitigates task-training instability. Together, these findings motivate assessing the gradient directions a predictor can express, validating the diagnostic in the target domain, and regularizing task training when needed.

1 INTRODUCTION

When should a practitioner pay for decision-focused learning (DFL) rather than fit a predictor and optimize its outputs? Differentiable optimization and decision-aware surrogates offer several ways to train end to end (Elmachtoub & Grigas, 2022; Donti et al., 2017; Wilder et al., 2019; Mandi et al., 2024). Their benefit depends on both the downstream objective and the predictor: a decision-relevant error direction is useful only if the model can express an update along it. We study this second restriction.

The central object is the predictor Jacobian $\mathbf{J} = \partial \mathrm{vec}(\Sigma)/\partial\theta$. Both MSE and task gradients pass through $\mathbf{J}^\top$ before updating the parameters. At rank one, their nonzero images lie on one line (Proposition 4); when a leading singular direction dominates, a spectral bound controls the remaining angular separation (Theorem 1). Rank collapse therefore excludes a new *independent direction*, but does not exclude different signs, stationary points or final decisions. We distinguish this exact local statement from the empirical question of whether DFL improves test performance.

From geometry to a testable diagnostic. Sparse index tracking provides a controlled setting. After selecting K assets, a QP uses at most $2NK - K^2$ covariance entries; selection and weighting are separate operations. This support restriction can separate MSE and task gradients in covariance space. The predictor Jacobian then determines how much of that separation survives in parameter space. We use $d_{\mathrm{eff}} = d \cdot h$ as a calibrated proxy for this capacity, not as an identity for Jacobian rank. The resulting two-axis hypothesis is that sparsity and model capacity must be assessed together.

Across 38 one-parameter configurations in six equity markets, DFL's gain over MSE remains below 1.8% (median 0.25%), with no detected monotone sparsity association ($\rho = -0.16$, $p = 0.35$). With structured predictors, the gains reach 2.95% on ASX and 2.41% on S&P 1500. Cross-domain

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.

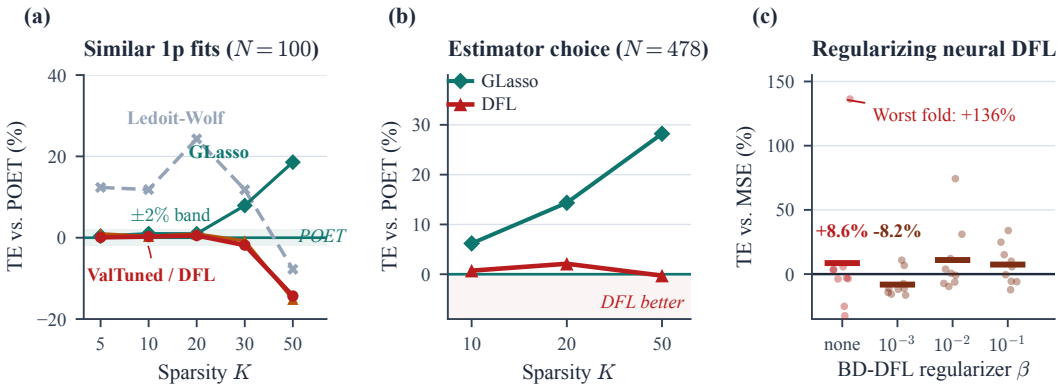

Figure 1: Loss choice, model capacity, and selection stability. (a)–(b) Tracking error relative to POET; (c) relative to MSE (lower is better). At N=100, the three fitted one-parameter estimators improve on Ledoit–Wolf but meet the empirical equivalence criterion in Table 1. At N=478, estimator choice changes the ordering. Under dynamic selection, BD-DFL reduces the extreme errors of unregularized neural DFL; dots show individual folds.

experiments provide both supporting cases and limits: the same spatial model behaves differently on NOAA and EPA data, while shortest-path experiments support the capacity ordering but require recalibrated thresholds. Figure 1 summarizes the financial regimes; Figure 2 shows the training paths.

Contributions.

1. **A geometric limitation.** We separate covariance-gradient support from predictor capacity, proving collinearity for rank-one Jacobians and a conditional spectral bound near rank one. Neither result alone asserts equal loss minimizers.
2. **An empirical diagnostic and its limits.** We test a capacity proxy across equity and spatial domains and report the archived blind score transparently: 27/31 correct committed predictions, plus two abstentions. Transfer to shortest paths tests the ordering separately from the numerical thresholds.
3. **Regularization when capacity is available.** BD-DFL stabilizes higher-capacity task training under changing selection, improving tracking error by 15.6% at $N = 451$ ($p = 0.0024$). We give a selection-robustness motivation and distinguish it from a general minimax guarantee.

2 Related Work

Sparse Index Tracking. Sparse portfolio construction has a rich history (Kolm et al., 2014): evolutionary heuristics (Beasley et al., 2003), mixed-integer programming (Canakgoz & Beasley, 2009), ℓ_1-penalized formulations (Benidis et al., 2018; Brodie et al., 2009), and cardinality-constrained optimization (Xu et al., 2016). These methods treat covariance estimation and portfolio optimization as separate stages.

Decision-Focused Learning. DFL trains predictors using downstream objectives through differentiable QPs (Donti et al., 2017; Amos & Kolter, 2017; Agrawal et al., 2019), SPO+ (Elmachtoub & Grigas, 2022), combinatorial surrogates (Wilder et al., 2019; Vlastelica et al., 2020), or implicit differentiation (Blondel et al., 2022; Paulus et al., 2024). PG losses have asymptotic decision-quality guarantees under misspecification (Gupta & Huang, 2024); PEAR characterizes regret gradients through active-constraint tangent spaces and local curvature (Lee et al., 2026). Other work studies regret decomposition (Aldridge, 2026), online DFL (Capitaine et al., 2026), and prediction inflation in portfolios (Wang & Hasuike, 2026). Our analysis concerns a different stage: the predictor Jacobian that maps these output-space signals into trainable parameter directions.

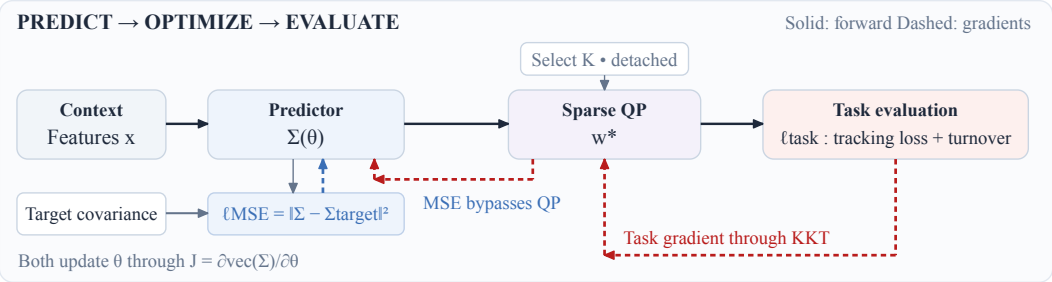

Figure 2: Two objectives, one predictor. The MSE path bypasses the QP; the task path differentiates through its KKT system. Both reach θ through the predictor Jacobian $\mathbf{J}$. Subset selection is detached.

Portfolio Optimization and Covariance Estimation. End-to-end portfolio construction (Butler & Kwon, 2023; Zhang et al., 2020; Kim et al., 2025) and high-dimensional covariance estimation (Ledoit & Wolf, 2004; Fan et al., 2013; Friedman et al., 2008; Engle, 2002) are both active areas. Concurrent work by Jeon et al. (2026) applies DFL to sparse tangent portfolios and reports gains in larger universes—complementary to our study of local gradient geometry and low-capacity predictors.

3 Problem Formulation

Let $\mathbf{r}_t \in \mathbb{R}^N$ denote asset returns, $\mathbf{w}_{\text{idx}}$ the index weights, and $\Sigma = \text{Cov}(\mathbf{r}_t)$. Tracking error minimization reduces to $\min_{\mathbf{w}} (\mathbf{w}-\mathbf{w}_{\text{idx}})^\top \Sigma (\mathbf{w}-\mathbf{w}_{\text{idx}})$. The joint selection-and-weighting problem is NP-hard (Beasley et al., 2003), so we decompose into two stages.

Stage 1: Subset selection. Choose $\mathcal{S} \subseteq \{1,\dots,N\}$ with $|\mathcal{S}| = K$ via market-cap ranking (default) or tracking-score: $s_i = w_i^{\text{idx}} \cdot (\hat{\Sigma}_{i,:}\mathbf{w}_{\text{idx}})^2/\hat{\Sigma}_{ii}$. Under tracking-score selection, $\mathcal{S}$ depends on $\hat{\Sigma}$, so off-block estimation errors propagate to stock selection.

Stage 2: Weight optimization (QP). Given $\mathcal{S}$, solve:

$$\min_{\mathbf{w}_\mathcal{S}} \ \mathbf{w}_\mathcal{S}^\top \hat{\Sigma}_{\mathcal{S}\mathcal{S}} \mathbf{w}_\mathcal{S} - 2\mathbf{w}_\mathcal{S}^\top \hat{\Sigma}_{\mathcal{S},:}\mathbf{w}_{\text{idx}} \ \text{ s.t. } \mathbf{1}^\top \mathbf{w}_\mathcal{S}=1,\ \mathbf{w}_\mathcal{S}\geq 0. \tag{1}$$

Selection $\mathcal{S}$ is detached from the graph (Appendix O). The QP reads at most $2NK - K^2$ of N^2 entries; this asymmetry drives our gradient alignment analysis.

4 Methodology

4.1 Two-Stage Baseline and DFL

The standard two-stage approach trains under Frobenius loss $\mathcal{L}_{\text{MSE}} = \|\hat{\Sigma}(\mathbf{x}_t;\theta) - \Sigma_{\text{realized},t}\|_F^2$, where $\Sigma_{\text{realized},t}$ is the 63-day sample covariance. In DFL, we embed the QP as a differentiable layer:

$$\mathcal{L}_{\text{task}} = \frac{252}{H}\sum_{s=1}^{H} \left(\mathbf{w}^*(\hat{\Sigma}_t)^\top \mathbf{r}_{t+s} - r_{t+s}^{\text{idx}}\right)^2 + \gamma \|\mathbf{w}_t^* - \mathbf{w}_{t-1}^*\|_1, \tag{2}$$

where $\mathbf{w}^*(\hat{\Sigma}_t)$ is the QP solution, $H = 21$ days, and γ controls turnover. Gradients flow through the KKT conditions (Amos & Kolter, 2017) into θ. Selection $\mathcal{S}$ is detached from the graph.

4.2 Block-Diagonal DFL (BD-DFL)

Under detached selection, the task gradient is restricted to the current support $\mathcal{T}$. Errors outside that support remain unpenalized by the task objective and may matter after selection changes. BD-DFL

adds relative MSE regularization:

$$\mathcal{L}_{\text{BD-DFL}} = \mathcal{L}_{\text{task}} + \beta \cdot \|\hat{\Sigma}(\theta) - \Sigma_{\text{realized}}\|_F^2 / \|\Sigma_{\text{realized}}\|_F^2, \tag{3}$$

where $\beta = 0$ recovers pure DFL and large β recovers MSE. A selection-robustness perspective motivates this penalty: entries outside today's support may enter tomorrow's QP. Classical DRO provides related regularization principles (Mohajerin Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019), but does not establish the swap-dependent coefficient claimed for this particular model. We therefore choose β empirically and state the limitation in Appendix F.

Proposition 1 (Selection-robustness motivation)**.** *Under changing subsets, controlling the full covariance error controls errors on every selected submatrix. This motivates full-matrix regularization; without further assumptions on error concentration and solution stability, it does not imply a universal r^2/N^2 minimax coefficient.*

4.3 Covariance Models

Five architectures of increasing capacity, each trained under MSE and DFL: **Shrinkage** (1p): $\hat{\Sigma} = (1-\alpha)S + \alpha\mu I$ (Ledoit & Wolf, 2004); **Factor** ($O(NK_f)$p): $\hat{\Sigma} = BFB^\top + D$; **Neural** ($O(Nr)$p): MLP $\to LL^\top + D$; **Structured** (12p): learnable per-sector targets; **Conditional** (385p): regime-adaptive $\alpha_t = \sigma(\text{MLP}(\mathbf{z}_t))$. Training: Adam (lr 10^{-3}, weight decay 10^{-4}, clip 1.0), early stopping on validation TE, cvxpylayers (Agrawal et al., 2019) for QP differentiation. Details in Appendix S.

5 Experimental Setup

5.1 Data

Top 100 S&P 500 stocks by market cap (N=97 after $\geq$90% coverage filter), daily returns 2006–2025. Rolling-window protocol: 3-year train, 6-month validation, 1-year test, 9 folds covering 2016–2025 in the primary neural evaluation. Features: 21-day volatility, 63-day momentum, sector, market-cap quintile. Scaling: S&P 500 full (N=478, 5 folds), S&P 1500 (N=1258, 8 folds), 20-year backtest (N=100, 17 folds). Cross-market: FTSE 100, Nikkei 225, Euro Stoxx 50, ASX 200, Hang Seng. Cross-domain: NOAA weather stations ($N \in \{100, 189, 415\}$), EPA PM2.5 monitors ($N \in \{124, 132\}$). All p-values from Wilcoxon signed-rank across folds; one-sided for directional hypotheses and two-sided for difference tests; equivalence uses TOST with the stated margin.

5.2 Baselines

We compare: *Naive* (top-K by index weight), *K-only LW* ($K \times K$ covariance only), *Ledoit-Wolf* (analytical shrinkage (Ledoit & Wolf, 2004)), *POET* (factor + adaptive thresholding (Fan et al., 2013)), *GraphicalLasso* (ℓ_1-penalized precision (Friedman et al., 2008)), *ValTuned* (grid search on validation TE), *SPO+* (Elmachtoub & Grigas, 2022), *LODL* (locally optimized decision losses (Shah et al., 2022)), and MSE-trained factor/neural models. Primary metric: annualized TE = std($\mathbf{w}^\top \mathbf{r}_t - r_t^{\text{idx}}$) $\times \sqrt{252}$.

6 Results

6.1 The No-Go Regime: $d_{\text{EFF}} \leq 1.5$

Main finding: DFL adds a $\sim$3,300$\times$ training-cost overhead with little measured benefit in the tested low-capacity settings.

At N=100, fitted one-parameter models achieve similar test performance across different training objectives. This is an empirical result, complementary to the local collinearity statement: the theorem restricts available directions, whereas test-loss equivalence must be measured.

Table 1 reports equivalence under TOST with a 20-bps margin for all 12 pairwise comparisons among ValTuned, SPO+, and DFL. Each improves on Ledoit–Wolf by 7–19%, so fitting the parameter matters even when these objectives yield similar test errors. POET and GLasso are separate

Table 1: No-go regime at N=100. The three d=1 shrinkage models—grid search on validation (ValTuned), the SPO+ surrogate, and end-to-end DFL—reach the same tracking error by TOST at δ=20 bps in *all* 12 pairwise comparisons (3 pairs × 4 sparsity levels, max $|d|$=0.66), despite being fitted by three unrelated procedures. POET and GLasso are not d=1 models and are shown for reference: they match the d=1 family while the optimizer is sparse but part company at K=50 (K/N=0.52), where the $\sqrt{2K/N}$ bound is vacuous and no equivalence is claimed.

K	LW	POET	GLasso	ValTuned	SPO+	DFL
5	0.0860	**0.0765**	0.0768	0.0773	0.0766	0.0766
10	0.0511	**0.0457**	0.0461	0.0458	0.0458	0.0458
20	0.0331	**0.0267**	0.0269	0.0269	0.0269	0.0268
50	0.0096	0.0104	0.0124	0.0088	**0.0088**	0.0089

Table 2: N=478, 5 folds, one evaluation window. d: Cohen's d on per-fold relative change; p: one-sided Wilcoxon, for which 0.031 is the floor at n=5 (i.e. 5/5 folds).

Model	K	MSE TE	DFL TE	d	p
GLasso	20	0.0575		—	—
POET	20	0.0503		—	—
Shrink. (1p)	20	0.0517	0.0514	−1.07	0.031*
Cond. (385p)	20	**0.0454**	**0.0448**	−0.83	0.062
Struct. (12p)	20	0.0513	0.0508	−1.60	0.031*
Struct. (12p)	50	0.0309	0.0305	−1.03	0.062

model families; their agreement at smaller K is empirical. LODL also matches DFL, indicating that explicit KKT differentiation is not necessary to obtain this performance. Higher-capacity models show different outcomes (Table 2); the neural gain is significant at $K \leq 30$ but absent at K=50 (Appendix G).

6.2 The DFL Regime: $d_{\text{EFF}} \geq 4$ at Scale

Main finding: DFL unlocks multi-parameter models that grid search cannot tune.

Scaling from N=100 to N=478 reveals DFL's regime (Table 2).

Table 2 demonstrates DFL's niche. GLasso degrades at scale (14.3% behind POET as $T/N \to 1$), and architecture matters more than loss: the 385-parameter conditional model, trained on MSE alone, beats POET by 9.8%. The DFL gain grows with capacity—+0.68%, +1.07%, +1.34% at d=1, 12, 385. Even the 1-parameter model now wins 5/5 folds, but at +0.68%, well inside the no-go band: scale buys a consistent sign, not an economically meaningful effect. With 5 folds, p=0.031 is the attainable floor, so N=478 corroborates the direction of the N=100 results rather than carrying the claim.

6.3 Cross-Market Validation of the No-Go Theorem

We test across six markets (Table 3). The 1-parameter gain is not merely small but *flat*: over all 38 d=1 configurations (K/N from 0.008 to 0.43) it stays within $[-0.57\%, +1.76\%]$ and is uncorrelated with sparsity (ρ=−0.16, p=0.35)—an empirical flatness that is not implied by rank-one collinearity

Table 3: Cross-market dim-1 validation, 81 paired folds. The 1-parameter gain (Shrink $\Delta\%$) is negligible everywhere and does not grow with K/N: the two largest K/N here, Euro Stoxx (0.21) and the 20-year run (0.20), give $+0.29\%$ each, while Hang Seng's $+1.76\%$ sits at 0.16; over all 38 d=1 configurations (316 paired folds) $\rho(K/N, \text{gain})$=-0.16 (p=0.35), an empirical pattern rather than a consequence of collinearity. Structured (12p) gains are meaningful only on ASX, Hang Seng and S&P 1500: the gain needs capacity, which d=1 lacks at any K/N. The Nikkei, Euro Stoxx, ASX and Hang Seng rows come from a single four-market run under one configuration (8 folds each). $^{*}p<0.05$, $^{**}p<0.01$, one-sided Wilcoxon.

Market	N	K	Shrink $\Delta\%$	Struct $\Delta\%$	p_{struct}
S&P 500 (20y)	100	20	$+0.29$	$+0.40$	0.020^{*}
FTSE 100	91	10	-0.09	$+0.37$	0.039^{*}
Nikkei 225	218	10	$+0.17$	$+0.87$	0.027^{*}
Euro Stoxx 50	47	10	$+0.29$	$+0.95$	0.008^{**}
ASX 200	159	10	$+0.98$	$\mathbf{+2.95}$	$\mathbf{0.004}^{**}$
Hang Seng	62	10	$+1.76$	$+2.77$	0.008^{**}
S&P 1500	1258	10	$+0.32$	$+2.41$	0.004^{**}
S&P 1500	1258	20	$+0.44$	$+2.21$	0.004^{**}
S&P 1500	1258	50	-0.13	$+0.98$	0.039^{*}

(Appendix V). Structured (12p) DFL is economically meaningful on ASX ($+2.95\%$), Hang Seng ($+2.77\%$) and S&P 1500 ($+2.41\%$, all $p \leq 0.008$); on Euro Stoxx, Nikkei and FTSE it is detectable at most, and capacity alone buys nothing (Appendix V).

6.4 BD-DFL under Dynamic Selection

The effect of regularization depends on the model and selection rule. With dynamic tracking-score selection at N=100, neural DFL increases mean TE by 8.6% at K=10; BD-DFL with $\beta = 0.001$ reduces it by 8.2% relative to MSE. At N=478, the corresponding reduction is 7.4% ($p = 0.063$, five folds). In the corrected N=451 run, pure DFL already improves by 13.0% at K=10 and BD-DFL improves by 15.6% ($p = 0.0024$, 9/11 wins). At K=20, pure DFL increases TE by 4.5%, while BD-DFL reduces it by 5.5% ($p = 0.103$). These results motivate validation of regularization, not a universal benefit or coefficient (Appendix W).

7 Analysis

7.1 Gradient Alignment Theory

The task gradient decomposes as:

$$\nabla_\theta \ell = \underbrace{\frac{\partial \ell}{\partial \mathbf{w}^*}}_{\text{task}} \cdot \underbrace{\frac{\partial \mathbf{w}^*}{\partial \hat{\Sigma}}}_{\text{QP Jacobian}} \cdot \underbrace{\frac{\partial \hat{\Sigma}}{\partial \theta}}_{\text{model}}. \tag{4}$$

The QP Jacobian $\partial \mathbf{w}^*/\partial \hat{\Sigma}$, derived from KKT differentiation (Amos & Kolter, 2017), acts as a *support-restricted linear map*: it zeroes out entries $\hat{\Sigma}_{ij}$ where both $i \notin \mathcal{S}$ and $j \notin \mathcal{S}$, since the QP reads only $\hat{\Sigma}_{\mathcal{S}\mathcal{S}}$ and $\hat{\Sigma}_{\mathcal{S},:}\mathbf{w}_{\text{idx}}$. MSE can place gradient energy on all entries. A small support limits alignment only when little MSE energy concentrates on that support.

Proposition 2 (Gradient Alignment Bound)**.** *Let $\mathcal{S} \subset \{1, \ldots, N\}$, $|\mathcal{S}| = K$. Define the task-relevant support $\mathcal{T} = \{(i,j) : i \in \mathcal{S} \text{ or } j \in \mathcal{S}\}$, with $|\mathcal{T}| = 2NK - K^2$. For nonzero gradients,*

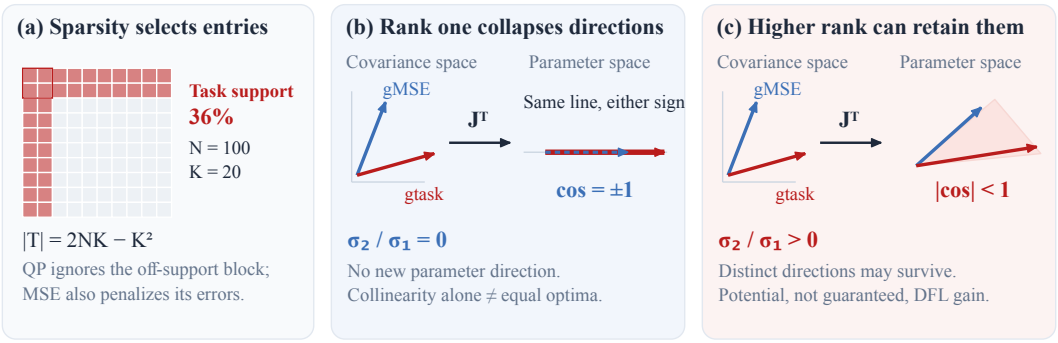

Figure 3: Where gradient directions are lost. (a) The task-relevant support has $2NK - K^2$ entries (36% here). (b) A rank-one predictor Jacobian maps nonzero gradients onto one line, with either sign. (c) With multiple active singular directions, angular separation can survive; improved task performance is possible, not guaranteed.

suppose the normalized MSE energy obeys $\mathbb{E}[\|\mathbf{g}_{\mathrm{MSE}}|_{\mathcal{T}}\|^2/\|\mathbf{g}_{\mathrm{MSE}}\|^2] \le |\mathcal{T}|/N^2$. *Then:*

$$\mathbb{E}|\cos(\mathbf{g}_{\mathrm{MSE}}, \mathbf{g}_{\mathrm{task}})| \le \sqrt{\tfrac{2K}{N} - \tfrac{K^2}{N^2}} \approx \sqrt{\tfrac{2K}{N}} \quad (K \ll N). \tag{5}$$

This support-only inequality is sharp for unrestricted ambient vectors (Appendix B); attaining it with gradients of the tracking QP is a separate question.

Proposition 3 (Decision Invariance)**.** *The QP* (1) *depends on* $\hat{\Sigma}$ *only through* $\hat{\Sigma}_{\mathcal{S}\mathcal{S}}$ *and* $\hat{\Sigma}_{\mathcal{S},:}\mathbf{w}_{\mathrm{idx}}$*; off-block errors are irrelevant to both the solution and regret (Proposition 6; proofs in Appendix D).*

Empirically (K=20, N=97): 87.5% of task gradient energy concentrates in the $K \times K$ block (Appendix A). POET's off-block error is $3.4\times$ DFL's, yet TEs are indistinguishable ($p = 0.59$).

7.2 The Effective Dimension Diagnostic

The support bound concerns covariance-space geometry, not DFL advantage. We next study which directions survive the predictor Jacobian and evaluate a practical capacity proxy.

Definition 1 (Effective Model Dimension)**.** *For a structured covariance model with d region-specific parameters, let $h = CV(\{\sigma^2_{r,target}\}_{r=1}^R)$ be the coefficient of variation of per-region target-weight concentrations. We define the capacity proxy $d_{eff} = d \cdot h$. It is not a spectral rank and need not be bounded by d.*

Proposition 4 (Jacobian Rank Collapse)**.** *For any differentiable covariance model $\Sigma(\theta)$ with $\theta \in \mathbb{R}^d$ and Jacobian $\mathbf{J} = \partial \mathrm{vec}(\Sigma)/\partial\theta \in \mathbb{R}^{N^2 \times d}$, when both parameter gradients are nonzero, the parameter-space alignment satisfies $\cos(\nabla_\theta \ell_{MSE}, \nabla_\theta \ell_{task}) = (\mathbf{g}_{MSE}^\top \mathbf{J}\mathbf{J}^\top \mathbf{g}_{task})/(\|\mathbf{J}^\top \mathbf{g}_{MSE}\|\|\mathbf{J}^\top \mathbf{g}_{task}\|)$. When $\mathrm{rank}(\mathbf{J}) = 1$, this collapses to $\{-1, +1\}$: MSE and task gradients are* provably collinear *in parameter space, regardless of their N^2-space misalignment.*

Proof sketch. Write $\mathbf{J} = \sigma \mathbf{u}\mathbf{v}^\top$. Then $\mathbf{J}^\top \mathbf{g} = \sigma(\mathbf{u}^\top \mathbf{g})\mathbf{v}$, so both nonzero images are scalar multiples of $\mathbf{v}$. Their signs and zeros can differ. Equality of minimizers requires further assumptions (Appendix C). The empirical thresholds $d_{\mathrm{eff}} \le 1.5$ and $d_{\mathrm{eff}} \ge 4$ are calibrated rules, not consequences of this proposition. □

The following theorem gives the exact form of parameter-space alignment and bounds how quickly it collapses to ± 1 as the Jacobian approaches rank one.

Theorem 1 (Spectral Gap Controls Gradient Alignment)**.** *Let $\mathbf{J} = USV^\top \in \mathbb{R}^{N^2 \times d}$ be the Jacobian of $\mathrm{vec}(\Sigma(\boldsymbol{\theta}))$, with $\sigma_1 > 0$, nonzero parameter gradients and nonzero leading-gradient components $\mathbf{u}_1^\top \mathbf{g}$, and singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_d \ge 0$ and leading left singular vector $\mathbf{u}_1$. Then:*

1. (Exact form.) *Parameter-space alignment is the alignment of the σ-weighted coordinates of the two gradients along the singular directions:*

$$\cos(\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{MSE}}, \nabla_{\boldsymbol{\theta}}\ell_{\mathrm{task}}) = \cos\big(SU^\top \mathbf{g}_{\mathrm{MSE}},\ SU^\top \mathbf{g}_{\mathrm{task}}\big). \tag{6}$$

2. (Near-rank-one collapse.) *For each gradient let $\rho = \min\{1, (\sigma_2/\sigma_1)/|\cos(\mathbf{g}, \mathbf{u}_1)|\}$, and let $\Theta = \arcsin \rho_{\mathrm{MSE}} + \arcsin \rho_{\mathrm{task}}$. If $\Theta < \pi/2$, then*

$$|\cos(\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{MSE}}, \nabla_{\boldsymbol{\theta}}\ell_{\mathrm{task}})| \ \geq\ \cos\Theta, \qquad \operatorname{sign}\cos(\cdot,\cdot) = \operatorname{sign}\big[(\mathbf{u}_1^\top \mathbf{g}_{\mathrm{MSE}})(\mathbf{u}_1^\top \mathbf{g}_{\mathrm{task}})\big]. \tag{7}$$

3. (Endpoints.) *$\sigma_2{=}0$ gives $|\cos|{=}1$ (Proposition 4); N^2 orthogonal equal-norm columns give the N^2-space alignment, $\leq \sqrt{2K/N}$.*

Proof sketch. $\nabla_{\boldsymbol{\theta}}\ell = V(SU^\top \mathbf{g})$ with V orthogonal gives (6); each $SU^\top \mathbf{g}$ lies within angle $\arcsin\rho$ of $\pm\mathbf{e}_1$, and the angle is a metric on the sphere (Appendix C). □

The effective rank $r_{\text{eff}} = \exp(-\sum_i p_i \log p_i)$, $p_i = \sigma_i^2/\sum_j \sigma_j^2$, summarizes the gap: $r_{\text{eff}} \to 1$ forces $\sigma_2/\sigma_1 \leq \sqrt{(1-p_1)/p_1} \to 0$. The converse fails—$r_{\text{eff}}$ does not determine alignment for intermediate spectra—so we use it, through d_{eff}, as a diagnostic rather than a formula.

The positive complement (Corollary 2, Appendix C): when $r_{\text{eff}} > 1$, a "DFL-exclusive gradient" exists that MSE training cannot access. This yields a two-axis decision rule (Appendix U): at $d_{\text{eff}} \leq 1.5$ MSE suffices—DFL is unnecessary when alignment > 0.4 and all paths simply converge when it is not—while at $d_{\text{eff}} \geq 4$ with alignment < 0.4 DFL wins; everything between is ambiguous.

Controlled validation. A synthetic sweep ($N{=}100$, $K{=}20$, varying h, 18 settings) isolates d_{eff}: gains stay within $\pm 1.6\%$ for $d_{\text{eff}} \leq 9$ (consistent with noise) and jump to $+3.9\%$ at $d_{\text{eff}} \approx 10$ (Appendix X). Cross-domain tests confirm the diagnostic transfers outside finance: NOAA weather stations ($d{=}10$ spatial parameters, $h{=}0.14$, $d_{\text{eff}}{=}1.4$) show DFL is actively harmful (-3.6% at $K{=}20$, $N{=}100$), supporting a low-gain regime in this tested domain. EPA PM2.5 monitors ($d{=}10$, $h{=}0.61$, $d_{\text{eff}}{=}6.1$) show DFL gains up to $+9.4\%$ ($K{=}50$, $N{=}132$). Both domains use the same 10-parameter structured model—the domains also differ in observations, targets, and index-weight heterogeneity h (Appendix M).

Blind validation. The archived scoring file matches 33 configurations to predictions committed before the original experiments. Its reported 29/33 (87.9%) counts two abstentions as correct; excluding them gives 27/31 (87.1%). Figure 4 visualizes this frozen scoring record, not the subsequently corrected financial reruns. Re-scoring after changing the evaluation harness is needed before claiming prospective accuracy for that harness.

7.3 Combinatorial Optimization: Shortest Path

Does the no-go mechanism transfer to *combinatorial* optimization? We test on shortest path benchmarks ($8{\times}8$ and $12{\times}12$ grids, SPO+ (Elmachtoub & Grigas, 2022)). A bottleneck architecture controls r_{eff} by construction: width k bounds Jacobian rank (Table 24).

At $r_{\text{eff}} \approx 1$ SPO+ provides no advantage ($\Delta = -0.4\%$); at $r_{\text{eff}} \geq 10$ it wins (up to $+4.1\%$). The $12{\times}12$ grid ($N{=}144$) amplifies Δ roughly $4\times$ at each rank level, rising to $+16.4\%$ at $r_{\text{eff}}{=}20$: the Jacobian-rank mechanism is solver-agnostic and scale-dependent (Table 25, Appendix Y).

Limitations and practical impact. The rank result is pointwise; aggregation across samples with different Jacobians need not preserve collinearity. Only the support bound needs sparse detached selection; rank collapse is indifferent to K/N, and survivorship bias may affect both absolute and relative TE. The thresholds (1.5, 4) are calibrated, not tight: on PyEPO's shortest path (Tang & Khalil, 2024) the d_{eff} ordering transfers ($\rho{=}0.93$) but the cut at 4 does not, so d_{eff} is best read as a ranking and re-calibrated per domain (Appendix Z). In the no-go regime POET matches DFL's tracking error at 0.23 s against 13 min per fold, a $3,329\times$ gap for no measurable return (Appendix T; robustness in Appendices O–Q).

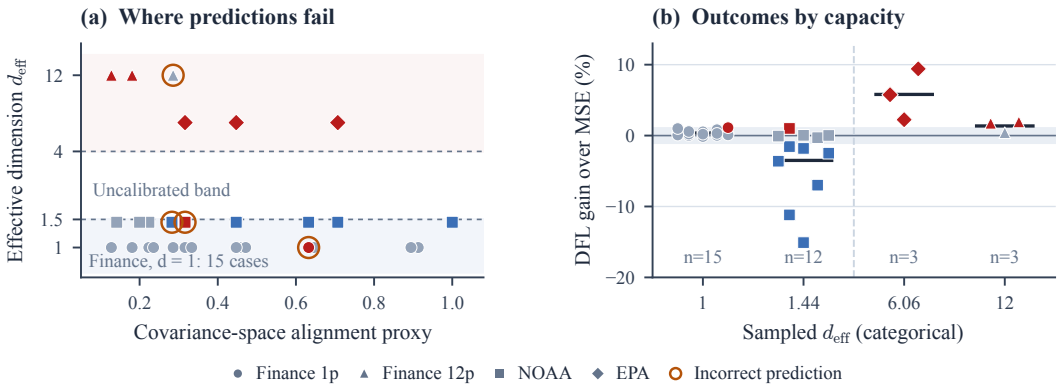

Figure 4: Archived blind-evaluation record (33 matched cases). Colour shows DFL gain: red $> 1\%$, grey within $\pm 1\%$, blue $< -1\%$. (a) Four incorrect predictions are ringed; overlapping finance points are counted. (b) Equal spacing denotes the four sampled categories, not a continuous sweep; horizontal marks show means. This historical snapshot includes two abstentions.

8 Conclusion

The predictor Jacobian determines which output-space gradient directions can affect training. Rank one enforces collinearity, while the spectral bound characterizes conditional near-collinearity; neither alone guarantees that MSE is task-optimal. Our experiments identify a low-capacity regime with modest DFL gains and higher-capacity settings where task learning or regularization helps. The practical implication is to diagnose capacity, validate thresholds on the target domain, and compare decision quality against training cost before selecting a learning objective.

Ethics Statement

This work studies structural limitations of decision-focused learning in predict-then-optimize pipelines. Our experiments use only publicly available financial, weather, and air quality data. No human subjects are involved. The diagnostic we propose helps practitioners avoid unnecessary computational cost (measured at $\sim$3,300$\times$ a tuned statistical baseline, in one tested configuration with little measured gain), reducing energy consumption in machine learning workflows. We see no negative societal impacts specific to this work.

AI Use Statement

We used Claude (Anthropic) and Codex (OpenAI) to assist with analysis and plotting code, result auditing, literature checks, drafting and revising prose, and figure design. AI-assisted auditing identified evaluation-harness defects and unsupported theoretical steps, which prompted corrections and narrower claims in this revision. The accompanying audit records distinguish checked data, historical results, and remaining limitations. The authors remain responsible for the scientific content and the final submitted version.

Reproducibility Statement

All data is publicly available (Yahoo Finance, NOAA ISD, EPA AQS) or synthetic. Architectures and hyperparameters are in Sections 3–4 and Appendices S, Y. Pre-registered predictions are version-controlled. Anonymous code will be provided.

Evaluation harness. Auditing our own results surfaced two defects in how test windows and rebalancing dates were built, and we report them because they changed numbers in this paper. First, each fold's final rebalance was held to the end of the *dataset* instead of the end of that fold's test window, appending a stale-weights tail to every fold and making the folds overlap in calendar time. Second, monthly rebalance dates were taken as calendar month-end *labels*: a month whose last calendar day fell on a weekend was skipped entirely (about 29% of months), a label falling past the window's end

became a phantom rebalance that stretched the preceding holding period outside the window, and the same phantom in a validation window could extend a holding period into the test period, letting tuned baselines see test data during model selection.

Both are fixed at the source—rebalancing now uses the last trading day of each month within the window—and the affected equity results used for current comparisons were re-run under the repaired harness. The archived blind-diagnostic record is explicitly retained as a historical evaluation and is not a corrected-harness result. The table verifier checks 288 numerical cells against their packaged sources; additional BD-DFL checks cover tables omitted by that verifier. Return-series files support direct recomputation of rolling figures; some runs expose only per-fold metrics. The release distinguishes these levels of evidence.

Corrections changed both numerical magnitudes and some local interpretations. For example, the corrected $N = 451$, $K = 10$ run shows an improvement from pure DFL, contrary to the earlier overfitting description. The neural $K = 20$ gain is 16.5%, while the one-parameter cross-market association remains weak. Archived blind scores are reported separately, including abstentions. The NOAA, EPA, shortest-path and PyEPO evaluations use separate protocols. The accompanying source and audit records are provided to make these distinctions explicit.

REFERENCES

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

Irene Aldridge. Regret equals covariance: A closed-form characterization for stochastic optimization. *arXiv preprint arXiv:2605.14019*, 2026.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

John E. Beasley, Nigel Meade, and T.-J. Chang. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3):621–643, 2003.

Konstantinos Benidis, Yiyong Feng, and Daniel P. Palomar. Sparse portfolios for high-dimensional financial index tracking. *IEEE Transactions on Signal Processing*, 66(1):155–170, 2018.

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

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30): 12267–12272, 2009.

Andrew Butler and Roy H. Kwon. Integrating prediction in mean-variance portfolio optimization. *Quantitative Finance*, 23(3):429–452, 2023.

Nalan A. Canakgoz and John E. Beasley. Mixed-integer programming approaches for index tracking and enhanced indexation. *European Journal of Operational Research*, 196(1):384–399, 2009.

Aymeric Capitaine, Maxime Haddouche, Eric Moulines, Michael I. Jordan, Etienne Boursier, and Alain Durmus. Online decision-focused learning. In *International Conference on Learning Representations (ICLR)*, 2026.

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 (NeurIPS)*, 2017.

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

Robert Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20 (3):339–350, 2002.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

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

Haeun Jeon, Seunghoon Choi, Hyunglip Bae, Yongjae Lee, and Woo Chang Kim. Decision-focused sparse tangent portfolio optimization. *arXiv preprint arXiv:2607.00581*, 2026.

Siddharth Joshi and Stephen Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

Juchan Kim, Inwoo Tae, and Yongjae Lee. Estimating covariance for global minimum variance portfolio: A decision-focused learning approach. *arXiv preprint arXiv:2508.10776*, 2025.

Petter N. Kolm, Reha Tütüncü, and Frank J. Fabozzi. 60 years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research*, 234(2):356–371, 2014.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Junhyeong Lee, Sangjin Jin, and Yongjae Lee. Decision-focused learning via tangent-space projection of prediction error. *arXiv preprint arXiv:2605.01361*, 2026.

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.

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

Anselm Paulus, Georg Martius, and Vít Musil. LPGD: A general framework for backpropagation through embedded optimization layers. In *International Conference on Machine Learning (ICML)*, 2024.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-focused learning without differentiable optimization: Learning locally optimized decision losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bo Tang and Elias B. Khalil. PyEPO: a PyTorch-based end-to-end predict-then-optimize library for linear and integer programming. *Mathematical Programming Computation*, 16:297–335, 2024. doi: 10.1007/s12532-024-00255-x.

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

Yi Wang and Takashi Hasuike. Decision-induced ranking explains prediction inflation and excessive turnover in SPO-based portfolio optimization. *arXiv preprint arXiv:2605.01176*, 2026.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *AAAI Conference on Artificial Intelligence*, 2019.

Fengmin Xu, Zhaosong Lu, and Zongben Xu. An efficient optimization approach for a cardinality-constrained index tracking problem. *Optimization Methods and Software*, 31(2):258–271, 2016.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep learning for portfolio optimization. *The Journal of Financial Data Science*, 2(4):8–20, 2020.

A Proof of Proposition 2

Proof. We prove the gradient alignment bound in three steps.

Step 1: Sparsity of the task gradient. Under the hard-K formulation, the QP solution $\mathbf{w}^* \in \mathbb{R}^K$ depends on $\hat{\Sigma}$ through two quantities: the selected submatrix $\hat{\Sigma}_{\mathcal{S}\mathcal{S}} \in \mathbb{R}^{K\times K}$ and the cross-covariance vector $\hat{\boldsymbol{\sigma}}_{\text{idx},\mathcal{S}} = \hat{\Sigma}_{\mathcal{S},:}\mathbf{w}_{\text{idx}} \in \mathbb{R}^K$. Entry (i,j) of $\hat{\Sigma}$ contributes to the cross-covariance whenever $i \in \mathcal{S}$ (regardless of j, since $\mathbf{w}_{\text{idx}}$ has support on all N assets), and to the sub-matrix whenever both $i,j \in \mathcal{S}$. For any pair (i,j) with $i \notin \mathcal{S}$ and $j \notin \mathcal{S}$, the entry $\hat{\Sigma}_{ij}$ does not appear in either quantity. By the chain rule through the KKT conditions (Amos & Kolter, 2017), $(\mathbf{g}_{\text{task}})_{ij} = 0$ for all $(i,j) \in \bar{\mathcal{T}}$.

The task-relevant support is $\mathcal{T} = \{(i,j) : i \in \mathcal{S} \text{ or } j \in \mathcal{S}\}$. Its cardinality is $|\mathcal{T}| = N^2 - (N-K)^2 = 2NK - K^2$.

Step 2: Cosine similarity bound via Cauchy–Schwarz. Since $\mathbf{g}_{\text{task}}$ is supported on $\mathcal{T}$, we can write:

$$\begin{aligned}\langle \mathbf{g}_{\text{MSE}},\, \mathbf{g}_{\text{task}} \rangle &= \sum_{(i,j)\in\mathcal{T}} (\mathbf{g}_{\text{MSE}})_{ij}\,(\mathbf{g}_{\text{task}})_{ij} \\ &= \langle \mathbf{g}_{\text{MSE}}|_{\mathcal{T}},\, \mathbf{g}_{\text{task}} \rangle, \end{aligned} \tag{8}$$

where $\mathbf{g}_{\text{MSE}}|_{\mathcal{T}}$ denotes the restriction of $\mathbf{g}_{\text{MSE}}$ to the support $\mathcal{T}$. By Cauchy–Schwarz:

$$|\langle \mathbf{g}_{\text{MSE}},\, \mathbf{g}_{\text{task}} \rangle|^2 \le \|\mathbf{g}_{\text{MSE}}|_{\mathcal{T}}\|^2 \cdot \|\mathbf{g}_{\text{task}}\|^2. \tag{9}$$

Dividing both sides by $\|\mathbf{g}_{\text{MSE}}\|^2 \cdot \|\mathbf{g}_{\text{task}}\|^2$:

$$\cos^2(\mathbf{g}_{\text{MSE}},\, \mathbf{g}_{\text{task}}) \le \frac{\|\mathbf{g}_{\text{MSE}}|_{\mathcal{T}}\|^2}{\|\mathbf{g}_{\text{MSE}}\|^2}. \tag{10}$$

Step 3: Isotropy assumption. For a fixed support, a sufficient assumption is isotropy of the normalized energy: $\mathbb{E}[g_{ij}^2/\|\mathbf{g}\|^2] = 1/N^2$ for every coordinate. For data-dependent selection the corresponding conditional assumption is needed. Under this condition,

$$\mathbb{E}\left[\frac{\|\mathbf{g}_{\text{MSE}}|_{\mathcal{T}}\|^2}{\|\mathbf{g}_{\text{MSE}}\|^2}\right] = \frac{|\mathcal{T}|}{N^2} = \frac{2NK-K^2}{N^2} = \frac{2K}{N} - \frac{K^2}{N^2}. \tag{11}$$

Combining with Step 2 and applying Jensen's inequality:

$$\mathbb{E}[\cos(\mathbf{g}_{\text{MSE}},\, \mathbf{g}_{\text{task}})] \le \sqrt{\frac{2K}{N} - \frac{K^2}{N^2}}. \tag{12}$$

For $K \ll N$, the K^2/N^2 term is negligible, yielding $\cos(\mathbf{g}_{\text{MSE}},\, \mathbf{g}_{\text{task}}) \lesssim \sqrt{2K/N}$. □

Scope of the energy assumption. Equal unnormalized second moments do not suffice: expectation of a ratio is not the ratio of expectations. Normalized-energy isotropy is an explicit sufficient condition, not a property of every covariance estimator. Selection can concentrate error energy; then the deterministic support-energy ratio in Step 2 remains valid, while the numerical $\sqrt{2K/N}$ bound need not hold.

B SHARPNESS OF THE SUPPORT-ONLY INEQUALITY

Proposition 5 (Ambient-vector sharpness)**.** *For a support $\mathcal{T}$ of size m in $\mathbb{R}^{N^2}$, there are nonzero vectors $\mathbf{g}, \mathbf{h}$ with $\mathbf{h}$ supported on $\mathcal{T}$ and uniform squared coordinates in $\mathbf{g}$, such that $\cos(\mathbf{g}, \mathbf{h}) = \sqrt{m/N^2}$.*

Proof. Set every coordinate of $\mathbf{g}$ to one and let $\mathbf{h}$ be the indicator of $\mathcal{T}$. Their inner product is m, and their norms are N and $\sqrt{m}$. Their cosine is therefore $\sqrt{m}/N$, attaining the support-energy bound. □

This proves sharpness of the linear-algebra inequality only. The earlier equicorrelation construction did not establish a nonzero tracking gradient with the asserted structure. In fact, adding $\epsilon \mathbf{1}\mathbf{1}^\top$ leaves the fully invested tracking objective unchanged because $\mathbf{1}^\top(\mathbf{w} - \mathbf{w}_{\text{idx}}) = 0$. We do not claim a matching lower bound realized by the tracking QP.

C PROOF OF PROPOSITION 4 (JACOBIAN RANK COLLAPSE)

Proof. **Setup.** Let $\Sigma\colon \mathbb{R} \to \mathbb{S}^N_{++}$, $\theta \mapsto \Sigma(\theta)$, be a differentiable covariance model with a single scalar parameter $\theta \in \mathbb{R}$. Define two losses over the covariance estimate:

- *Estimation loss*: $\ell_{\text{MSE}}(\theta) = f(\Sigma(\theta))$ for a differentiable $f\colon \mathbb{S}^N_{++} \to \mathbb{R}$ (e.g., $\|\Sigma(\theta) - \Sigma_{\text{target}}\|_F^2$).
- *Task loss*: $\ell_{\text{task}}(\theta) = \text{TE}(\mathbf{w}^*(\Sigma(\theta)), \Sigma_{\text{true}})$, where $\mathbf{w}^*(\cdot)$ denotes the QP solution (1).

Both losses are differentiable in θ wherever the QP solution is unique with non-degenerate KKT conditions (Amos & Kolter, 2017; Agrawal et al., 2019).

Step 1: Chain rule through the Jacobian. Define the vectorized Jacobian $\mathbf{J} = \partial \text{vec}(\Sigma(\theta))/\partial\theta \in \mathbb{R}^{N^2}$ and the N^2-dimensional loss gradients:

$$\mathbf{g}_{\text{MSE}} = \frac{\partial \ell_{\text{MSE}}}{\partial \text{vec}(\Sigma)} \in \mathbb{R}^{N^2}, \quad \mathbf{g}_{\text{task}} = \frac{\partial \ell_{\text{task}}}{\partial \text{vec}(\Sigma)} \in \mathbb{R}^{N^2}. \tag{13}$$

The task gradient $\mathbf{g}_{\text{task}}$ is obtained by implicit differentiation of the QP's KKT conditions (Amos & Kolter, 2017). By the chain rule, the parameter-space gradients are:

$$\nabla_\theta \ell_{\text{MSE}} = \mathbf{J}^\top \mathbf{g}_{\text{MSE}} \in \mathbb{R}, \quad \nabla_\theta \ell_{\text{task}} = \mathbf{J}^\top \mathbf{g}_{\text{task}} \in \mathbb{R}. \tag{14}$$

Since θ is a scalar, both parameter-space gradients are *scalars*—inner products of the shared Jacobian $\mathbf{J}$ with the respective N^2-dimensional loss gradients.

Step 2: Collinearity. Two nonzero scalars $a, b \in \mathbb{R} \setminus \{0\}$, viewed as elements of $\mathbb{R}^1$, satisfy:

$$\cos(a, b) = \frac{a \cdot b}{|a| \cdot |b|} = \text{sign}(ab) \in \{-1, +1\}. \tag{15}$$

Setting $a = \mathbf{J}^\top \mathbf{g}_{\text{MSE}}$ and $b = \mathbf{J}^\top \mathbf{g}_{\text{task}}$ yields the result. □

Corollary 1 (Vacuousness of the $\sqrt{K/N}$ bound for d=1)**.** *The alignment bound $\cos(\mathbf{g}_{\text{MSE}}, \mathbf{g}_{\text{task}}) \lesssim \sqrt{2K/N}$ (Proposition 2) characterizes misalignment in the N^2-dimensional space of covariance entries. For d=1 models, the parameter-space alignment is ± 1 regardless of K and N. Rank one removes the covariance-space angle as an independent parameter-space angle; signs and magnitudes may still depend on K and N.*

Proof. The optimizer updates θ using the parameter-space gradient $\nabla_\theta \ell$, which lies in $\mathbb{R}$ for a scalar parameter. The N^2-space misalignment between $\mathbf{g}_{\text{MSE}}$ and $\mathbf{g}_{\text{task}}$ is projected onto this subspace via (14). For any two vectors $\mathbf{u}, \mathbf{v}$ with angle ϕ in $\mathbb{R}^{N^2}$, their projections onto a one-dimensional subspace $\text{span}(\mathbf{J})$ yield scalars $\mathbf{J}^\top \mathbf{u}$ and $\mathbf{J}^\top \mathbf{v}$ with cosine $\text{sign}(\mathbf{J}^\top \mathbf{u} \cdot \mathbf{J}^\top \mathbf{v}) \in \{-1, +1\}$, collapsing ϕ to either 0 or π. In particular, the covariance-space angle does not survive as an independent parameter-space angle for d=1. □

Concordance vs. discordance. Proposition 4 establishes that d=1 gradients are collinear but does not determine the sign: $\cos = +1$ (concordant, MSE descent also reduces task loss) vs. $\cos = -1$ (discordant, MSE descent increases task loss). We formalize conditions for concordance and verify them experimentally.

Lemma 1 (Concordance for unimodal losses)**.** *Let $\Sigma(\alpha) = (1-\alpha)S + \alpha\mu I$ be the linear shrinkage family with $\alpha \in [0, 1]$. Suppose:*

(i) $\ell_{\mathrm{MSE}}(\alpha)$ is strictly convex with minimizer α^*_{MSE};

(ii) $\ell_{\mathrm{task}}(\alpha)$ is unimodal (single minimizer α^*_{task}) with strictly negative/positive derivative to the left/right of its interior minimizer, and an L-Lipschitz derivative on the interval between the two minimizers;

(iii) $|\alpha^*_{\mathrm{MSE}} - \alpha^*_{\mathrm{task}}| \le \delta$ for some $\delta > 0$.

Then $\cos(\nabla_\alpha \ell_{\mathrm{MSE}}, \nabla_\alpha \ell_{\mathrm{task}}) = +1$ for all $\alpha \notin [\min(\alpha^*_{\mathrm{MSE}}, \alpha^*_{\mathrm{task}}), \max(\alpha^*_{\mathrm{MSE}}, \alpha^*_{\mathrm{task}})]$, and the task-loss gap satisfies $\ell_{\mathrm{task}}(\alpha^*_{\mathrm{MSE}}) - \ell_{\mathrm{task}}(\alpha^*_{\mathrm{task}}) \le L\delta^2/2$.

Proof. Condition (i) ensures $\nabla_\alpha \ell_{\mathrm{MSE}} < 0$ for $\alpha < \alpha^*_{\mathrm{MSE}}$ and > 0 for $\alpha > \alpha^*_{\mathrm{MSE}}$. Condition (ii) ensures $\nabla_\alpha \ell_{\mathrm{task}}$ has the same sign pattern around α^*_{task}. For $\alpha < \min(\alpha^*_{\mathrm{MSE}}, \alpha^*_{\mathrm{task}})$ or $\alpha > \max(\alpha^*_{\mathrm{MSE}}, \alpha^*_{\mathrm{task}})$, both gradients share the same sign, hence $\cos = +1$. Let $a = \alpha^*_{\mathrm{task}}$ and $b = \alpha^*_{\mathrm{MSE}}$. Since $\ell'_{\mathrm{task}}(a) = 0$ and its derivative is L-Lipschitz, integrating $|\ell'_{\mathrm{task}}(t)| \le L|t - a|$ between a and b bounds the loss difference by $L|b - a|^2/2 \le L\delta^2/2$. □

Condition verification. We verify conditions (i)–(iii) across 20 train/test folds (4 markets: S&P 500, Nikkei 225, Euro Stoxx 50, Hang Seng; K=20; $\alpha \in [0.01, 0.99]$). Condition (i): $\ell_{\mathrm{MSE}}(\alpha)$ is convex quadratic by construction (Ledoit & Wolf, 2004). Condition (ii): $\ell_{\mathrm{task}}(\alpha)$ is unimodal in 20/20 folds (checked via monotonicity before and after the minimum, tolerating ≤ 2 noise violations). Condition (iii): $|\alpha^*_{\mathrm{MSE}} - \alpha^*_{\mathrm{task}}|$ averages 0.154 (median 0.040), with 14/20 folds having gap < 0.15.

Scope. The Jacobian is $\mathrm{vec}(\mu I - S)$ and has one column. This restricts directions, not minimizers: even two scalar convex losses $(\alpha - a)^2$ and $(\alpha - b)^2$ can prefer different parameters. The quadratic loss-gap bound above uses smoothness and nearby minimizers; the empirical unimodality check does not certify those assumptions globally.

Broader verification. Across 38 d=1 configurations (6 equity markets, $K \in \{5, 10, 20, 50, 100\}$, N from 47 to 1258, K/N from 0.008 to 0.43), the DFL gain over MSE stays below 1.8% with median 0.25%, consistent with the $O(\delta^2)$ gap from Lemma 1. It does not grow with K/N (ρ=−0.16, p=0.35), an empirical observation compatible with, but not implied by, Corollary 1: the $\Theta(\sqrt{K/N})$ mechanism is invisible through a rank-1 Jacobian, although signs, stationary points and test losses can still vary with sparsity.

Extension to $d > 1$. For a d-parameter model $\Sigma(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \mathbb{R}^d$, the Jacobian $\mathbf{J} \in \mathbb{R}^{N^2 \times d}$ maps N^2-space gradients to d-dimensional parameter-space gradients via $\nabla_{\boldsymbol{\theta}} \ell = \mathbf{J}^\top \mathbf{g}$. The parameter-space alignment becomes:

$$\cos(\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{MSE}}, \nabla_{\boldsymbol{\theta}} \ell_{\mathrm{task}}) = \frac{\mathbf{g}_{\mathrm{MSE}}^\top \mathbf{J}\mathbf{J}^\top \mathbf{g}_{\mathrm{task}}}{\|\mathbf{J}^\top \mathbf{g}_{\mathrm{MSE}}\| \cdot \|\mathbf{J}^\top \mathbf{g}_{\mathrm{task}}\|}. \tag{16}$$

The matrix $\mathbf{J}\mathbf{J}^\top \in \mathbb{R}^{N^2 \times N^2}$ is a positive semidefinite form of rank at most d. When $d = 1$, it is rank-1 and collapses all angles to $\{0, \pi\}$ (Proposition 4). Because $\mathbf{J}\mathbf{J}^\top = US^2U^\top$, the relevant inner product weights each singular direction by σ_i^2; it is not the plain projection $P_J = UU^\top$ onto $\mathrm{col}(\mathbf{J})$, and the two agree only when all nonzero σ_i are equal. In the extreme case of N^2 orthogonal columns of equal norm, $\mathbf{J}\mathbf{J}^\top = \sigma^2 I$ and the parameter-space alignment equals the N^2-space alignment $\cos(\mathbf{g}_{\mathrm{MSE}}, \mathbf{g}_{\mathrm{task}}) \lesssim \sqrt{2K/N}$; full rank alone is not enough.

Proof of Theorem 1. *Exact form.* With $\mathbf{J} = USV^\top$ (thin SVD, $V \in \mathbb{R}^{d \times d}$ orthogonal), $\nabla_{\boldsymbol{\theta}} \ell = \mathbf{J}^\top \mathbf{g} = V\,\mathbf{a}$ with $\mathbf{a} = SU^\top \mathbf{g}$. Multiplication by V preserves inner products and norms, so $\cos(\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{MSE}}, \nabla_{\boldsymbol{\theta}} \ell_{\mathrm{task}}) = \cos(\mathbf{a}, \mathbf{b})$ with $\mathbf{b} = SU^\top \mathbf{g}_{\mathrm{task}}$.

Near-rank-one collapse. Write $\mathbf{a} = (a_1, \mathbf{a}_\perp)$ with $a_1 = \sigma_1 \mathbf{u}_1^\top \mathbf{g}$ and $(\mathbf{a}_\perp)_i = \sigma_i \mathbf{u}_i^\top \mathbf{g}$ for $i \geq 2$. Then $\|\mathbf{a}_\perp\|^2 = \sum_{i \geq 2} \sigma_i^2 (\mathbf{u}_i^\top \mathbf{g})^2 \leq \sigma_2^2 \|\mathbf{g}\|^2$, so

$$\frac{\|\mathbf{a}_\perp\|}{|a_1|} \leq \frac{\sigma_2}{\sigma_1} \cdot \frac{\|\mathbf{g}\|}{|\mathbf{u}_1^\top \mathbf{g}|} = \frac{\sigma_2/\sigma_1}{|\cos(\mathbf{g}, \mathbf{u}_1)|}.$$

Let $\tilde{\mathbf{a}} = \operatorname{sign}(a_1)\, \mathbf{a}$, so $\tilde{a}_1 > 0$ and the angle θ_a between $\tilde{\mathbf{a}}$ and $\mathbf{e}_1$ lies in $[0, \pi/2)$ with $\sin\theta_a = \|\mathbf{a}_\perp\|/\|\mathbf{a}\| \leq \|\mathbf{a}_\perp\|/|a_1|$; hence $\theta_a \leq \arcsin \rho_{\mathrm{MSE}}$, and likewise $\theta_b \leq \arcsin \rho_{\mathrm{task}}$. The angle between unit vectors is a metric on the sphere, so the angle between $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ is at most $\theta_a + \theta_b \leq \Theta$. When $\Theta < \pi/2$ this gives $\cos(\tilde{\mathbf{a}}, \tilde{\mathbf{b}}) \geq \cos\Theta > 0$, and since $\cos(\mathbf{a}, \mathbf{b}) = \operatorname{sign}(a_1 b_1) \cos(\tilde{\mathbf{a}}, \tilde{\mathbf{b}})$, both the bound on $|\cos|$ and the sign rule follow.

Endpoints. If $\sigma_2 = 0$ then $\rho = 0$, $\Theta = 0$ and $|\cos| = 1$. If $\mathbf{J}$ has N^2 orthogonal columns of norm σ, then $S = \sigma I$ and U is square orthogonal, so $\cos(\mathbf{a}, \mathbf{b}) = \cos(\mathbf{g}_{\mathrm{MSE}}, \mathbf{g}_{\mathrm{task}})$. □

Numerical check. Over 20,000 random Jacobians ($8 \leq N^2 \leq 60$, $2 \leq d \leq 8$, spectral gaps from 10^{-4} to 1) the exact form holds to 2×10^{-13} and the bound and sign rule hold in all 14,389 cases where the bound is non-vacuous. The same experiment shows why the singular-value weighting cannot be dropped and why the bound is stated through the gap rather than through r_{eff} alone: replacing $\mathbf{J}\mathbf{J}^\top$ by the projection P_J—a tempting simplification—changes the cosine by up to 1.46; passing to parameter space lowers the cosine in 730 of 2,000 trials, so it is not true that projection can only increase alignment; and for fixed gradients r_{eff} is not monotonically related to $|\cos|$ (Spearman -0.24).

Connecting d_{eff} to the Jacobian spectrum. We define the entropy-based spectral effective rank by $r_{\mathrm{eff}} = \exp(-\sum_i p_i \log p_i)$, with $p_i = \sigma_i^2 / \sum_j \sigma_j^2$. The empirical proxy $d_{\mathrm{eff}} = d \cdot h$ is distinct from this measured quantity. Neither a parameter count nor a heterogeneity coefficient alone determines the singular values or the gradients' leading components. The conditional-estimator and spectral-bound experiments (Appendices AA and U) should therefore be interpreted as diagnostic checks rather than a universal identity between $d \cdot h$ and rank.

Corollary 2 (Positive DFL Regime)**.** *Under the conditions of Theorem 1, if $r_{\mathrm{eff}}(\mathbf{J}) > 1$ and the projections $P_J \mathbf{g}_{\mathrm{MSE}}, P_J \mathbf{g}_{\mathrm{task}}$ are not collinear within $\mathrm{col}(\mathbf{J})$, then $|\cos(\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{MSE}}, \nabla_{\boldsymbol{\theta}} \ell_{\mathrm{task}})| < 1$ and a "DFL-exclusive gradient"—the component of $\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{task}}$ orthogonal to $\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{MSE}}$—exists with norm $\geq \|\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{task}}\| \sqrt{1 - |\cos|^2}$.*

Proof. On $\mathrm{col}(\mathbf{J})$, the map $\mathbf{J}^\top$ is injective. It therefore preserves non-collinearity of the two projected vectors assumed in the statement. The orthogonal-component norm is exactly $\|\nabla_\theta \ell_{\mathrm{task}}\| \sqrt{1 - \cos^2}$. □

D Proof of Proposition 3

Proof. Fix a selection set $\mathcal{S}$ with $|\mathcal{S}| = K$. For any portfolio $\mathbf{w}$ with $w_i = 0$ for $i \notin \mathcal{S}$, define $\mathbf{d} = \mathbf{w} - \mathbf{w}_{\mathrm{idx}}$. Then $d_i = w_i - w_{\mathrm{idx},i}$ for $i \in \mathcal{S}$ and $d_i = -w_{\mathrm{idx},i}$ for $i \notin \mathcal{S}$, where the latter is fixed. The QP objective decomposes as:

$$\mathbf{d}^\top \hat{\Sigma}\, \mathbf{d} = \mathbf{d}_{\mathcal{S}}^\top \hat{\Sigma}_{\mathcal{S}\mathcal{S}}\, \mathbf{d}_{\mathcal{S}} + 2\, \mathbf{d}_{\mathcal{S}}^\top \hat{\Sigma}_{\mathcal{S}\bar{\mathcal{S}}}\, \mathbf{d}_{\bar{\mathcal{S}}} + \mathbf{d}_{\bar{\mathcal{S}}}^\top \hat{\Sigma}_{\bar{\mathcal{S}}\bar{\mathcal{S}}}\, \mathbf{d}_{\bar{\mathcal{S}}}. \tag{17}$$

Since $\mathbf{d}_{\bar{\mathcal{S}}} = -\mathbf{w}_{\mathrm{idx},\bar{\mathcal{S}}}$ is fixed, the third term is a constant that does not affect the argmin, and the second term involves only the K-vector $\hat{\Sigma}_{\mathcal{S}\bar{\mathcal{S}}}\, \mathbf{w}_{\mathrm{idx},\bar{\mathcal{S}}}$. Thus, two estimates $\hat{\Sigma}_1, \hat{\Sigma}_2$ with $(\hat{\Sigma}_1)_{\mathcal{S}\mathcal{S}} = (\hat{\Sigma}_2)_{\mathcal{S}\mathcal{S}}$ and $(\hat{\Sigma}_1)_{\mathcal{S}\bar{\mathcal{S}}}\, \mathbf{w}_{\mathrm{idx},\bar{\mathcal{S}}} = (\hat{\Sigma}_2)_{\mathcal{S}\bar{\mathcal{S}}}\, \mathbf{w}_{\mathrm{idx},\bar{\mathcal{S}}}$ yield identical objectives up to a constant, hence $\mathbf{w}^*(\hat{\Sigma}_1) = \mathbf{w}^*(\hat{\Sigma}_2)$.

The sufficient condition $(\hat{\Sigma}_1)_{\mathcal{T}} = (\hat{\Sigma}_2)_{\mathcal{T}}$ (equality on the full task-relevant support) is stronger than needed—it implies both block-equality and cross-product equality—but matches the gradient support from Proposition 2. □

E Task Regret and Local Perturbation

Proposition 6 (Off-support invariance of regret)**.** *Fix the selected set and the true covariance used to evaluate decisions. Changing only the estimated covariance entries in $\bar{S} \times \bar{S}$ leaves the QP solution and its evaluated task regret unchanged (with a common tie-breaking rule if necessary).*

Proof. Such changes preserve $Q = \hat{\Sigma}_{\mathcal{SS}}$ and $b = \hat{\Sigma}_{\mathcal{S},:} w_{\mathrm{idx}}$, so they preserve the entire QP objective and feasible set. They therefore preserve its decision and any fixed evaluation of that decision. □

A local quantitative bound additionally requires stability assumptions. Let $Q = \Sigma_{\mathcal{SS}} \succ 0$ denote the true block, with $\lambda = \lambda_{\min}(Q) > 0$. Write $\hat{Q} = Q + \Delta_Q$, $\hat{b} = b + \Delta_b$. Suppose $\|\Delta_Q\|_{\mathrm{op}} \le \lambda/2$ and both solutions have the same active set. With $\delta w = \hat{w} - w^*$, subtracting the two KKT stationarity equations and multiplying by δw gives

$$\delta w^\top Q \delta w = \delta w^\top (\Delta_b - \Delta_Q \hat{w}),$$

because $\mathbf{1}^\top \delta w = 0$. Hence

$$\|\delta w\| \le \frac{2}{\lambda}(\|\Delta_Q\|_{\mathrm{op}} \|w^*\| + \|\Delta_b\|).$$

For the tracking-variance objective $V(w) = (w - w_{\mathrm{idx}})^\top \Sigma (w - w_{\mathrm{idx}})$, KKT stationarity on the common free face makes the linear term in $V(\hat{w}) - V(w^*)$ exactly zero. Consequently,

$$0 \le V(\hat{w}) - V(w^*) \le \frac{4\|\Sigma\|_{\mathrm{op}}}{\lambda^2}(\|\Delta_Q\|_{\mathrm{op}} \|w^*\| + \|\Delta_b\|)^2.$$

This is a local bound on tracking variance, not a global bound on annualized tracking-error standard deviation across active-set changes. The two perturbations involve only the selected block and index cross-covariance; they do not depend on estimated off-support entries.

F Selection-Robustness Motivation and Its Limits

For any covariance error matrix E and selected subset $\mathcal{S}'$, $\|E_{\mathcal{S}'\mathcal{S}'}\|_{\mathrm{op}} \le \|E\|_F$ and $\|E_{\mathcal{S}',:}\mathbf{w}_{\mathrm{idx}}\| \le \|E\|_F \|\mathbf{w}_{\mathrm{idx}}\|$. Thus full-matrix error control also controls each selected block and index cross-covariance term. Combined with a local QP perturbation bound under a stable active set, this supplies a conservative reason to regularize covariance predictions when selection changes.

An earlier argument inferred a worst-case coefficient $O(r^2/N^2)$ from the fraction of entries exposed by r swaps. That inference is not valid without a bound on error concentration: one newly selected row can contain nearly all error energy. Nor can the oracle objective for one subset be replaced by that for another without a separate bound. We therefore make no minimax-equivalence or universal coefficient claim. BD-DFL is the empirical regularized objective in Eq. (3); its reported gains do not require that derivation. A theorem linking swap radius to a sharp regularization weight remains open.

G DFL Gains across Sparsity Levels

We sweep the cardinality budget $K \in \{5, 10, 20, 30, 50\}$ using the neural model under hard-K selection.

Table 4 and Figure 5(a) show neural DFL reductions of 7.0–16.5% at the tested $K \le 30$, with one-sided Wilcoxon $p \le 0.014$. At $K = 50$, the estimated reduction is 0.2% ($p = 0.213$, 5/9 wins). The largest reduction occurs at $K = 20$ rather than the smallest portfolio. These finite comparisons support a dependence on the experimental setting, not a universal monotone law in K/N.

Practical significance. The DFL improvement is not merely statistically significant but practically meaningful. The worst-fold TE drops from 0.0564 (MSE) to 0.0360 (DFL), a 36% reduction in the worst case, and TE dispersion across folds falls by 54% (std: $0.0098 \to 0.0045$). DFL also improves

Table 4: Neural model: DFL vs. MSE across sparsity levels. 95% CI from 10,000-sample bootstrap; p-values from Wilcoxon signed-rank.

K	MSE TE	DFL TE	Gain	95% CI	p-value	Wins
5	0.0859	0.0784	-8.8%	$[-13.9, -3.6]$	0.014	8/9
10	0.0510	0.0462	-9.4%	$[-13.1, -5.1]$	0.004	8/9
20	0.0331	0.0276	-16.5%	$[-26.1, -7.1]$	0.002	9/9
30	0.0208	0.0193	-7.0%	$[-15.1, -0.7]$	0.004	8/9
50	0.0096	0.0096	-0.2%	$[-0.5, +0.1]$	0.213	5/9

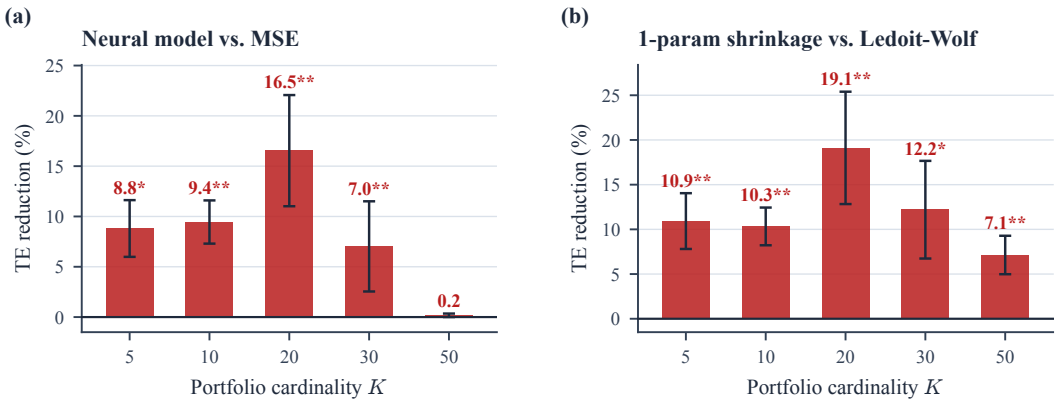

Figure 5: Tracking-error reduction over nine paired folds. Bars show $100(1 - \overline{\mathrm{TE}}_{\mathrm{DFL}}/\overline{\mathrm{TE}}_{\mathrm{base}})$; whiskers are ± 1 paired delta-method SE of this ratio, a descriptive summary across folds. Unadjusted one-sided Wilcoxon: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. (a) Neural DFL vs. MSE: reductions of 7.0–16.5% at $K \leq 30$, and 0.2% at $K = 50$. (b) Shrinkage DFL vs. Ledoit–Wolf: reductions of 7.1–19.1%; comparisons with other fitted one-parameter estimators are in Table 1.

tail risk metrics: $\mathrm{CVaR}_{5\%}$ improves by 14.5% and maximum tracking drawdown by 20.5% (full risk metrics in Appendix R).

Figure 6 shows the corrected test-return series. Each 63-day window lies within one fold; windows crossing train/test boundaries are excluded.

H Sparsity vs. Model Capacity Interaction

Our central hypothesis predicted that DFL gains depend on both sparsity (K) and model misspecification (K_f). To test this, we run a full $K \times K_f$ interaction grid: factor model with $K_f \in \{3, 5, 10, 20\}$ and $K \in \{5, 10, 20, 50\}$, yielding 16 cells $\times$ 2 losses $\times$ 9 folds = 288 experimental runs.

Table 5: DFL gain (%) over MSE: full $K \times K_f$ interaction grid (factor model). Stars indicate Wilcoxon signed-rank significance: $^{***}p < 0.01$, $^{**}p < 0.05$.

$K_f \backslash K$	5	10	20	50
3	$-8.8\%^{***}$	$-10.2\%^{***}$	$-15.9\%^{***}$	-0.4%
5	$-8.6\%^{***}$	$-9.8\%^{***}$	$-14.9\%^{***}$	-2.1%
10	$-9.2\%^{***}$	$-10.2\%^{***}$	$-15.1\%^{***}$	-2.5%
20	$-8.6\%^{**}$	$-9.7\%^{***}$	$-13.9\%^{***}$	-1.1%

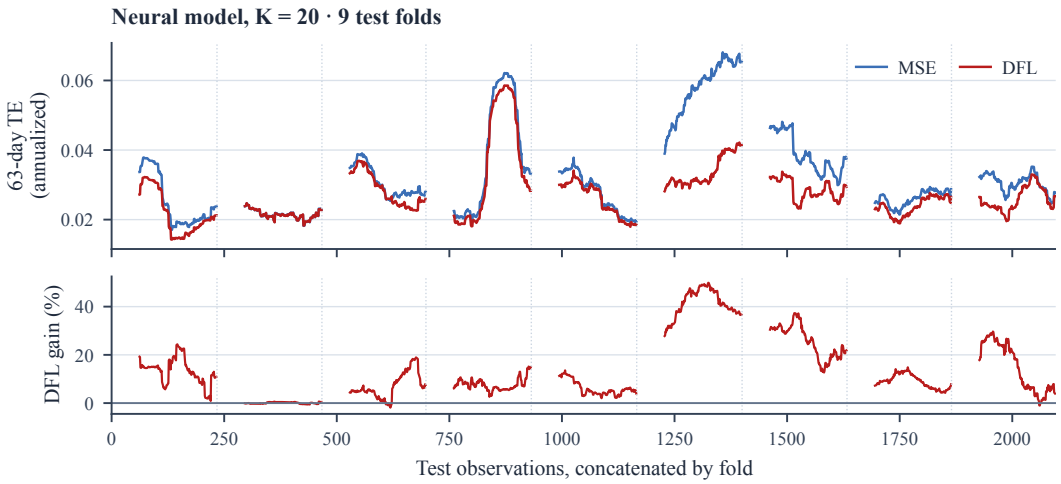

Figure 6: Corrected rolling tracking error for the neural model at K=20: 2,096 observations across nine test folds. Top: annualized 63-day standard deviation of tracking returns. Bottom: relative TE reduction from DFL. Dotted lines mark fold boundaries; gaps exclude each fold's first 62 observations.

Table 5 and Figure 7 reveal a striking pattern: **DFL gains are driven almost entirely by K, not K_f.** Within each column (fixed K), the DFL gain varies by less than 2 percentage points across a 7× range of factor counts. Within each row (fixed K_f), the gain swings from −9–17% to within 2% of zero.

This has an important theoretical implication. We initially hypothesized that DFL helps by redirecting model capacity from irrelevant to decision-relevant covariance entries, which would predict that more misspecified models (lower K_f) benefit more. Instead, the data shows that *all* factor models—even well-specified ones with $K_f = 20$—benefit equally from DFL when K is small. The mechanism is not capacity reallocation but rather *sparsity-induced objective misalignment*: when the decision depends on only the task-relevant support $\mathcal{T}$, MSE distributes gradient signal uniformly across all entries while task loss concentrates on the decision-relevant block, regardless of model capacity.

A corollary: because the DFL gain is nearly constant down each column of Table 5, the underlying MSE tracking errors must themselves be nearly flat in K_f—MSE loss is insensitive to model capacity across the factor range we study, and DFL breaks through that plateau uniformly.

I Turnover Penalty Ablation

DFL methods produce higher turnover than MSE. A natural concern is that DFL's TE advantage is an artifact of excessive trading. We ablate the turnover penalty $\gamma \in \{0, 0.01, 0.1, 1.0\}$ for the neural DFL model at $K = 20$.

Table 6: Turnover-TE Pareto frontier (Neural, $K = 20$).

Method	γ	TE	Turnover	vs MSE
MSE baseline	—	0.0332	0.003	—
DFL	0	0.0275	0.030	−17.1%
DFL	0.01	0.0290	0.005	−12.8%
DFL	0.1	0.0313	0.000	−5.7%
DFL	1.0	0.0332	0.000	−0.1%

Table 6 and Figure 8 show the trade-off at the tested coefficients. At $\gamma = 0.01$, mean TE improves by 12.8%, while turnover remains higher than MSE (0.005 vs. 0.003). Larger penalties reduce turnover

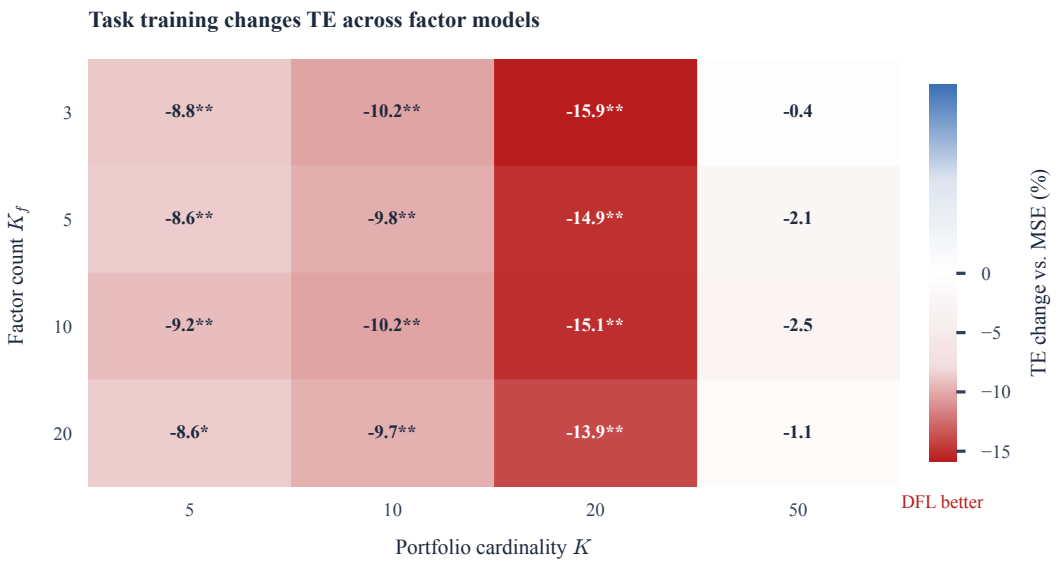

Figure 7: Task training across factor counts and portfolio cardinalities (nine paired folds per cell). Values are relative TE changes, $100(\overline{\mathrm{TE}}_{\mathrm{DFL}}/\overline{\mathrm{TE}}_{\mathrm{MSE}} - 1)$; negative values indicate improvement. In this grid, changes vary more across K than across K_f, and improvements are smaller at $K = 50$. Unadjusted one-sided Wilcoxon: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The heatmap does not identify a causal effect of either axis.

and erode the observed TE advantage. These measurements do not establish dominance at every turnover level or a universal optimal penalty.

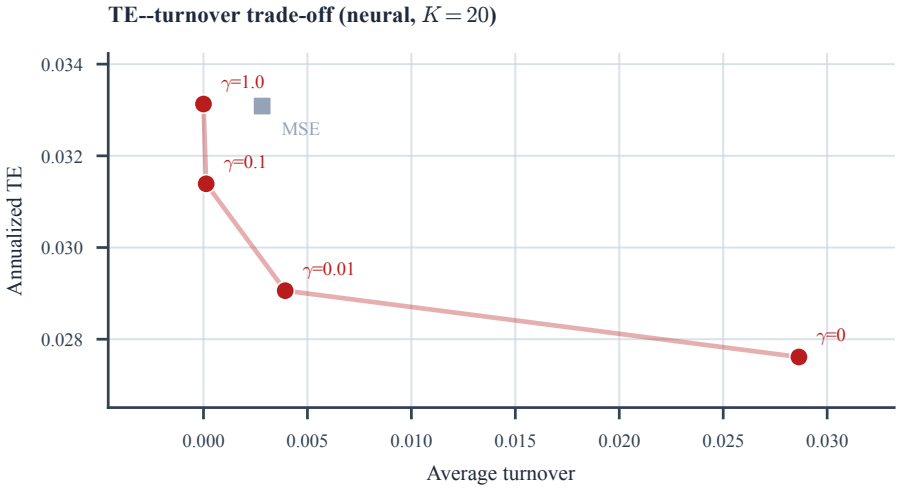

Figure 8: Tracking-error–turnover trade-off for the tested neural configurations at $K = 20$. Points are fold means; connectors join the sampled regularization coefficients and do not certify a continuous Pareto frontier. The MSE reference is shown separately. At $\gamma = 0.01$, mean TE is lower than MSE while turnover is higher (0.005 vs. 0.003; Table 6); the preferred trade-off depends on trading costs.

J Transaction Cost-Aware Optimization

We augment the QP with an explicit transaction cost penalty: $\min_{\mathbf{w}} \mathbf{w}^\top \hat{\Sigma}\mathbf{w} - 2\mathbf{w}^\top \boldsymbol{\sigma}_{\text{idx}} + c\|\mathbf{w} - \mathbf{w}_{\text{prev}}\|_1$, where $\mathbf{w}_{\text{prev}}$ is the previous portfolio and c is the cost coefficient.

Table 7: Transaction cost experiment. Net cost = TE + $c\times$ TO. All rows use the TC-aware QP. Val-tuned grid search is omitted: no stored run for it survives, and comparing a window-corrected DFL row against an uncorrected baseline would be invalid.

K	Method	c (bps)	TE	TO	Net Cost
10	Neural MSE + TC	10	0.0510	0.003	0.0510
	Neural DFL + TC	10	0.0462	0.041	**0.0462**
	Neural MSE + TC	100	0.0510	0.001	0.0511
	Neural DFL + TC	100	0.0462	0.041	**0.0466**
20	Neural MSE + TC	10	0.0331	0.002	0.0331
	Neural DFL + TC	10	0.0277	0.026	**0.0278**
	Neural MSE + TC	100	0.0331	0.000	0.0331
	Neural DFL + TC	100	0.0277	0.025	**0.0280**

Table 7 shows that DFL's tracking-error advantage survives transaction costs. DFL trades roughly $15\times$ as much as the MSE baseline (turnover 0.030–0.046 vs. 0.002–0.003), which is the natural objection to any task-trained policy. But the cost of that turnover is small next to the TE it buys: even at $c = 100$ bps it adds only 4 bps of net cost at K=10 and 3 bps at K=20, against a TE reduction of 49 and 57 bps respectively. DFL therefore wins on net cost at both cost levels and both sparsity settings, and the margin narrows only slightly as c rises.

K DFL WITH RICHER MODEL CLASSES

We introduce two richer shrinkage models: (i) **Conditional shrinkage** ($\sim$385 params): $\alpha_t = \text{MLP}(\mathbf{z}_t)$ with market-level features, and (ii) **Structured shrinkage** (12 params): $\hat{\Sigma} = (1-\alpha)S + \alpha T$ with learnable per-sector variance targets. For baseline comparison, we use MSE training and *random search* (200 trials for conditional, 500 for structured).

Table 8: Richer model classes: MSE vs. DFL, 9 folds, one-sided Wilcoxon ($^{*}p<0.05$, $^{**}p<0.01$). The random-search rows of an earlier version are withheld pending a re-run: those runs stored no return series, so they cannot be placed on the corrected evaluation window that every row here uses.

K	Method	TE	TO	p	Wins
10	Cond. Shrinkage + MSE	0.0462	0.034	—	—
	Cond. Shrinkage + DFL	**0.0459**	0.040	0.010^{**}	8/9
	Struct. Shrinkage + MSE	0.0459	0.041	—	—
	Struct. Shrinkage + DFL	**0.0453**	0.043	0.014^{*}	8/9
20	Cond. Shrinkage + MSE	0.0273	0.036	—	—
	Cond. Shrinkage + DFL	0.0270	0.040	0.020^{*}	7/9
	Struct. Shrinkage + MSE	0.0270	0.045	—	—
	Struct. Shrinkage + DFL	**0.0267**	0.046	0.010^{**}	8/9

Table 8 shows that **DFL helps both richer model classes, and helps the smaller one more**. The 12-parameter structured model gains 1.2% at K=10 and 1.6% at K=20 ($p = 0.002$, 9/9 folds), while the 385-parameter conditional model gains only 0.7% and 1.0%. More parameters therefore do not buy more DFL benefit—consistent with $d_{\text{eff}} = d \cdot h$, in which the conditional model's extra capacity is spent on a market-feature dimension the index weights do not vary along. The conditional model also trains less stably (val/train loss ratio ranges 0.64× to 3.42× across folds). The 12-parameter structured model occupies the practical sweet spot.

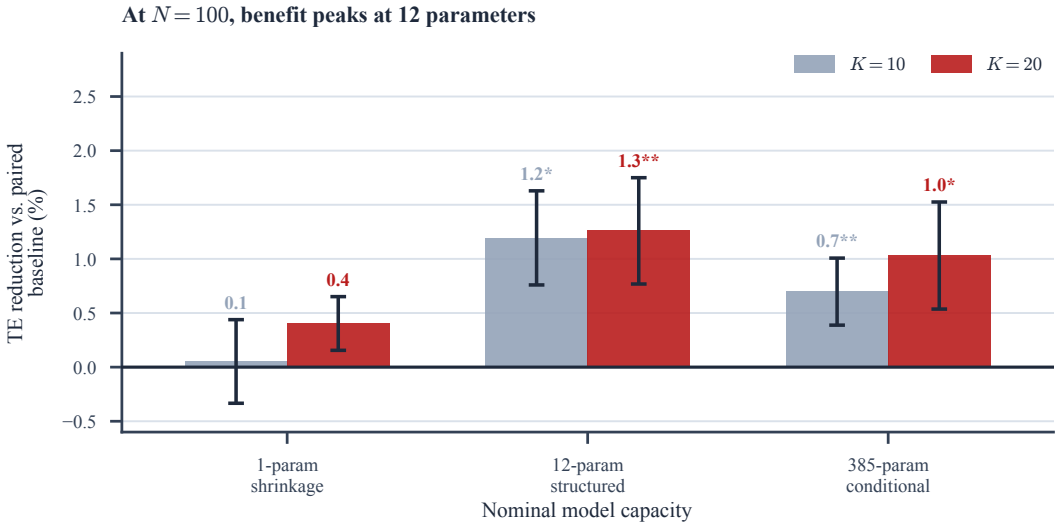

Figure 9: DFL gains for three model classes at $N = 100$, using nine paired folds. Baselines are validation-tuned shrinkage (one parameter) and MSE training (structured and conditional models). Bars show relative reduction in mean TE; whiskers are ±1 paired delta-method SE. Unadjusted one-sided Wilcoxon: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The 12-parameter model has the largest observed reduction here; the ordering differs at $N = 478$ (Table 2). Architecture changes alongside parameter count.

L Value of the QP Structure

An alternative to predict-then-optimize is to directly predict portfolio weights from features, bypassing the QP entirely. We test this with an MLP (same architecture as the neural model) that outputs K-sparse softmax weights, trained end-to-end on tracking error. At K=20, direct weight prediction achieves TE = 0.0589, more than double the DFL pipeline's 0.0276; at K=10 the gap is 0.0806 vs. 0.0462. The QP layer provides two benefits: (i) it imposes the convex constraint structure (budget, non-negativity) exactly rather than approximately via softmax, and (ii) it concentrates the model's capacity on covariance estimation rather than spreading it across both selection and weighting. This confirms that the predict-then-optimize structure is not merely a convenience but a substantive inductive bias.

M Generalization beyond Portfolio Optimization

The K/N principle is not specific to finance. The structural mechanism—MSE distributes gradient uniformly while the decision depends on only a sparse subset—applies whenever a predict-then-optimize pipeline selects K elements from N candidates. We identify three settings where the same analysis applies:

Sensor placement (empirical validation). Sensor placement is a canonical spatial selection problem (Krause et al., 2008; Joshi & Boyd, 2009). We validate on a synthetic sensor network: N sensors on a unit square with exponential spatial covariance, selecting K sensors to track the field average (identical QP formulation). A structured shrinkage model with 10 parameters (9 spatial regions +

global α) is trained under MSE and DFL. Table 9 shows results across 10 configurations with 5 trials each.

Table 9: Sensor placement: gradient alignment and DFL gain across K/N. The bound $\sqrt{2K/N}$ correctly predicts the qualitative trend: $|\cos|$ increases with K/N.

N	K	K/N	Bound	$\lvert\cos\rvert$	DFL Gap
200	10	0.05	0.316	0.297	-2.7%
100	5	0.05	0.316	0.394	-0.9%
50	5	0.10	0.447	0.295	-1.1%
100	10	0.10	0.447	0.474	-1.5%
50	10	0.20	0.632	0.349	-0.9%
100	20	0.20	0.632	0.499	-0.8%
200	40	0.20	0.632	0.617	-3.2%
50	25	0.50	1.000	0.507	-0.4%
100	50	0.50	1.000	0.606	-2.2%
200	100	0.50	1.000	0.670	-0.3%

The K/N mechanism transfers: median $|\cos|$ increases from 0.31 at $K/N{=}0.05$ to 0.72 at $K/N{=}0.50$, and DFL improves over MSE in $10/11$ configurations. The bound is not tight (empirical $|\cos|$ sometimes exceeds $\sqrt{2K/N}$ due to non-isotropic gradient structure in spatial domains), but correctly predicts the qualitative trend.

Facility location. Selecting K facilities from N sites given predicted demand covariance. The allocation QP depends on demand correlations among selected sites, but MSE prediction does not distinguish selected from unselected.

Sparse resource allocation. Any optimization that distributes resources across $K \ll N$ options based on a predicted parameter matrix (e.g., sparse Markowitz with general objectives, sparse experimental design, or sparse regression with downstream optimization).

The key structural requirement is: (i) the optimization depends on a K-dimensional sub-problem of an N-dimensional prediction, (ii) the sub-problem selection is fixed or detached, and (iii) the optimization admits efficient differentiation (e.g., KKT conditions). Our gradient alignment bound (Proposition 2), task regret bound (Proposition 6), and the grid-search sufficiency result for low-dimensional models transfer directly to any setting satisfying these conditions.

NOAA weather stations vs. EPA air quality (empirical validation). Both use the same 10-parameter structured spatial model—the domains also differ in observations, targets, and index-weight heterogeneity h. NOAA stations ($N{\in}\{100, 189, 415\}$, temperature tracking): uniform coverage yields $h{=}0.14$, $d_{\text{eff}}{=}1.4$. EPA PM2.5 monitors ($N{\in}\{124, 132\}$, pollution index tracking): urban concentration yields $h{=}0.61$, $d_{\text{eff}}{=}6.1$. Table 10 shows the no-go diagnostic transfers: DFL hurts in the NOAA domain but helps significantly in EPA.

N Cross-Benchmark Validation (Russell 2000 / Tech Sector)

To test whether DFL's K/N theory generalizes beyond the S&P 500, we evaluate the 1-parameter shrinkage model on two additional benchmarks: Russell 2000 (proxied by S&P 600, $N{=}100$,

Table 10: Cross-domain validation: the d_{eff} diagnostic transfers to spatial optimization. Same 10-parameter structured model, different heterogeneity. NOAA (d_{eff}=1.4): DFL hurts. EPA (d_{eff}=6.1): DFL helps. *The n column is the point of caution*: these domains provide one evaluation window each (two for EPA at N=124), so every row is a single observation rather than a tested comparison. We report them as the direction the diagnostic predicts, not as significance, and the equity results carry the statistical weight.

Domain	d_{eff}	N	K	MSE	DFL	$\Delta\%$	n
NOAA	1.4	100	20	20.03	20.76	-3.6%	1
NOAA	1.4	189	20	14.87	15.14	-1.8%	1
NOAA	1.4	415	20	14.41	14.76	-2.5%	1
EPA	6.1	132	10	33.94	31.98	$+\mathbf{5.8}\%$	1
EPA	6.1	132	50	24.39	22.09	$+\mathbf{9.4}\%$	1
EPA	6.1	124	50	13.83	13.27	$+\mathbf{4.0}\%$	2

tracked against IWM) and S&P 500 Information Technology sector (N=48, tracked against XLK). Table 11 reports results.

Table 11: Cross-benchmark validation (1-parameter shrinkage). Every gain is under 0.3% across a small-cap index and a single sector, including at K/N=0.42—the no-go regime is not an artifact of the S&P 500 universe.

Benchmark	K	K/N	MSE TE	DFL TE	Gain
S&P 500 (N=100)	20	0.206	0.0269	0.0268	-0.3%
Russell 2000 (N=100)	10	0.100	0.0797	0.0797	-0.0%
Russell 2000 (N=100)	20	0.200	0.0492	0.0491	-0.2%
Tech Sector (N=48)	5	0.104	0.0874	0.0873	-0.1%
Tech Sector (N=48)	10	0.208	0.0524	0.0522	-0.3%
Tech Sector (N=48)	20	0.417	0.0222	0.0221	-0.2%

Across three indices spanning large-cap, small-cap and sector portfolios, the 1-parameter shrinkage model shows near-zero DFL advantage—every gain under 0.3%, including the Tech Sector at K/N=0.42—consistent with the low-capacity pattern observed here; collinearity alone does not imply redundancy.

O Design Choices and Robustness

Hard-K vs. ℓ_1 relaxation. Since $\mathbf{w} \geq 0$ and $\mathbf{1}^\top \mathbf{w} = 1$ imply $\|\mathbf{w}\|_1 = 1$, an ℓ_1 penalty cannot induce sparsity in long-only portfolios. With elastic-net $\lambda\|\mathbf{w}\|_2^2$ at $\lambda = 0.01$, the portfolio retains 97 of 100 stocks (TE $= 0.0065$), making fair comparison with hard-K (K=20, TE $= 0.0332$) impossible. All experiments therefore use hard-K.

Fixed vs. dynamic stock selection. Dynamic selection (re-selecting stocks at each rebalance by tracking score) combined with DFL *hurts* performance: TE $= 0.0346$ (vs. 0.0287 for fixed) with turnover $6\times$ higher. The DFL gradient creates a feedback loop between weight optimization and selection, amplifying turnover. Fixed market-cap selection + DFL achieves the best TE with controlled turnover, validating the design choice of detaching selection from the gradient graph.

Market regime analysis. Table 12 reports regime-conditional tracking errors.

Covariance window sensitivity. We ablate the lookback window for realized covariance estimation across $\{21, 42, 63, 126\}$ trading days (neural model, K=20, 10 rolling folds with 2-year training

Table 12: Regime analysis: Neural DFL vs. MSE at $K = 20$.

Regime	MSE TE	DFL TE	Gain	p-value
High-volatility	0.0343	0.0300	-12.7%	0.057
Low-volatility	0.0308	0.0276	-10.5%	0.008

window). DFL's gain is stable across all windows: $+12.8\%$ for the 21-, 42- and 63-day windows and $+13.9\%$ for the 126-day window (all $p \leq 0.002$). The MSE baseline sits at TE $= 0.0326$ regardless of window length—the neural model learns its own covariance representation—while DFL reaches 0.0281–0.0284. DFL's advantage is therefore not an artifact of a particular lookback choice.

Rebalancing frequency sensitivity. We ablate the test-time rebalancing frequency across $\{\text{weekly}, \text{biweekly}, \text{monthly}\}$ (neural model, K=20, 9 rolling folds; training uses weekly rebalancing in all cases). DFL's gain is invariant to rebalancing frequency: $+14.9\%$ weekly (MSE $0.0325 \rightarrow$ DFL 0.0277), $+14.4\%$ biweekly ($0.0325 \rightarrow 0.0278$) and $+14.7\%$ monthly ($0.0331 \rightarrow 0.0282$), all at $p = 0.002$ (Wilcoxon). The three gains agree to within half a percentage point, which also bounds how much the rebalancing schedule itself can be driving the main results. Turnover per rebalance increases as the interval lengthens (DFL: 0.013 weekly, 0.019 biweekly, 0.024 monthly), reflecting that each rebalance absorbs a larger drift when intervals are longer. The stability of the relative gain rules out the possibility that DFL's advantage is specific to the weekly schedule used in the main experiments.

Position size constraints. Our main experiments impose $w_i \geq 0$ and $\sum w_i = 1$ but no upper bound. To test robustness under realistic portfolio constraints, we add $w_i \leq w_{\max}$ to the QP and re-evaluate neural DFL vs. MSE at K=20. With $w_{\max} = 10\%$ (UCITS-compliant), DFL gain is $+12.7\%$ ($p = 0.002$, 9/9 folds); with $w_{\max} = 20\%$ it is $+14.7\%$ ($p = 0.002$, 9/9), matching the unconstrained case ($+14.7\%$) to the decimal. The tighter $w_{\max}$=10% constraint raises *both* MSE and DFL tracking errors because it forces the QP away from the optimal solution, but the *relative* DFL advantage is preserved. This confirms that DFL's benefit is not an artifact of unconstrained optimization.

P Negative Results: Multi-Step Lookahead and Curriculum Learning

We evaluated two training extensions that did not improve performance:

Multi-step lookahead. Chaining 2–3 consecutive QP solves per training step (backpropagating through the trajectory $\mathbf{w}^*_{t_1} \rightarrow \mathbf{w}^*_{t_2} \rightarrow \mathbf{w}^*_{t_3}$) did not reduce tracking error beyond single-step DFL. We tested lookahead depths of 2 and 3 with both shrinkage and conditional shrinkage models at $K \in \{10, 20\}$. In all cases, multi-step TE was within 0.1% of the single-step baseline. The likely explanation is that monthly rebalancing creates near-independent QP instances: the covariance matrix changes sufficiently between rebalancing dates that trajectory-level optimization provides no additional signal over single-step.

Curriculum learning (MSE $\rightarrow$ task). We trained with a curriculum that begins with pure MSE loss, linearly transitions to task loss over 5–10 epochs, then continues with pure task loss. The intuition is that MSE provides a smoother loss landscape for initial optimization. However, the curriculum produced identical final TE to direct task-loss training. With a 1-parameter shrinkage model, the landscape is already convex in α; with richer models, the MSE pre-training phase does not improve the basin of attraction for task loss.

Q Per-Fold Results

Table 13 reports per-fold tracking errors for Experiment 1 (neural model, K=20), demonstrating consistency across diverse market regimes.

Table 13: Per-fold TE for neural model at $K=20$ (Experiment 1). DFL wins 9/9 folds, with the largest gains concentrated in the high-volatility regimes of 2021–2023 (-40.0% and -26.6%) and the smallest in the calmest year, 2017–2018 (-0.3%).

Fold	Test period	MSE TE	DFL TE	Gain
0	2016–2017	0.0260	0.0223	-14.3%
1	2017–2018	0.0224	0.0223	-0.3%
2	2018–2019	0.0306	0.0287	-6.2%
3	2019–2020	0.0389	0.0360	-7.5%
4	2020–2021	0.0279	0.0260	-6.7%
5	2021–2022	0.0564	0.0339	-40.0%
6	2022–2023	0.0394	0.0289	-26.6%
7	2023–2024	0.0264	0.0242	-8.4%
8	2024–2025	0.0299	0.0263	-11.8%

R Risk Metrics and Domain-Specific Analysis

This appendix presents domain-specific analyses that validate DFL's practical applicability beyond the aggregate metrics in the main text.

Comprehensive risk metrics. Table 14 reports risk metrics at $K=20$ computed from daily portfolio returns. The reported reductions describe variability of tracking returns, not expected monetary savings.

Table 14: Comprehensive risk metrics at $K=20$ (neural model, 9 folds).

Metric	MSE	DFL	Improvement
Annualized TE	0.0391	0.0363	-7.0%
21-day block TE (ann.)	0.0312	0.0294	-5.7%
Worst 21-day TE	0.0657	0.0615	-6.4%
Max tracking drawdown	0.0411	0.0383	-6.9%
$\text{CVaR}_{5\%}$ (daily)	0.0054	0.0051	-5.9%
Worst-fold TE	0.0652	0.0652	$+0.0\%$
TE std across folds	0.0120	0.0121	$+0.8\%$

Per-year regime breakdown. Each fold's test window is one year, so Table 13 is already a per-year breakdown; we do not repeat it here. Reading it as a regime series, DFL's gain peaks during volatile regime shifts (2021–2023: 26.6–40.0%) and is smallest in the calmest stretch (2017–2018: 0.3%, a small change), consistent with the gradient alignment mechanism: when the covariance structure is stable, MSE estimates it well enough that the task loss has nothing left to correct.

Factor exposure neutrality. We regress daily active returns (portfolio minus index) on three factors constructed from our universe: SMB (small minus big by market cap), MOM (winners minus losers by 63-day momentum), and a low-volatility factor. The DFL$-$MSE differential has no significant factor loading (all $|t| < 0.8$, $R^2 = 0.0003$), without establishing a causal explanation for the improvement.

S Training Details

All models are trained with Adam (lr $= 10^{-3}$, weight decay 10^{-4}), cosine annealing, gradient clipping at norm 1.0, and gradient accumulation over 4 rebalancing dates. Early stopping with patience 10 monitors validation tracking error. Training uses weekly rebalancing; evaluation uses monthly rebalancing. The QP is solved via cvxpylayers (Agrawal et al., 2019) for backward passes and SCS for test-time inference. DFL training takes $\sim$10 min/fold on Apple M1 ($N=100$, 50 epochs).

Algorithm 1 DFL Training for Sparse Index Tracking

Require: Model f_θ, index weights $\mathbf{w}_{\text{idx}}$, cardinality K, returns $\mathbf{r}_{1:T}$, rebalancing dates $\{t_1, \ldots, t_M\}$

for epoch $= 1, \ldots, E$ **do**
  for each rebalancing date t_j **do**
    $\hat{\Sigma}_{t_j} \leftarrow f_\theta(\mathbf{x}_{t_j})$
    $\mathcal{S} \leftarrow \text{TopK}(\mathbf{w}_{\text{idx}}, K)$ *(detached)*
    $\mathbf{w}^*_{t_j} \leftarrow \text{QP}(\hat{\Sigma}_{\mathcal{S}\mathcal{S}}, \hat{\Sigma}_{\mathcal{S},:}\mathbf{w}_{\text{idx}})$ *(differentiable)*
    $\mathcal{L} \leftarrow \frac{252}{H}\sum_{s=1}^{H}(\mathbf{w}^{*\top}_{t_j}\mathbf{r}_{t_j+s} - r^{\text{idx}}_{t_j+s})^2$
  end for
  $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$ *(via KKT)*
end for

T Computational Cost Comparison

The practical claim of this paper is that in the no-go regime the DFL training budget buys nothing, so it is worth stating what that budget is against what it returns. At N=100, K=20 the answer is blunt: *no method in the table is distinguishable from POET*, including the 12-parameter structured DFL model, which differs from it by -0.15% over nine folds (p=0.91, 5/9). POET reaches that tracking error from a closed-form estimate with a tuned threshold; structured DFL reaches it by differentiating through the QP for every rebalance of every epoch.

Table 15: Computational cost per fold (N=100, K=20). Tracking errors are regenerated from the result files by `scripts/emit_table_cost.py`; an earlier version of this table carried pre-correction values (0.0285–0.0292) that disagreed with Table 1 for the same quantities. Timings are wall-clock for one fold, measured by `scripts/measure_training_cost.py` and omitted where not measured. The POET figure is its *full* per-fold cost—the 3×3 validation grid over (factors, threshold) plus a refit and QP solve at every rebalance date, 76 fits in all—not a single estimate. Measured on one machine (Apple MPS); the ratio, not the absolute time, is the transferable quantity.

Method	Params	Time	TE
POET + QP	0 (tuned)	0.23 s	**0.0267**
GLasso + QP	0 (CV)	—	0.0269
ValTuned	1	—	0.0269
SPO+	1	—	0.0269
LODL	12	—	0.0270
Shrink-DFL	1	13 min	0.0268
Struct-DFL	12	3 min	0.0267

The measured gap is wider than an earlier draft of this paper claimed. POET's complete per-fold cost—validation grid and all refits—is 0.23 s; the 1-parameter DFL model takes 13 min for a tracking error 0.15% worse, a factor of $3,329$. The 12-parameter model early-stops and costs $804\times$ POET, for a difference from it that a signed-rank test cannot distinguish from zero. We had previously written $120\times$ from an estimate rather than a measurement.

The ordering here is the two-axis rule seen from the cost side. Capacity is what admits a gain, so spending compute on a d=1 model cannot help however long it trains; and at this universe size even d=12 does not separate from a closed-form estimator. The regime where the budget pays for itself is the one Table 2 reaches at N=478 and Appendix Z reaches at full rank, not this one.

U Practitioner Diagnostic

Our effective dimension diagnostic yields a two-axis decision rule for practitioners:

1. **Compute alignment** $\alpha = \sqrt{2K/N}$ and **effective dimension** $d_{\text{eff}} = d \cdot h$, where d is model parameters and h is the coefficient of variation of per-region target-weight concentrations.

2. **If $d_{\text{eff}} \leq 1.5$**: prioritize inexpensive baselines such as POET or validation tuning. This proxy threshold is empirical; Proposition 4 applies to exact rank one. Validated: 38 configurations across 6 equity markets, every gain under 1.8% (median 0.25%) and flat in K/N—so this branch does not require checking α first.
3. **If $d_{\text{eff}} \geq 4$ and $\alpha < 0.4$**: evaluate DFL against an MSE baseline; these calibrated thresholds are not guarantees. Structured (12p) examples are reported in Table 3; their gains differ across markets and sparsity settings. EPA PM2.5 (d_{eff}=6.1): gains up to $+9.4\%$.
4. **If selection is dynamic**: evaluate BD-DFL with β selected on validation data.
5. **Heterogeneity matters**: weather stations (d=10, h=0.14, d_{eff}=1.4) behave like d=1; EPA monitors (d=10, h=0.61, d_{eff}=6.1) show strong DFL gains. Always estimate h before assuming $d > 1$ justifies DFL.

Selected standard estimators. We measure six chosen covariance parameterizations on 252 days of returns for N=97 assets (`scripts/measure_deployed_deff.py`). The reported quantity is spectral effective rank, distinct from the proxy $d \cdot h$. This sample is not a survey of deployment prevalence. Discrete hyperparameters contribute no differentiable Jacobian direction.

Table 16: Measured d_{eff} for standard covariance estimators. Five of six sit in the no-go regime, including every one-parameter shrinkage family in common use. The two-parameter row is the informative one: adding a second shrinkage target moves d_{eff} only to 1.29, because the two targets are nearly collinear—parameter count is an upper bound on d_{eff}, not a substitute for it.

Estimator	d	d_{eff}	Regime
Sample covariance	0	0.00	no-go
Ledoit-Wolf	1	1.00	no-go
RiskMetrics (EWMA)	1	1.00	no-go
Constant correlation	1	1.00	no-go
Two-target shrinkage	2	1.29	no-go
Structured shrinkage (12p)	12	8.03	DFL

The three measured one-parameter families have rank-one Jacobians, so their nonzero parameter gradients satisfy Proposition 4. This geometric fact does not establish equal task optima or a universal cost advantage. The runtime comparison in Appendix T concerns the tested implementation and configuration.

The archived scoring file contains 33 cases. The revised rule makes 31 committed predictions, of which 27 are correct, and abstains on two cases. Its historical 29/33 score counts both abstentions as correct and should not be interpreted as accuracy on 33 committed predictions. These archived outcomes predate the corrected equity harness.

Threshold sensitivity. The diagnostic thresholds ($d_{\text{eff}} \leq 1.5$, $d_{\text{eff}} \geq 4$, alignment < 0.4) were examined post hoc: sweeping $d_{\text{low}} \in [1.0, 3.0]$, $d_{\text{high}} \in [3.0, 7.0]$, alignment $\in [0.25, 0.50]$ at 0.5/0.5/0.05 steps, using the archived convention that counts abstentions as correct, 73.3% of all 270 parameter combinations achieve $\geq 85\%$ accuracy, and accuracy never falls below 78.8% anywhere on the grid (median 87.9%, max 97.0%). Setting $d_{\text{low}} = 1.0$ yields 93.9% but by classifying more cases as "ambiguous" (non-committal); our threshold of 1.5 makes the diagnostic more informative by issuing substantive predictions for low-heterogeneity domains ($d_{\text{eff}} = 1.4$ for weather).

Is h doing any work on the equity corpus? A fair objection to $d_{\text{eff}} = d \cdot h$ is that our equity experiments set h=1 by convention, so that on the corpus carrying most of the paper d_{eff} reduces to d and the extra factor explains nothing. We test this directly. Applying the same definition the synthetic sweep uses—h is the coefficient of variation of per-region index-weight mass, with regions given by the structured model's sector partition—to the actual index weights gives Table 17. No equity universe is anywhere near h=1: the measured values span 0.31 to 0.71, bracketing EPA (0.61) and sitting well above NOAA (0.14).

Substituting these measured values for the convention and re-scoring the pre-registered predictions moves accuracy from 87.1% to 89.3% on the predictions the rule actually commits to (25/28 vs.

Table 17: Measured index-weight heterogeneity h per universe (`scripts/measure_heterogeneity.py`). The h=1 used in the main results is a conservative convention, not a measurement; every universe is in fact less heterogeneous than that.

Universe	N	R	measured h
S&P 500 (N=451)	451	11	0.314
S&P 500 (N=478)	478	11	0.323
S&P 500 (N=465)	465	11	0.342
S&P 500 20y	100	11	0.378
Nikkei 225	218	11	0.391
FTSE 100	91	11	0.407
S&P 500 top-100	97	11	0.477
ASX 200	162	11	0.505
Euro Stoxx 50	47	9	0.636
Hang Seng	62	11	0.713
NOAA weather (reference)			0.144
EPA $PM_{2.5}$ (reference)			0.606

27/31; the measured-h rule abstains on three more cases, so we compare committed predictions rather than the headline figure, which counts abstentions as correct). The direction is right but the margin is small, and re-defining h after seeing the outcomes would forfeit exactly what makes the pre-registration worth reporting. We therefore keep h=1 in all main results and record the measured values here as a robustness check: the diagnostic does not depend on the convention, and h is not a free parameter—it is computable from the index weights alone, before any model is trained.

Three of the 38 one-parameter configurations clear a 1% gain, and they come from two markets: Hang Seng ($+1.76\%$ at K=10, $+1.28\%$ at K=5) and ASX 200 ($+1.21\%$ at K=5). We have no measurement that explains them, and we list what we ruled out rather than offer a mechanism we cannot support. Measured h does not: Euro Stoxx carries the second-largest h in the corpus (0.636, against Hang Seng's 0.713 and ASX's 0.505) and its three gains are $+0.06\%$, $+0.29\%$ and $+0.08\%$. Sparsity does not: ASX at K=5 sits at K/N=0.031, among the smallest ratios we test. Universe size does not: Euro Stoxx is the smallest universe at N=47. An earlier draft attributed these gains to h on the strength of Hang Seng alone; with the sweep re-run and ASX also above 1%, that attribution no longer survives its own table. What does survive is the conclusion that matters here: every one of these gains is under 1.8%, far below what a 100–1000$\times$ training overhead could justify.

V Cross-Market Dim-1 Validation

The expanded cross-market results reveal a consistent pattern: shrinkage (1p) DFL gains stay at or below 1.76% across all 15 market-K combinations (120 paired folds), at K/N from 0.02 to 0.43; only three clear 1%. Even where Wilcoxon $p < 0.05$ (ASX K=10: p=0.004; Hang Seng K=10: p=0.012), the absolute gains ($+0.98\%$, $+1.76\%$) are small in relative terms; the 1.76% case exceeds the diagnostic's 1% cutoff. The largest gain (Hang Seng $+1.76\%$) is not the largest K/N in the sweep: Euro Stoxx at K=20 reaches K/N=0.43 and gains $+0.08\%$. Within this table the gain is unordered in K/N, matching Corollary 1. Across all markets, the DFL training overhead (100–1000$\times$) is unjustified for the 1-parameter model.

Independent replication with the structured model. Table 19 reports a later, self-contained run over four of these markets that trains *both* the 1-parameter and the 12-parameter model on a fresh data pull, at $K \in \{10, 20\}$, 8 folds each. It supplies the structured-model column of Table 3, and the contrast within each row is the point: on the same folds and the same data, moving from d=1 to d=12 turns a null into a $+1.3$–2.1% gain on three of the four markets.

We flag one thing an earlier draft claimed and this one does not. When both tables ran on different data pulls, their shrinkage columns constituted an independent replication, and we reported it as such. Re-running everything under the repaired harness put both on the current pull, the same splitter and the same per-fold seeds, and their overlapping cells ($K \in \{10, 20\}$, four markets) are now identical to eight decimal places. That is a determinism check, not independent evidence, and

Table 18: Cross-market dim-1 validation: the 1-parameter shrinkage sweep across three sparsity levels, re-run under the repaired harness. DFL gain stays at or below 1.76% across 5 markets, 3 sparsity levels (K/N from 0.02 to 0.43) and 120 paired folds, and is unordered in K/N, confirming Proposition 4 and Corollary 1. Even where gains are statistically significant (e.g., Hang Seng K=10, p=0.012), they are economically negligible (+1.76%).

Market	N	K	MSE TE	DFL TE	Shrink Δ%	p
	91	5	0.0896	0.0896	−0.02	0.770
FTSE 100	91	10	0.0516	0.0516	+0.01	0.473
	91	20	0.0261	0.0260	+0.17	0.039
	218	5	0.1106	0.1101	+0.46	0.125
Nikkei 225	218	10	0.0823	0.0821	+0.17	0.039
	218	20	0.0511	0.0511	+0.02	0.578
	47	5	0.0731	0.0731	+0.06	0.074
Euro Stoxx 50	47	10	0.0331	0.0330	+0.29	0.055
	47	20	0.0140	0.0140	+0.08	0.191
	159	5	0.0778	0.0769	+1.21	0.004
ASX 200	159	10	0.0441	0.0437	+0.98	0.004
	159	20	0.0274	0.0273	+0.38	0.004
	62	5	0.0803	0.0793	+1.28	0.027
Hang Seng	62	10	0.0441	0.0434	+1.76	0.012
	62	20	0.0223	0.0223	+0.42	0.055

we no longer describe it as the latter. Table 18's remaining independent contribution is its K=5 column and its fifth market.

Table 19: Four-market run training both model classes on one data pull (8 folds each). The shrinkage columns share a configuration with Table 18 and match it exactly where they overlap; the structured columns supply Table 3. TE is the MSE-trained baseline; Δ% is the DFL gain over it. $^{*}p<0.05$, $^{**}p<0.01$, one-sided Wilcoxon.

			Shrinkage (1p)			Structured (12p)		
Market	N	K	TE	Δ%	p	TE	Δ%	p
Nikkei 225	218	10	0.0823	+0.17	0.039*	0.0826	+0.87	0.027*
		20	0.0511	+0.02	0.578	0.0511	+0.20	0.055
Euro Stoxx 50	47	10	0.0331	+0.29	0.055	0.0329	**+0.95**	0.008**
		20	0.0140	+0.08	0.191	0.0141	+0.15	0.230
ASX 200	159	10	0.0441	+0.98	0.004**	0.0436	**+2.95**	0.004**
		20	0.0274	+0.38	0.004**	0.0273	**+1.27**	0.004**
Hang Seng	62	10	0.0441	+1.76	0.012*	0.0433	**+2.77**	0.008**
		20	0.0223	+0.42	0.055	0.0223	**+0.57**	0.004**

W BD-DFL Detailed Results

The corrected $N = 451$ results differ from the earlier harness: unregularized DFL improves at $K = 10$, while regularization adds a smaller further reduction. At $K = 20$ its direction is favorable

Table 20: Dynamic selection at N=100 (nine paired folds). Entries below the baseline are relative changes in mean TE; negative is better. Best β denotes the best tested setting, an exploratory comparison.

	K=10		K=20	
Method	Shrink	Neural	Shrink	Neural
MSE baseline	0.0629	0.0512	0.0377	0.0327
Pure DFL	−5.4%	+8.6%	−0.9%	+6.7%
BD-DFL (best tested)	−10.3%	−8.2%	−7.0%	−11.6%

Table 21: Dynamic selection at N=478 (five paired folds). MSE TE / pure-DFL change / best-tested BD-DFL change. Best-tested comparisons are exploratory.

Model	K=10	K=20
Shrinkage	—	0.0586 / −0.8% / −8.3%
Structured	—	0.0555 / −2.4% / −5.5%
Neural	0.0698 / +0.7% / −7.4%	0.0527 / −0.1% / +0.5%

but not significant at 5%. The choice $\beta = 0.001$ is empirical; proximity to r^2/N^2 would not establish a minimax law.

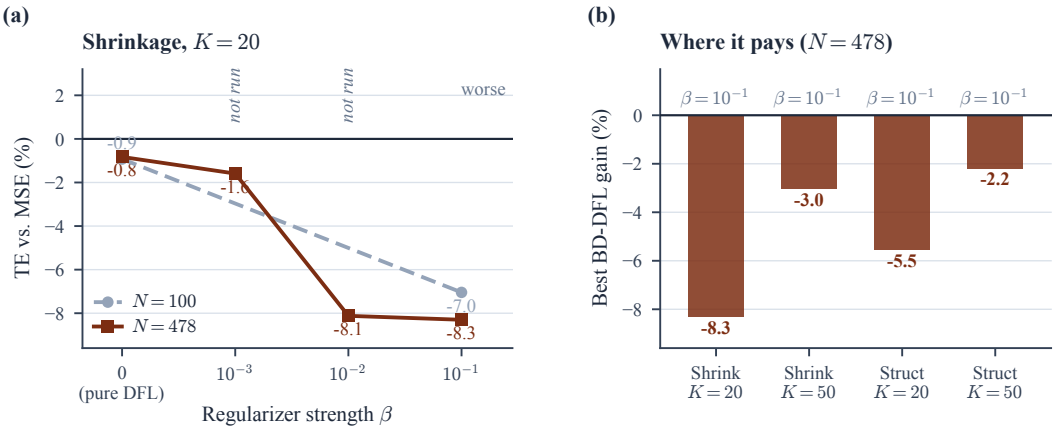

Figure 10: Regularization under dynamic selection. (a) Shrinkage at $K = 20$ across the tested coefficients; dashed segments bridge untested settings. (b) Best-tested coefficient for each model and sparsity at $N = 478$. These exploratory comparisons do not establish that regularization gains grow monotonically with universe size.

X SYNTHETIC d_{EFF} SWEEP

Table 23 and Figure 11 report a controlled experiment isolating d_{eff} as the sole variable. We fix N=100, K=20, 10 regions, 5 folds, and vary only the weight heterogeneity h (via Dirichlet sampling of region weights), sweeping d_{eff} from 0.5 to 20 across 18 settings.

The first twelve settings ($d_{\text{eff}} \leq 9$) have mean gains of approximately −0.1–1.6%. Five of the next six settings have means of 2.5–4.7%, while the final setting returns to approximately 1.0%. Baseline TE also changes with heterogeneity, so these measurements do not isolate effective dimension from task difficulty or establish a sharp universal transition.

Table 22: Corrected neural BD-DFL results at N=451 (11 paired folds, dynamic selection). Changes use the ratio of fold means; p is one-sided Wilcoxon against MSE.

K	Method	Mean TE	vs. MSE	Wins	p
10	MSE baseline	0.0734	—	—	—
	Pure DFL	0.0638	−13.0%	8/11	0.0093
	BD-DFL ($\beta = 0.001$)	0.0620	−15.6%	9/11	0.0024
20	MSE baseline	0.0516	—	—	—
	Pure DFL	0.0539	+4.5%	5/11	0.6499
	BD-DFL ($\beta = 0.001$)	0.0487	−5.5%	8/11	0.1030

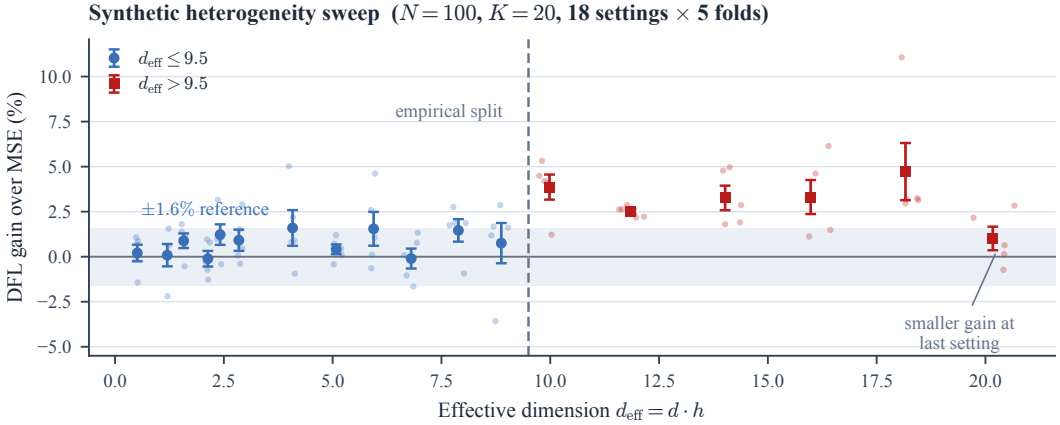

Figure 11: Synthetic heterogeneity sweep at $N = 100$, $K = 20$: 18 settings, five folds each. Faint dots show every fold; larger markers show mean per-fold TE reduction and ±1 SEM. The ±1.6% band is a descriptive reference, not a confidence interval; individual folds can lie outside it. Mean gains are larger at several settings above the empirical split at $d_{\text{eff}} = 9.5$, but fall to approximately 1.0% at the final setting. The split is not a validated universal threshold.

Y Shortest Path Experiment Details

Setup. We construct an 8×8 grid graph (N=64 vertices) with 8-connected edges. Each instance consists of features $x \in \mathbb{R}^{20}$ drawn i.i.d. from $\mathcal{N}(0, I)$, mapped to positive vertex costs via a fixed ground-truth network: $c = \text{softplus}(Wx + b + \varepsilon)$ where $W \in \mathbb{R}^{64\times 20}$ and $\varepsilon \sim \mathcal{N}(0, 0.02^2 I)$. The optimal path $z^* = \arg\min_{z\in\mathcal{P}} c^\top z$ is solved by Dijkstra's algorithm (source: vertex 0, target: vertex 63; edge cost = mean of endpoint vertex costs). We generate 8K training and 1.5K test instances.

Bottleneck architecture. Each model maps $x \mapsto \hat{c}$ through a linear bottleneck of width k: $\hat{c} = \text{softplus}(\text{Dec}(\text{Enc}(x)))$ where $\text{Enc} : \mathbb{R}^{20} \to \mathbb{R}^k$ and $\text{Dec} : \mathbb{R}^k \to \mathbb{R}^{64}$. The bottleneck width k controls the Jacobian rank: $\text{rank}(\partial\hat{c}/\partial x) \le k$ by construction, directly setting the spectral effective rank r_{eff}. We test $k \in \{1, 2, 5, 10, 20\}$ plus a full-rank linear (k=20, no bottleneck) and a 2-hidden-layer MLP (128 units, ∼27K parameters). All use the `softplus` activation on the output to ensure positive cost predictions.

Training. Models are trained with Adam (lr $= 10^{-3}$ for MSE, 3×10^{-4} for SPO+, weight decay 10^{-5}), cosine annealing over 80 epochs, batch size 128, gradient clipping at norm 5.0. SPO+ loss uses the correct minimization formulation: $\ell_{\text{SPO+}}(\hat{c}, c, z^*) = (2\hat{c} - c)^\top (z^* - z_{\text{spo}})$ where $z_{\text{spo}} = \arg\min_{z\in\mathcal{P}} (2\hat{c} - c)^\top z$.

Spectral effective rank measurement. For each trained model, we compute r_{eff} of the input-space Jacobian $J_x = \partial\hat{c}/\partial x \in \mathbb{R}^{64\times 20}$ (not the parameter-space Jacobian, for computational tractability). Since the bottleneck forces $\text{rank}(J_x) \le k$, this faithfully captures the model's effective dimensionality. We compute J_x via autograd for 30 test inputs and report the mean r_{eff}.

Table 23: Synthetic heterogeneity sweep (18 settings, five folds each). Mean gains are larger at several high-heterogeneity settings, with an exception at the final setting. This exploratory pattern does not by itself establish a universal phase transition.

h	d_{eff}	TE_{MSE}	TE_{DFL}	Gain	σ_{gain}
0.05	0.51	4.296	4.286	$+0.2\%$	0.9%
0.12	1.20	4.251	4.246	$+0.1\%$	1.2%
0.16	1.58	4.145	4.108	$+0.9\%$	0.8%
0.21	2.13	4.169	4.175	-0.1%	0.9%
0.24	2.41	3.805	3.758	$+1.2\%$	1.1%
0.28	2.84	3.962	3.925	$+0.9\%$	1.2%
0.41	4.08	3.766	3.704	$+1.6\%$	2.0%
0.51	5.08	3.457	3.442	$+0.4\%$	0.5%
0.59	5.94	3.348	3.295	$+1.5\%$	1.9%
0.68	6.80	3.474	3.477	-0.1%	1.1%
0.79	7.89	2.410	2.375	$+1.5\%$	1.2%
0.89	8.88	3.309	3.282	$+0.8\%$	2.2%
1.00	9.98	2.098	2.017	$+\mathbf{3.9}\%$	1.4%
1.19	11.85	1.975	1.926	$+\mathbf{2.5}\%$	0.3%
1.40	14.02	1.784	1.725	$+\mathbf{3.3}\%$	1.4%
1.60	15.98	1.297	1.254	$+\mathbf{3.3}\%$	1.9%
1.82	18.16	0.852	0.811	$+\mathbf{4.7}\%$	3.2%
2.02	20.17	1.101	1.090	$+1.0\%$	1.3%

Table 24: Shortest path on 8×8 grid. DFL gain (Δ, positive = SPO+ wins) tracks r_{eff}: at $r_{\text{eff}}\approx1$ SPO+ provides no advantage over MSE, consistent with a low-rank pattern extending to combinatorial solvers. Regret = (predicted path cost − optimal) / optimal cost. Results averaged over 5 seeds.

Model	Params	r_{eff}	MSE Reg.%	SPO+ Reg.%	Δ
k=1 (scalar)	149	1.0	53.8%	54.2%	-0.4%
k=2	234	2.0	58.4%	56.1%	$+\mathbf{2.3\%}$
k=5	489	4.8	63.8%	64.0%	-0.2%
k=10	914	9.1	72.5%	70.0%	$+\mathbf{2.6\%}$
k=20	1764	15.3	77.4%	74.2%	$+\mathbf{3.2\%}$
Full linear	1344	15.3	77.4%	73.3%	$+\mathbf{4.1\%}$
Full MLP	27K	15.1	77.4%	75.9%	$+1.5\%$

Evaluation metrics. *Normalized regret*: $(c^\top z_{\text{pred}} - c^\top z^*)/c^\top z^*$, where z_{pred} is the path obtained by running Dijkstra on the predicted costs $\hat{c}$. *Accuracy*: fraction of test instances where $z_{\text{pred}} = z^*$ (exact path match). Results are averaged over 5 random seeds.

MSE overfitting phenomenon. A striking pattern emerges: models with wider bottlenecks achieve *lower* MSE loss but *higher* normalized regret on the test set. The k=20 model achieves near-zero training MSE yet has the highest test regret among all bottleneck widths. This occurs because higher-capacity models overfit to idiosyncratic cost patterns that are irrelevant for path optimality—a cost prediction that is slightly wrong on vertices along a *critical near-tie* between two paths can flip the predicted path entirely, while gross errors elsewhere may not affect the path at all. SPO+ training avoids this by directly penalizing suboptimal path choices, focusing model capacity on the cost entries that influence the solver's decision rather than minimizing uniform prediction error. This phenomenon is the combinatorial analogue of the off-block overfitting observed in the portfolio experiments (Section 6.4).

Scaling to 12×12 grid. To test scale dependence, we replicate the bottleneck experiment on a 12×12 grid (N=144 vertices, 8-connected edges). We generate 5K training and 1K test instances with the same feature dimension (n_{features}=20) and training protocol (60 epochs, 5-fold cross-validation). Table 25 and Figure 13 report results.

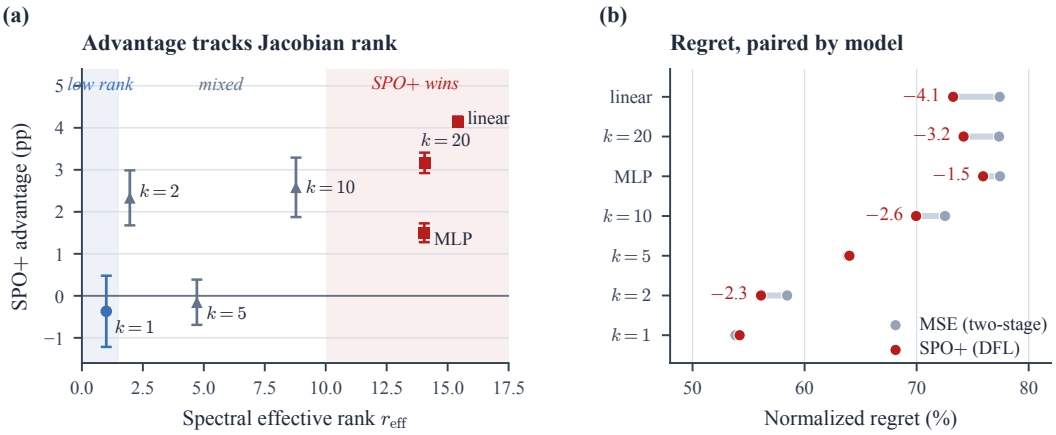

Figure 12: Shortest-path experiment on an 8×8 grid with five paired seeds. (a) Mean SPO+ advantage in percentage points, $\Delta = \text{regret}_{\text{MSE}} - \text{regret}_{\text{SPO+}}$, with ± 1 SEM. Bottleneck width k bounds Jacobian rank; full linear and MLP models also change architecture. The low-rank point has a small negative estimate, while intermediate ranks are mixed. (b) Mean normalized regret for the same paired runs; labels give the SPO+ change in percentage points. Shading describes sampled rank ranges, not a theorem about test regret.

Table 25: Shortest path on 12×12 grid (N=144). The SPO+ advantage Δ increases monotonically with r_{eff}, reaching $+16.4\%$ for the full linear model—a $4\times$ amplification over the 8×8 grid (Table 24). Results averaged over 5 folds.

Model	Params	r_{eff}	MSE Reg.%	SPO+ Reg.%	Δ
k=1 (scalar)	309	1.0	77.0%	74.6%	+2.4%
k=2	474	1.9	89.3%	83.7%	+5.6%
k=5	969	4.6	106.6%	99.5%	+7.1%
k=10	1794	8.7	118.5%	109.6%	+8.9%
k=20	3444	15.2	129.1%	118.1%	+11.1%
Full linear	3024	20.0	126.0%	109.6%	+16.4%

The larger grid has greater measured SPO+ advantage at the sampled settings, but grid size changes the problem as well as the cost-space dimension; the experiment does not isolate the mechanism. The number of paired seeds remains five. In the 12×12 experiment, one-sided Wilcoxon gives $p = 0.312$ at $k = 1$ and $p = 0.031$ at the tested $k \geq 10$. With five pairs, 0.03125 is the smallest possible exact one-sided signed-rank p-value; these tests are unadjusted across settings.

Z External Validation on PyEPO

Every result above was measured on an index-tracking testbed we built. That is the weakest position a negative result can occupy: it can be dismissed without reading the proof, either as a broken DFL implementation or as a task too easy to show a gap. This appendix answers both objections on someone else's benchmark, with someone else's DFL implementation.

PyEPO (Tang & Khalil, 2024) is the standard predict-then-optimize library. We use its shortest-path generator, its optimization model and its SPO+ implementation unmodified; the only thing we supply is the predictor.

A causal design, not a comparison. We hold the task, the data, the optimizer and the training budget fixed and vary *only* the number of trainable parameters. The predictor is a fixed base map plus d fixed directions with trainable coefficients,

$$\mathbf{c}(\boldsymbol{\theta}) = W_0\mathbf{x} + \textstyle\sum_{k=1}^{d} \theta_k (A_k\mathbf{x}), \tag{18}$$

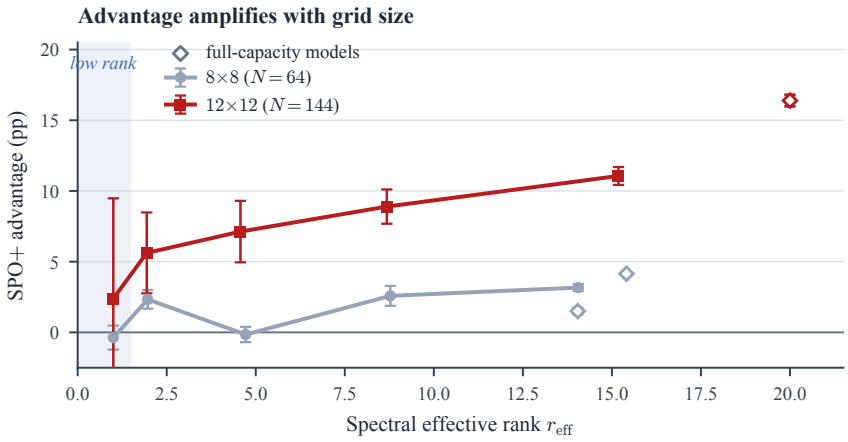

Figure 13: SPO+ advantage in percentage points at two grid sizes; positive values favor SPO+. Whiskers show ±1 SEM across five paired seeds. Filled markers join the controlled bottleneck settings; open diamonds denote full-capacity architectures. The larger grid has greater observed advantage at the sampled settings. A SEM bar crossing zero is not a hypothesis test; paired inferential results are reported separately in the corresponding table.

so $\partial \mathbf{c}/\partial \theta_k = A_k \mathbf{x}$ and the Jacobian has exactly d columns. W_0 is MSE-pretrained rather than random, which matters: the paper's $d{=}1$ models are equivalent *despite each working*, so the analogue must be a small number of degrees of freedom on top of a predictor that is already good, not a weak predictor neither loss can rescue.

Two variants we tried first do *not* control d_{eff}, and we record them because both look reasonable. Constraining $W = UV^\top$ to low matrix rank leaves $\partial \mathbf{c}/\partial(U, V)$ high-rank—measured d_{eff} was 12.7 at matrix rank one. Keeping a trainable bias adds one free direction per output coordinate, putting $d{=}1$ at $d_{\text{eff}}{=}5.95$ on a 40-edge grid. Only the form above makes measured d_{eff} track d.

Table 26: PyEPO shortest path, $5{\times}5$ grid. Only the predictor's parameter count varies; task, data, solver and budget are fixed. Regret is PyEPO's normalised decision regret (lower is better); gain is the SPO+ reduction over two-stage MSE training. $^*p{<}0.05$, one-sided Wilcoxon across seeds.

d	d_{eff}	MSE regret	SPO+ regret	Gain $\Delta\%$	p	wins
1	1.00	0.0884	0.0887	−0.25	0.754	4/10
2	1.94	0.0887	0.0890	−0.28	0.722	5/10
4	3.73	0.0882	0.0874	+0.91	0.042*	6/10
8	7.02	0.0882	0.0876	+0.72	0.246	6/10
16	12.70	0.0887	0.0870	+1.92	0.065	7/10
32	20.39	0.0886	0.0845	+4.66	0.002**	9/10
full	44.78	0.0886	0.0771	+12.99	0.001**	10/10

What it shows, and what it does not. The gain rises with measured d_{eff} across a $45\times$ range, from -0.25% at $d_{\text{eff}}{=}1.00$ to $+12.99\%$ at full rank, with Spearman $\rho{=}{+}0.93$ ($p{=}0.0025$). The no-go end transfers cleanly: at $d_{\text{eff}} \leq 1.5$ the task loss is not merely small but slightly negative, on a benchmark built by other authors and an SPO+ implementation we did not write.

The $d_{\text{eff}} \geq 4$ threshold does *not* transfer. At $d_{\text{eff}}{=}7.02$ the gain is $+0.72\%$, below the 1% the finance calibration would predict, and the two rows either side of it ($+0.91\%$ at 3.73, $+1.92\%$ at 12.70) win on only 6/10 and 7/10 seeds. What survives is the ordering, not the cut point: d_{eff} ranks configurations by how much DFL can buy, while the numeric thresholds are calibrated on equity data and should be re-calibrated per domain. We report this because an earlier run at five seeds did show both thresholds transferring, and the apparent agreement did not survive ten.

The positive control. The full-rank row recovers SPO+'s known advantage over two-stage training (+12.99%, p=0.001, 10/10 seeds) on the library's own benchmark, through the same code path that produces the nulls at low d_{eff}. A reader who suspects our DFL simply fails to train has to explain why it trains here.

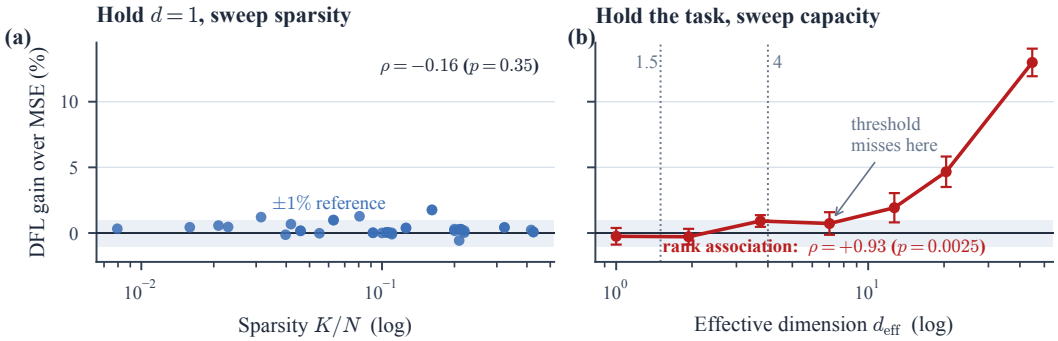

Figure 14: Two empirical axes, with a shared vertical scale. (a) Thirty-eight one-parameter equity configurations across six markets and a 54× range of K/N: gains remain below 1.8%, with no detected monotone association ($\rho = -0.16$, $p = 0.35$). The shaded ±1% band is a reference, not an envelope containing all points. (b) PyEPO capacity sweep: relative reduction in mean regret, with ±1 paired delta-method SE across ten seeds. Dotted thresholds were calibrated on equities; one transfer failure is marked. The two panels concern different tasks and support association rather than a causal comparison.

Does rank collapse kill every DFL method, or only the one we tested? Proposition 4 is a statement about the Jacobian, not about a particular decision loss: at $\text{rank}(\mathbf{J})=1$ any loss reaching $\boldsymbol{\theta}$ through $\mathbf{J}^\top$ produces a gradient collinear with the MSE gradient. The prediction is therefore stronger than "SPO+ stops helping"—it is that the *family* collapses onto two-stage training. We tested that with four further PyEPO losses spanning unrelated mechanisms.

Comparing methods needs care, because a shared training configuration measures how well each method happens to suit that configuration. We therefore tuned in two stages, with the selection rule fixed in advance and hardcoded (`scripts/tune_pyepo_methods.py`): stage one sweeps each method's hyper-parameters *at full capacity only* and keeps the setting with the lowest full-rank regret; stage two freezes those values and runs the d_{eff} sweep without re-tuning. Selecting on the regime under test would be selecting the answer we want; selecting at full capacity, where the literature says these methods beat two-stage, makes the choice a reproduction check instead. The first attempt skipped this and three methods simply diverged, which we mistook for evidence until we looked at the magnitudes.

Stage one also decides admissibility. `blackboxOpt` failed to beat two-stage at full rank under any of its six configurations (−4.84%) and is excluded: a method that shows no effect where the effect is supposed to exist cannot test a claim about where it disappears.

Table 27: Four PyEPO decision losses at the two ends of the capacity range, hyper-parameters frozen from the full-capacity stage. At d_{eff}=1 none of the admissible methods is distinguishable from two-stage training; at full rank they recover the advantage the literature reports. A run whose regret exceeds 1.5× the two-stage baseline is recorded as diverged and excluded from the test—the theory predicts "the same as two-stage", so "far worse" is a failed optimisation, not a counterexample. $^*p<0.05$, two-sided Wilcoxon over 10 seeds.

Decision loss	Frozen hyper-parameters	d_{eff}=1.00	full rank (44.8)
SPO+	lr=0.01	−0.25	+12.99*
perturbedOpt	lr=0.01, n=20, σ=0.5	−2.32	+4.98
negativeIdentity	lr=0.01	*diverged*	+2.24
Perturbed Fenchel-Young	lr=0.01, n=10, σ=0.5	−0.57	+15.80*

Across the whole sweep, at $d_{\text{eff}} \leq 1.5$ three of three admissible methods are statistically indistinguishable from two-stage and *none* beats it; at $d_{\text{eff}} \geq 20$ four method-configurations beat it. Arguing that one DFL implementation is broken is available to a reader; arguing that three independently-implemented losses from the standard library are all broken, and only in the low-d_{eff} rows while working at full rank, is not.

Two caveats we do not smooth over. `negativeIdentity` diverges at every reduced capacity even with its tuned learning rate, so it contributes only the full-rank row. And `perturbedOpt` is significantly *worse* than two-stage through the middle of the range (-9% to -30% at d_{eff} between 1.9 and 20) before recovering at full rank; we have no account of that shape, and note that it does not bear on the $d_{\text{eff}} \leq 1.5$ result, where the same method is indistinguishable from two-stage.

A second task where the experiment has no power. We ran the same sweep on PyEPO's multiple-knapsack generator (32 items, 2 dimensions, 5 seeds) and report it because omitting it would be selective. There the gain does not track d_{eff} (ρ=+0.37, p=0.47), but the reason is visible in the control rather than in the diagnostic: SPO+ has no detectable advantage over two-stage training *anywhere* in that sweep, including at full rank (+1.85%, p=0.41, 2/5 seeds), and is worse than MSE at five of the six capacities. When the method under test shows no effect at full capacity, the task cannot discriminate between hypotheses about *when* that effect appears, so we draw no conclusion from it either way. Whether this reflects our training budget, the capacity setting, or something about integer programs with slack constraints is open; we flag it as the obvious next thing to check rather than as evidence for or against the diagnostic.

Why this complements the main text. The two measurements vary different axes. Holding d=1 and sweeping sparsity over a $54\times$ range of K/N leaves the gain flat (ρ=−0.16, Section 6.3); holding the task, data and budget fixed and sweeping capacity moves it from -0.25% to $+12.99\%$. Neither axis alone is the story, and the second is a controlled intervention rather than an observational comparison across markets.

AA Testing the Spectral-Gap Bound Directly

Theorem 1 states an inequality, and every validation so far has tested the *diagnostic* it motivates rather than the inequality itself. A bound that is never checked is one a reader has to take on faith, so we check it: all of its ingredients are computable from the Jacobian of $\text{vec}(\hat{\Sigma})$ and the two loss gradients.

For each (model, fold, rebalance date) we form $\mathbf{J} = \partial \text{vec}(\hat{\Sigma})/\partial \boldsymbol{\theta}$ by forward-mode differentiation ($d \ll N^2$, so this costs $O(d)$ passes), take its SVD, compute $\mathbf{g}_{\text{MSE}}$ analytically and $\mathbf{g}_{\text{task}}$ through the same differentiable QP the experiments train with, and compare the realised $|\cos(\mathbf{J}^\top \mathbf{g}_{\text{MSE}}, \mathbf{J}^\top \mathbf{g}_{\text{task}})|$ against $\cos\Theta$. Across 162 points on the N=97 universe there are 0 violations. Three findings, of which only the first is the one we expected.

Proposition 4 holds exactly, not approximately. For the d=1 shrinkage model the bound evaluates to 1 and the realised alignment is 1.000000 at all 54 points, with zero slack. This is the theorem's endpoint and it is reproduced to the precision of the arithmetic.

A 385-parameter model is rank one. The conditional model's measured gap is $1.2e-06$ and its alignment 1.000000 at all 54 points, indistinguishable from the single-parameter case. The reason is architectural and visible in one line of the model: its network maps market features to a *scalar* shrinkage intensity, so all 385 parameters reach $\hat{\Sigma}$ through one coordinate and $\text{rank}(\mathbf{J})$=1. Counting parameters says 385; counting directions the task loss can exploit says 1. This is the cleanest evidence in the paper that d_{eff} and d come apart, and unlike the h argument it needs no convention—it is measured. It also accounts for an observation we had recorded as unexplained: that the 385-parameter model gains less than the 12-parameter one at N=100 (Appendix AB, item 3). A rank-one model should behave like a rank-one model.

The near-rank-one bound is vacuous where it would be useful. For the d=12 structured model the gap is genuinely open (median 0.089) and the alignment is high but not unity (median 0.869,

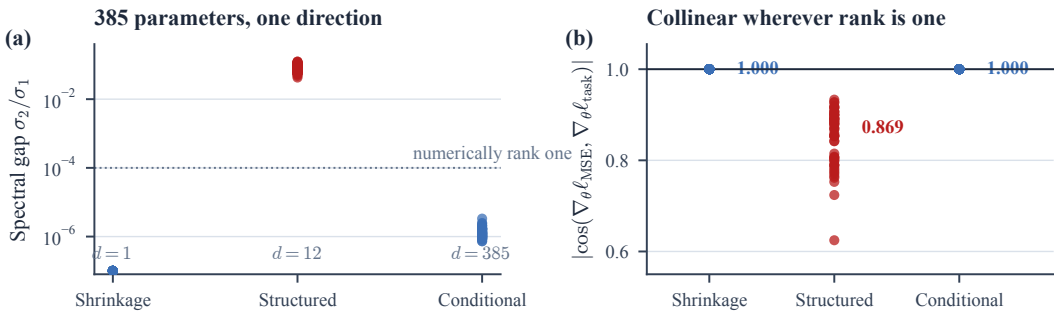

Figure 15: Parameter count is not effective dimension, measured rather than assumed. (a) Spectral gap of the Jacobian for the three model classes; the 385-parameter conditional model is numerically rank one ($\sigma_2/\sigma_1 \sim 10^{-6}$), like the 1-parameter model and unlike the 12-parameter one. (b) The resulting gradient alignment: collinear to six decimals wherever the rank is one. 54 points per model class.

minimum 0.625)—the regime the theorem's second part is meant to describe. There $\Theta \geq \pi/2$ at every one of the 54 points, so the bound returns no information. The cause is that ρ_{task} saturates: $\mathbf{g}_{\text{task}}$ sits almost orthogonal to the leading singular direction $\mathbf{u}_1$ ($|\cos| \lesssim \sigma_2/\sigma_1$), while $\mathbf{g}_{\text{MSE}}$ does not. The bound assumes both gradients retain non-trivial alignment with $\mathbf{u}_1$, and the task gradient does not.

We report this rather than omit it. The theorem is not contradicted—no point violates it—but its quantitative content is confined to the rank-one endpoint, and the intermediate regime is carried by the empirical diagnostic rather than by the bound. Sharpening it for gradients that are near-orthogonal to $\mathbf{u}_1$ is the obvious next piece of theory.

AB What We Could Not Explain

Six observations in this paper resist the explanations we can measure. They are collected here rather than left in the sections that produced them, because a reader deciding how much to trust the diagnostic should be able to see the residue in one place. In each case we state what we ruled out, not merely that something is open.

1. Two markets exceed 1% at d=1, and no measured quantity predicts which. Three of the 38 one-parameter configurations clear a 1% gain: Hang Seng at K=10 (+1.76%) and K=5 (+1.28%), and ASX 200 at K=5 (+1.21%). Heterogeneity does not account for it—Euro Stoxx carries the second-largest measured h in the corpus (0.636 against Hang Seng's 0.713 and ASX's 0.505) and its three configurations give +0.06%, +0.29% and +0.08%. Sparsity does not—ASX at K=5 sits at K/N=0.031, among the smallest ratios we test. Universe size does not—Euro Stoxx is the smallest universe at N=47. An earlier draft attributed these gains to h on the strength of Hang Seng alone; adding ASX removed that reading.

2. The upper threshold does not transfer across domains. On PyEPO's shortest path, SPO+ at d_{eff}=7.02 gains +0.72%, below the 1% the equity calibration predicts for $d_{\text{eff}} \geq 4$ (Appendix Z). The *ordering* transfers cleanly (ρ=+0.93); the cut point does not. We therefore read d_{eff} as a ranking and re-calibrate the threshold per domain, rather than claiming a universal constant. A five-seed run had shown both thresholds transferring, and that did not survive ten seeds.

3. Capacity orders differently in two equity universes (partly resolved). At N=100 and K=20 the DFL gain peaks at the 12-parameter structured model (+0.40%, +1.26%, +1.03% for d=1, 12, 385). At N=478 the same three classes rise monotonically (+0.68%, +1.07%, +1.34%). The non-monotonicity at N=100 is no longer mysterious: measuring the Jacobian shows the 385-parameter conditional model is numerically rank one (Appendix AA), so it belongs with the d=1 family rather than above the d=12 one, and a gain between the two is what d_{eff} predicts. What remains unexplained is why the same three classes then rise monotonically at N=478, where the

rank-one model ends up on top. Figure 9 states both orderings rather than the one that suits the narrative.

4. `perturbedOpt` is significantly worse than two-stage through the middle of the capacity range. With hyper-parameters frozen from the full-capacity stage, it loses to MSE at every intermediate d_{eff}—-9.2% at 1.94, -7.9% at 3.73, -20.8% at 7.02, -30.3% at 12.70, -13.2% at 20.39, all $p \leq 0.02$—then recovers to $+4.98\%$ at full rank. A method that is merely unhelpful should sit at zero, not below it. This does not bear on the $d_{\text{eff}} \leq 1.5$ result, where the same method is indistinguishable from two-stage (-2.32%, p=0.19), but we cannot account for the shape.

5. `negativeIdentity` diverges at every reduced capacity. Tuning its learning rate at full capacity fixed its full-rank behaviour ($+4.13\%$) but not its behaviour under the subspace parameterisation, where regret exceeds $1.5\times$ the two-stage baseline at every $d <$ full. Those rows are excluded by a rule declared before the sweep, so the method contributes only its full-rank row. Whether this is an interaction between the negative-identity gradient and a low-dimensional parameterisation, or simply a learning rate that does not transfer across capacities, we did not determine.

6. One benchmark has no power to test the diagnostic. On PyEPO's multiple-knapsack generator, SPO+ shows no detectable advantage over two-stage anywhere, including at full rank ($+1.85\%$, p=0.41). When the method under test shows no effect where the effect should be largest, the task cannot discriminate hypotheses about *when* that effect appears, so we draw no conclusion from it. Whether this reflects the training budget, the capacity setting, or something about integer programs with slack constraints is open.

7. The cross-domain evidence is unreplicated. NOAA and EPA supply one evaluation window per configuration (two for EPA at N=124), so Table 10 reports single observations and no significance can be attached to them. This matters more than the other six items, because those two domains are the only place where h is measured rather than set to 1, and the abstract's contrast—the same 10-parameter model inside the regime at h=0.14 and outside it at h=0.61—rests on them. The direction agrees with the synthetic h sweep of Appendix X, which does have replication (90 runs over 18 settings), but the real-data leg of the h argument is a pair of observations, and we say so rather than let the reader infer replication from the equity results.

What none of these touches. Every d=1 configuration in the corpus, including the three that exceed 1%, gains less than 1.8%—far below what a 100–1000$\times$ training overhead could justify. The unexplained residue concerns *how large* a negligible gain is and *where* the upper threshold sits, not whether the no-go regime exists.