CORRELATION STRUCTURE GOVERNS SCALARIZATION BIAS IN A GROUP-RELATIVE REINFORCEMENT LEARNING SYSTEM FOR PORTFOLIO OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper introduces Pareto-GRPO, a group-relative reinforcement learning system with verifiable rewards for portfolio rebalancing under four objectives: Sharpe ratio, negative maximum drawdown, diversification ratio, and negative turnover. At each rebalancing step the policy samples a group of candidate weight vectors, scores them with a closed-form auditable objective, and reduces the resulting objective matrix to a per-candidate advantage through one of two interchangeable schemes: a Minkowski-weighted (MW) linear combination, or a one-time computation of an in-group rank-and-crowding-distance statistic adapted from NSGA-II. Working under an equicorrelation asset model, this paper applies a differential-geometric treatment to the mean-variance frontier and a spectral decomposition to the asset covariance matrix. Using this framework, this paper derives a closed-form relationship between average pairwise asset correlation and the curvature of the Sharpe-drawdown trade-off surface. This paper demonstrates that as asset correlation falls, the region of this surface that cannot be reached via linear reduction grows monotonically. This analysis is extended to a finite-sample statement, built from Chernoff-type concentration bounds on a two-component sampling mixture, of what an in-group rank statistic can actually recover of that region. Further results establish an unbiasedness and variance-reduction property of the group-relative baseline, a trust-region bound on the clipped surrogate, and a consistency result for the Bradley-Terry pairwise variant as a maximum-likelihood estimator of the group's latent dominance ranking. Together these results yield a testable law: linear reduction should track rank-based reduction in concentrated, high-correlation markets, and increasingly lag it as the universe diversifies. The law is tested with a walk-forward protocol across six equity universes spanning several continents, using Wilcoxon signed-rank tests with Holm-Bonferroni correction and a Friedman rank test across markets. The empirical pattern matches the theoretical prediction on all six markets (S&P 500 tech subset, NASDAQ 100 subset, NIFTY 50, Russell 1000 Value, Dow 30, STOXX Europe 50): linear reduction leads on the four higher-correlation panels and rank-based reduction leads on the two lowest-correlation panels, though the cross-market Friedman statistic, while directionally consistent, remains underpowered at this sample size. A secondary job-scheduling instance, analyzed via a submodularity argument for the hypervolume indicator, indicates the reduction machinery is not specific to finance.

1 INTRODUCTION

A portfolio manager rebalancing a book of N assets is, whether she names it this way or not, solving a vector optimization problem: return per unit of risk should be high, the worst drawdown shallow, the book diversified, and transaction costs small. These goals share no common currency and routinely conflict. The classical response, due to Markowitz (1952), folds return and variance into a scalar through a risk-aversion coefficient. This is elegant, and it is also lossy in a precise sense developed in Section 3: a linear combination of objectives can only recover solutions on the convex hull of the true Pareto frontier (Das & Dennis, 1997). Whatever lies in a concave fold of the frontier is invisible to a linear method no matter how its weights are tuned.

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.

Reinforcement learning has become a common tool for learning rebalancing policies directly from data, and a smaller literature pairs RL training loops with population-based evolutionary reductions, most commonly NSGA-II (Deb et al., 2002), which in its textbook form maintains a population across generations and can in principle trace the full frontier, including non-convex regions. The system studied here does not implement that textbook algorithm. What is implemented is a group-relative policy optimization loop, in the style popularized for language-model post-training (Shao et al., 2024), in which the "reduction" step performs a single non-dominated sort and crowding-distance computation over one sampled group of $G = 16$ candidate portfolios per rebalancing step, with no generations, no population carried across steps, and no evolutionary variation operator. The label NSGA-II is used throughout only because the ranking rule (non-dominated sorting followed by crowding-distance tie-breaking) is lifted directly from Deb et al. (2002); no claim of a population-based evolutionary search is intended.

This distinction changes what the comparison is actually a comparison of. A textbook NSGA-II run, given enough generations and population, converges to the entire Pareto frontier; a single non-dominated sort over $G = 16$ i.i.d. samples is instead a finite-sample estimator of local frontier structure, bounded by how many samples land near a given region. Section 3 restates the convex-hull limitation of linear scalarization (Theorem 1), derives how the unreachable region's size depends on average pairwise correlation under an equicorrelation model (Proposition 1), sharpens this to an exact curvature functional (Theorem 2), and gives a finite-sample detection result (Proposition 2) with a concentration bound (Theorem 3) for the actual two-component sampler.

1.1 A Correlation-Based Answer

The central theoretical claim is that the shape of the achievable objective region, and hence the severity of the linear reduction's blind spot, is a property of the asset universe's covariance structure, not of the reduction method. Under the equicorrelation model, where every pair of N assets shares correlation ρ, the mean-variance normalizing constant $A(\rho) = \mathbf{1}^\top \Sigma^{-1} \mathbf{1}$ satisfies $A(\rho) = \frac{N}{\sigma^2(1+(N-1)\rho)}$, so the minimum-variance portfolio's variance $1/A(\rho)$ rises linearly in ρ while the frontier's curvature falls as ρ falls. This paper shows that the fraction of the four-objective frontier lying outside the convex hull spanned by any linear weighting is a strictly decreasing function of ρ, and sharpens this to an exact closed form for the frontier's curvature functional. Concentrated, highly correlated markets have nearly convex trade-off surfaces on which linear reduction is already close to optimal; diversified, low-correlation markets have genuinely non-convex surfaces on which rank-based reduction has room to find solutions linear reduction cannot reach, provided the sampled group is large enough to notice.

1.2 Why an RL Framework with Verifiable Rewards

A reward is termed *verifiable* here if it is a closed-form, auditable function of a fixed window of realized market data, never a learned critic acting as an oracle. At each step t, a group of G candidates is scored on $f(w) = (\mathrm{SR}(w), -\mathrm{MDD}(w), \mathrm{DR}(w), -\mathrm{TO}(w)) \in \mathbb{R}^4$. This is what makes the correlation curvature law testable: a learned reward would entangle the frontier's shape with the critic's, and the equicorrelation argument would no longer describe the optimization target. No critic is present; the group-relative baseline plays a critic's role at the cost of G evaluations per step instead of one (Lemma 3).

2 Related Work

Markowitz (1952) established the mean-variance efficient frontier, a one-parameter slice of a Pareto frontier indexed by risk aversion; the Sharpe ratio (Sharpe, 1966) collapses it to a tangency point, and Risk Parity (Qian, 2005) offers an equal-marginal-risk-contribution alternative. NSGA-II (Deb et al., 2002), with its non-dominated sort and crowding-distance operator, is the evolutionary workhorse most often paired with financial objective vectors beyond the mean-variance pair; its successor NSGA-III has been shown to recover more diverse, higher-Sharpe portfolios than classical mean-variance optimization when skewness and kurtosis are added as objectives (Weirstrass et al., 2025). Those systems run the full generational algorithm; the system studied here borrows only its per-generation ranking rule and applies it once, to a single sampled group, per rebalancing step. The closest prior system couples NSGA-II with an RL loop via gray relational coefficients and reports

a Sharpe ratio near 1.92 on a single backtest (Wang et al., 2026); a related line learns the NSGA-II selection operator itself from data (Zhu & Lou, 2025). Neither holds a competing linear reduction fixed against the evolutionary one inside the same pipeline, which is the controlled comparison the correlation curvature law needs to be falsifiable.

The optimizer run here, sampling a group of candidates per state, computing a group-relative advantage, and taking a clipped policy-gradient step with no learned critic, is the group-relative policy optimization (GRPO) family introduced for reasoning-oriented language-model post-training (Shao et al., 2024; Guo et al., 2025). GRPO replaces the value function of proximal policy optimization (Schulman et al., 2017) with the empirical mean (and, in some variants, standard deviation) of rewards within a sampled group, removing the need for a critic network at the cost of requiring G evaluations per update. Here the trade is favorable: computing $f(w)$ for $G = 16$ candidates from an already-realized return window is essentially free relative to a critic forward pass, so GRPO's sample-inefficiency is not the liability it is in a language-model rollout setting.

The dominant RL portfolio paradigm scalarizes to a single reward, usually the Sharpe ratio itself (Jiang et al., 2017; Li et al., 2019). The limitation is structural, not empirical, and is the object of Theorem 1: no linear weighting can reach a Pareto-optimal point strictly inside a concave fold of the frontier (Das & Dennis, 1997), and that literature does not predict, from properties of the asset universe alone, how large the unreachable region will be, a gap this paper's curvature results fill. Recent methodological work emphasizes that walk-forward-validated Sharpe ratios sit substantially below backtest-only claims and documents RL policies collapsing once moved to genuine out-of-sample evaluation; Salotra & Pinsky (2026) makes this case directly for RL-based sector rotation, showing a simple momentum benchmark outperforming a PPO ensemble once evaluated under strict chronological validation with a frozen-policy holdout extension, and this paper's reporting standard, and its expanding/rolling fold structure, follows that precedent. Outside finance, comparative studies of NSGA-II, MOEA/D, and RL-based multi-objective optimizers on routing and scheduling instances confirm that swapping the Pareto reduction while holding the training loop fixed produces measurably different frontier approximations.

3 Theoretical Framework

3.1 The Four-Objective Vector

Let $w \in \Delta^{N-1}$, the simplex over N assets, and let $\{R_t\}$ denote the resulting daily return series. With risk-free rate R_f, the annualized Sharpe ratio is $\mathrm{SR}(w) = \frac{\mathbb{E}[R_p]-R_f}{\sigma_p}\sqrt{252}$ with $\sigma_p = \sqrt{w^\top \Sigma w}$. Maximum drawdown over the value path is $\mathrm{MDD}(w) = \max_t \frac{\max_{s\le t} V_s - V_t}{\max_{s\le t} V_s}$. The diversification ratio of Choueifaty & Coignard (2008), $\mathrm{DR}(w) = \frac{\sum_i w_i \sigma_i}{\sigma_p} \ge 1$, measures how much weighted average single-asset volatility is cancelled by imperfect correlation. Turnover is $\mathrm{TO}_t = \sum_i |w_{i,t} - w_{i,t^-}|$. The vector to be jointly maximized is $f(w) = (\mathrm{SR}(w), -\mathrm{MDD}(w), \mathrm{DR}(w), -\mathrm{TO}(w))$, with dominance $w_1 \prec w_2$ iff $f_k(w_1) \ge f_k(w_2)$ for all k with strict inequality for at least one, and Pareto frontier $\mathcal{P}^* = \{w \in \Delta^{N-1} : \nexists\, w' \prec w\}$.

3.2 Linear Reduction and Its Reachable Set

The Minkowski-weighted (MW) reduction fixes $\lambda \in \Delta^3$ and solves $w^*(\lambda) = \arg\max_w \sum_k \lambda_k f_k(w)$.

Theorem 1 (Reachable set of a linear reduction, after Das & Dennis (1997))**.** *For any λ in the interior of Δ^3, $f(w^*(\lambda)) \in \partial\,\mathrm{conv}(f(\Delta^{N-1}))$. Consequently any Pareto-optimal $w^\dagger$ with $f(w^\dagger) \notin \mathrm{conv}(f(\Delta^{N-1}))$ is unreachable by any choice of λ.*

A full proof is given in Appendix C. This is the classical scalarization-bias result, restated with a full proof because the geometric content, not merely the conclusion, is what the correlation argument below needs.

3.3 Correlation and the Curvature of the Reachable Set

The mean-variance pair inside f is isolated first, since SR and MDD are both driven through $\sigma_p^2 = w^\top \Sigma w$. Under the equicorrelation model every asset has variance σ^2 and every pair correlation $\rho \in (0,1)$, so $\Sigma = \sigma^2[(1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top]$.

Lemma 1 (Spectrum of the equicorrelation matrix)**.** *Σ has exactly two distinct eigenvalues, $\mu_1 = \sigma^2(1+(N-1)\rho)$ with multiplicity 1 on $\mathbf{1}/\sqrt{N}$, and $\mu_2 = \sigma^2(1-\rho)$ with multiplicity $N-1$ on $\mathbf{1}^\perp$. Its condition number $\kappa(\Sigma) = \mu_1/\mu_2$ is strictly increasing in ρ and diverges as $\rho \to 1^-$.*

Proposition 1 (Correlation controls frontier curvature)**.** *Under the equicorrelation model, $A(\rho) = \mathbf{1}^\top \Sigma^{-1}\mathbf{1} = \frac{N}{\sigma^2(1+(N-1)\rho)}$, the global minimum-variance portfolio's variance $V_{\min}(\rho) = 1/A(\rho)$ is strictly increasing in ρ, and the frontier's variance-to-mean curvature $\partial^2\sigma_p^2/\partial m^2 = 1/A(\rho)$ at the minimum-variance point is strictly increasing in $\rho \in (0,1)$.*

Corollary 1 (Non-convexity shrinks as correlation rises)**.** *Because* SR *and* MDD *are monotone transforms of $\sigma_p^2(m)$ along the mean-variance boundary, the Lebesgue measure of $\mathrm{conv}(f(\Delta^{N-1})) \setminus f(\Delta^{N-1})$ is a decreasing function of ρ.*

Promoting this to an exact functional: writing $B = \mathbf{1}^\top\Sigma^{-1}\mu$, $C = \mu^\top\Sigma^{-1}\mu$, $A = A(\rho)$, the variance-minimizing frontier is a hyperbola $\sigma_p^2(m) = \frac{Am^2 - 2Bm + C}{AC - B^2}$ with vertex at $m^\dagger = B/A$, $\sigma_p^2(m^\dagger) = 1/A$ (Lemma 5, Appendix C).

Theorem 2 (Closed-form curvature functional)**.** *The signed curvature of $m \mapsto (\sigma_p(m), m)$ is $\kappa(m) = \frac{|\sigma_p''(m)|}{(1+\sigma_p'(m)^2)^{3/2}}$ with $\sigma_p'(m) = \frac{Am - B}{(AC-B^2)\sigma_p(m)}$, and at the vertex $\kappa(m^\dagger) = A^{3/2}(AC - B^2)^{-1}$. Under the equicorrelation model, holding μ fixed as ρ varies, $\kappa(m^\dagger)$ inherits the strict monotonicity of $A(\rho)$, confirming Corollary 1 with an exact functional form rather than a directional statement at one point.*

Remark 1. *The equicorrelation model is a stylization; real markets have heterogeneous pairwise correlations. It is used to obtain a closed-form, monotone relationship, and is best read as generating a testable directional prediction: on a more concentrated, higher-correlation universe, MW should track the rank-based reduction closely; on a more diversified universe, the rank-based reduction should be able to open a measurable lead. Section 7 tests this by ranking the six market panels on realized average pairwise correlation.*

3.4 From a Frontier-Search Guarantee to a Finite-Sample Statistic

Theorem 1 and Corollary 1 describe what a linear reduction can and cannot reach relative to the *entire* achievable set. A textbook NSGA-II, run for enough generations, approximates this set arbitrarily well. The system documented in Section 4 instead draws one group of $G = 16$ candidates per step and applies non-dominated sorting once, to that group, a meaningfully weaker apparatus, and it is worth stating exactly how much weaker.

Assumption 1. *Near a non-convex indentation of width δ (the fraction of the range of f_1 over which the indentation is Pareto-optimal) and depth κ (the normalized perpendicular distance to the convex-hull chord), a single sample from the current policy lands in the indentation with probability δ, independently across samples from a common sampling component.*

Proposition 2 (Group size and indentation detection)**.** *Under Assumption 1, the probability that a group of G i.i.d. samples contains at least one point inside an indentation of width δ is $P_{\mathrm{detect}}(G,\delta) = 1-(1-\delta)^G$, and the required group size for target detection probability $1-\alpha$ is $G^\dagger(\delta,\alpha) = \lceil \log\alpha / \log(1-\delta) \rceil$. At $G = 16$, the smallest indentation width detected with probability at least 0.95 satisfies $\delta \geq 1 - 0.05^{1/16} \approx 0.174$: indentations narrower than roughly 17% of the local objective range are, on average, invisible to a single group.*

The implemented sampler is in fact a two-component mixture: $G/2$ members drawn from a near-policy, low-noise Dirichlet, and $G/2$ from a wide, exploratory Dirichlet.

Theorem 3 (Detection under the mixture sampler)**.** *Let δ_n, δ_w be the per-draw hitting probabilities of the near and wide components. The exact detection probability is $P^{\mathrm{mix}}_{\mathrm{detect}}(G,\delta_n,\delta_w) = 1 - (1 -$*

Table 1: Pareto-GRPO training loop, a five-component protocol. Only Reduce differs between the MW and rank-based instantiations.

ENCODE(t)	State s_t built from a 60-day lookback: per-asset mean return, volatility, 30-day momentum, and average cross-correlation, a $5N$-dimensional vector.
SAMPLE(s_t)	Feed-forward network with temperature-scaled softmax gives a mean weight vector; $G = 16$ candidates are drawn by Dirichlet perturbation, half from a low-noise near-policy component and half from a wide exploratory component.
REWARD	Advance with the realized return window, evaluate $f(w_i)$ for every candidate, normalize each component with an EMA tracker (momentum 0.99).
REDUCE	MW: $\sum_k \lambda_k \tilde{f}_{i,k}$. NSGA-II: rank by dominance, break ties by crowding distance, reward increasing in rank and crowding. Composite: rank-based selection, MW-scaled reward. Scalar-PPO: fixed-weight linear baseline distinct from MW's weights.
UPDATE(θ)	Group-relative advantage $A_i = (r_i - \bar{r})/(\hat{\sigma}_r + \epsilon)$ (or Bradley-Terry loss for +BT variants), clipped surrogate against an EMA-synced reference policy, Adam step. No critic, no TD target.

$\delta_n)^{G/2}(1-\delta_w)^{G/2}$*, and with* $\bar{\delta} = \frac{1}{2}(\delta_n + \delta_w)$*, the miss probability satisfies the Chernoff-type bound*

$$1 - P^{\text{mix}}_{\text{detect}} \leq \exp\left(-\frac{G\bar{\delta}}{2}\right),$$

which recovers the exponential-in-G decay of Proposition 2 regardless of how $\bar{\delta}$ is split between the two components (Corollary 2, Appendix C).

Proofs appear in Appendix C. Over the many rebalancing steps of a full training run, the effective detection probability compounds; a closed-form model of this compounding is not derived here and is the most direct theoretical extension the present results motivate.

3.5 THE RANK STATISTIC AS A U-STATISTIC OF LOCAL DOMINANCE

Within a group $\{w_1, \ldots, w_G\}$, non-dominated sorting assigns $\text{rank}(w_i) = 1 + |\{w_j \in P : w_j \prec w_i\}|$, with ties broken by crowding distance $d_i = \sum_k \frac{f_k(w_{i+1}) - f_k(w_{i-1})}{f_k^{\max} - f_k^{\min}}$. Because $\text{rank}(w_i) - 1 = \sum_{j \neq i} \mathbb{1}[w_j \prec w_i]$ is a one-sample U-statistic in the kernel $h(w, w') = \mathbb{1}[w' \prec w]$, it has conditional mean $(G-1)q(w_i)$ and variance $(G-1)q(w_i)(1-q(w_i))$ where $q(w_i) = \Pr_{w'}(w' \prec w_i)$ (Lemma 6, Appendix C): once rescaled by $1/(G-1)$, the rank is a consistent, binomial-variance estimator of the population dominance quantile, so increasing G both improves detection of a specific indentation (Proposition 2) and reduces sampling noise in the rank estimate itself.

3.6 CLASSICAL REFERENCE POINTS

Three non-RL baselines anchor the comparison: Markowitz mean-variance optimization with tangency solution $w_{\tan} = \Sigma^{-1}(\mu - R_f \mathbf{1})/\mathbf{1}^\top \Sigma^{-1}(\mu - R_f \mathbf{1})$; Equal Weight, $w_i = 1/N$; and Risk Parity, equalizing total risk contributions $\text{TRC}_i = w_i(\Sigma w)_i/\sigma_p$. Under the equicorrelation model, Risk Parity coincides exactly with Equal Weight (Lemma 7, Appendix C).

4 THE PARETO-GRPO CONSTRUCTION

Pareto-GRPO poses portfolio rebalancing as a multi-objective decision process solved by group-relative policy optimization (Shao et al., 2024), with a verifiable, deterministic reward and a group-empirical advantage baseline rather than a learned value function. Table 1 specifies the five components of the training loop exactly as implemented: state encoding, group sampling, reward computation, reduction, and update. Only REDUCE differs between the MW and rank-based (NSGA-II) runs, which is the controlled substitution the correlation curvature law requires.

4.1 GROUP-RELATIVE ADVANTAGE AND THE CLIPPED SURROGATE

With $\bar{r} = \frac{1}{G}\sum_i r_i$, $\hat{\sigma}_r = \text{std}(\{r_i\})$, the advantage is $A_i = (r_i - \bar{r})/(\hat{\sigma}_r + \epsilon)$.

Lemma 2 (Unbiasedness of the group baseline)**.** *For G i.i.d. group members drawn from the policy's sampling distribution at s_t, $\mathbb{E}[\bar{r} \mid s_t] = \mathbb{E}_{w \sim \Pi_\theta(\cdot|s_t)}[r(w)] =: V(s_t)$, so $\bar{r}$ is an unbiased estimator of exactly the quantity a state-value critic would target.*

Lemma 3 (Variance reduction versus a constant baseline)**.** *Let $\hat{g}_{\text{const}}$ and $\hat{g}_{\text{grp}}$ denote the single-sample policy-gradient estimators with a state-independent constant baseline and with the group-relative baseline, respectively. When $\nabla_\theta \log \pi_\theta(w_i|s_t)$ has approximately constant magnitude across the group and $b_0 = V(s_t)$ is the optimal constant baseline,*

$$\text{Var}[\hat{g}_{\text{grp}}] \approx \text{Var}[\hat{g}_{\text{const}}] \cdot \left(1 - \tfrac{1}{G}\right),$$

i.e. the group-relative estimator recovers a $(1 - 1/G)$ fraction of the variance reduction an oracle constant baseline provides, at zero bias, at $G = 16$ a reduction factor of $15/16$.

With importance ratio $\eta_i(\theta) = \pi_\theta(w_i|s_t)/\pi_{\theta_{\text{ref}}}(w_i|s_t)$, the clipped surrogate is

$$L^{\text{CLIP}}(\theta) = \frac{1}{G} \sum_i \min\left(\eta_i(\theta)A_i,\ \text{clip}(\eta_i(\theta), 1-\epsilon, 1+\epsilon)A_i\right), \quad \epsilon = 0.20.$$

Proposition 3 (Pessimism of the clipped group-relative surrogate)**.** *For every i, θ, $\min(\eta_i(\theta)A_i, \text{clip}(\cdot)A_i) \leq \eta_i(\theta)A_i$, with equality except when $A_i > 0, \eta_i > 1 + \epsilon$ or $A_i < 0, \eta_i < 1 - \epsilon$. Hence L^{CLIP} lower-bounds the unclipped surrogate for every θ, and the two agree, with the same gradient, at $\theta = \theta_{\text{ref}}$.*

Proofs of Lemmas 2 to 3 and Proposition 3 appear in Appendix D. The reference policy is mixed toward the current policy by an EMA, $\theta_{\text{ref}} \leftarrow 0.90\theta_{\text{ref}} + 0.10\theta$, every 21 optimizer steps, and a KL penalty against the reference policy with a coefficient cycling through a fixed schedule is added to the loss; no entropy bonus is present, and exploration is driven entirely by the two-component Dirichlet sampling scheme.

4.2 The Bradley-Terry Pairwise Variant

For the +BT configurations, the group-relative advantage is replaced by a pairwise loss over the group's dominance relation, avoiding collapse of the partial order induced by f into a scalar before it reaches the policy gradient. A soft dominance probability $P_{ij} = \frac{1}{4}\sum_k \sigma\big((\tilde{f}_{i,k} - \tilde{f}_{j,k})/\tau\big)$, $\tau = 0.3$, feeds a Bradley-Terry loss $L^{\text{BT}} = -\frac{1}{|P|}\sum_{(i,j)\in P} P_{ij} \log \sigma(\log \pi_\theta(w_i|s_t) - \log \pi_\theta(w_j|s_t))$ over pairs with $P_{ij} > 0.5$.

Proposition 4 (Population minimizer of the Bradley-Terry loss)**.** *If P_{ij} is replaced by its idealized population counterpart under a Bradley-Terry generative model with latent strengths $u_i^\star$, the population risk is convex in u and is uniquely minimized, up to a shared additive constant, at $u = u^\star$.*

Proof in Appendix D. This clarifies that the +BT gradient step, in the population limit, drives $\log \pi_\theta(w_i|s_t)$ toward differences reproducing the group's pairwise dominance-margin probabilities exactly, fitting a full set of pairwise comparisons rather than a scalar target.

4.3 Objective Normalization and Policy Network

Each objective is normalized by an EMA tracker with momentum 0.99, a biased, geometrically-weighted estimator with an effective averaging window of 199 updates (Lemma 8, Appendix D). The policy network is a three-layer feed-forward network ($256 \to 128 \to 64$, LayerNorm and GELU) projecting to N logits followed by a temperature-scaled softmax; no separate critic network exists.

5 Market Datasets

Six equity universes of ten tickers each are built from daily adjusted closing prices spanning roughly 2014-2026: A (S&P 500 technology subset), C (NASDAQ 100 subset), E (NIFTY 50), D (Russell 1000 Value subset), B (Dow 30), and F (STOXX Europe 50), listed from highest to lowest realized average pairwise correlation $\bar{\rho}$ (full details in Table 4, Appendix A). Dataset A, the technology-concentrated S&P subset, is used for policy training and the full six-strategy ablation; its metrics are train-set metrics, reported separately from the five genuinely independent test panels B through F.

Because the empirical test hinges on ranking panels by $\bar{\rho}$, Appendix D gives a concentration bound (Proposition 5) showing that, for $N = 10$ and $T \approx 2500$ trading days, the standard error of $\bar{\rho}$ is on the order of 0.01-0.02, small relative to the gaps separating the six panels.

Table 2: Full evaluation, Dataset A, train set, ten-seed mean±std, vs. Equal Weight.

Strategy	Sharpe±sd	Sortino	Calmar	AnnRet	MDD
NSGA-II	0.507 ± 0.195	0.627	0.21	16.1%	56.9%
NSGA-II+BT	0.529 ± 0.220	0.653	0.22	16.6%	56.9%
Composite	1.418 ± 0.116	1.858	1.00	33.7%	29.6%
Composite+BT	1.415 ± 0.105	1.858	1.01	33.8%	29.6%
MW	1.259 ± 0.154	1.655	0.71	32.3%	39.9%
Scalar-PPO	1.113 ± 0.083	1.448	0.64	32.0%	43.9%
Markowitz MVO	1.043	1.320	0.67	24.3%	30.5%
Risk Parity	1.026	1.291	0.66	24.2%	30.5%
Equal Weight	1.035	1.309	0.69	25.0%	30.5%

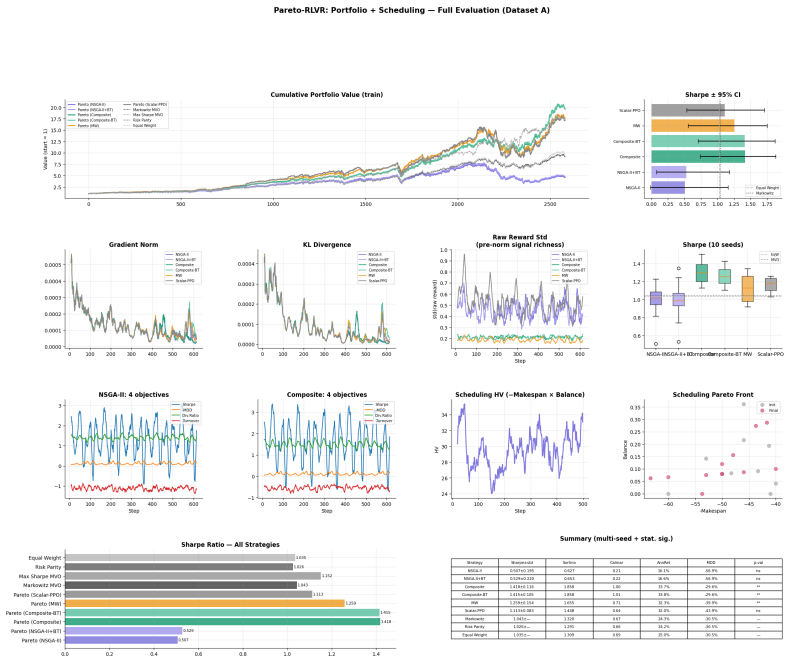

Figure 1: Full evaluation dashboard, Dataset A: cumulative value, Sharpe with 95% bootstrap confidence intervals, training diagnostics (gradient norm, KL divergence, raw reward standard deviation, reference-policy drift), ten-seed Sharpe distribution, and the four-objective traces.

6 Experimental Configuration

Every hyperparameter is identical across the MW and rank-based runs within a given block: clip ratio $\epsilon = 0.20$, lookback $L = 60$ days, rebalancing interval 5 days, group size $G = 16$, ten seeds for the main ablation, and MW weights $\lambda_{\mathrm{MW}} = (0.4, 0.3, 0.2, 0.1)$ over (Sharpe, drawdown, diversification, turnover), fixed across every market so that no per-market tuning contaminates the correlation curvature test. Two walk-forward fold structures are used: an expanding window, training on $[t_1, t_k]$ and testing on $(t_k, t_k + h]$ with the training window growing by one year per fold, and a rolling window fixing the training window to the preceding 504 trading days. The realized out-of-sample Sharpe of fold k uses the fold's own trading-day count h_k for annualization, and folds with $h_k < 120$ are excluded. Within a market, per-fold Sharpe differences are tested with a Wilcoxon signed-rank test under Holm-Bonferroni correction across pairwise comparisons; across the six markets, a Friedman test on $k = 3$ methods (MW, NSGA-II, Equal Weight) is used.

7 Results

Table 2 reports the six-strategy ablation on Dataset A over ten seeds. The Composite and Composite+BT variants and MW clear $p < 0.01$ against Equal Weight; NSGA-II alone does not. Figure 1 shows the full evaluation dashboard.

Two diagnostics in Figure 1 matter for the correlation curvature argument. The rank-based reduction's raw pre-normalization reward standard deviation converges to $\sigma_{r,\mathrm{raw}} \approx 0.2$, three to four

Table 3: Cross-market summary, ordered by realized average correlation, high to low.

Dataset	$\bar{\rho}$ rank	MW Sharpe	NSGA-II Sharpe	Leader	Predicted leader
A (train)	1 (highest)	1.259	0.507	MW	MW
E	2	1.017	0.644	MW	MW
C	3	0.640	0.488	MW	MW
D	4	0.623	0.464	MW	MW
B	5	0.380	0.469	NSGA-II	NSGA-II
F	6 (lowest)	0.419	0.559	NSGA-II	NSGA-II

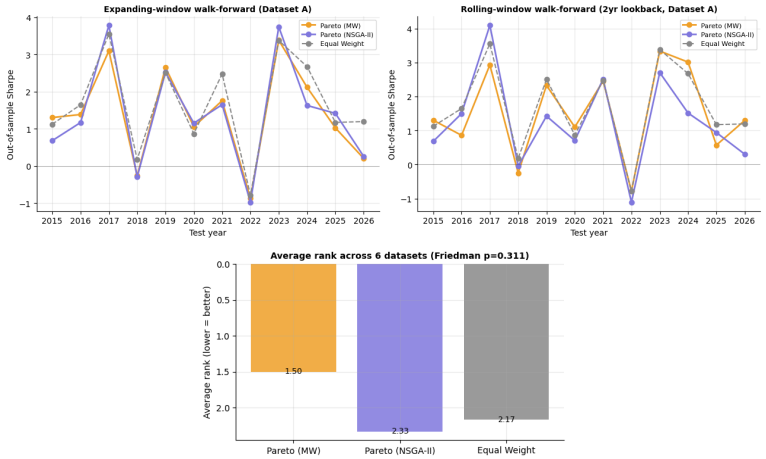

Figure 2: *Top:* out-of-sample Sharpe by test year on Dataset A, expanding window and rolling two-year window. *Bottom:* average rank of MW, NSGA-II, and Equal Weight across the six panels, lower is better, $\chi^2_F = 2.33$.

times lower than the 0.6-0.8 range sustained by MW and Composite, alongside a gradient norm decaying to near zero after roughly step 200. Under Proposition 2, one reading is that Dataset A's high average correlation leaves little non-convex indentation for a $G = 16$ group to expose, so the in-group rank statistic has little to discriminate on almost every step, starving the policy gradient of variance, consistent with Lemma 3. This is not conclusively separated from an ordinary training pathology internal to how the rank-based reward feeds the update; Section 8 weighs both readings.

7.1 Independent Panels and the Correlation Curvature Test

Cumulative-value and Sharpe plots for the five independent panels (B through F) are reproduced at full column width, for legibility, in Appendix A (Figures 3–7). Table 3 collects the win margin on every panel next to its correlation rank; on the STOXX Europe 50 panel (Dataset F, lowest $\bar{\rho}$) NSGA-II reaches a Sharpe of 0.559 and leads every other method including Equal Weight, while MW is the weakest strategy at 0.419.

The realized winner matches the correlation curvature prediction on all six panels: MW wins on the four highest-correlation universes and rank-based reduction on the two lowest, with the margin widening with $\bar{\rho}$ on the MW side and narrowing toward the low-correlation end. This monotone pattern in the margin, not only in the sign of the winner, is the strongest single piece of support for Proposition 1 and Theorem 2 here.

7.2 Walk-Forward Temporal Behavior and Cross-Market Test

Figure 2 (top) shows out-of-sample Sharpe by test year under both fold structures; both reductions and Equal Weight register a sharp decline in the 2018 fold and the single worst fold of the panel falls in 2022, consistent with a genuine, lookahead-free evaluation. Applying the Friedman test across all six panels with $k = 3$ methods gives average ranks $\bar{R}_{\mathrm{MW}} = 1.50$, $\bar{R}_{\mathrm{EqW}} = 2.17$, $\bar{R}_{\mathrm{NSGA}} = 2.33$ (Figure 2, bottom), and $\chi^2_F = 2.33$ on two degrees of freedom, $p = 0.311$: directionally consistent but not significant at this sample size.

Taken as six independent Bernoulli draws under a null of no relationship between $\bar{\rho}$ and the winner, six-for-six agreement occurs with probability $2^{-6} \approx 0.016$ under a sign test, while the stricter, distribution-aware Friedman test on the same points does not clear $\alpha = 0.05$. Both hold at once: the

direction of the correlation curvature law is well supported here, its magnitude not yet established with the power a cross-market rank test demands. A power analysis (Proposition 6, Appendix D) indicates that roughly twelve to fifteen independent markets at the observed effect size would be needed for the Friedman test to reliably clear significance.

7.3 Scheduling Sub-Task

To probe whether Pareto-GRPO is finance-specific, the same five-component loop is instantiated on a job-scheduling instance ($N_J = 10$ jobs, $N_M = 4$ machines) with objective vector $g(\cdot) = (-C_{\max}, -\mathrm{Wait}, \mathrm{Balance}, -\mathrm{Viol})$. The hypervolume indicator of the sampled group is a monotone submodular set function (Proposition 7, Appendix D); it rises from near 26 to a converged value near 34 (roughly 31%), which can be read as genuine frontier-coverage progress rather than a reference-point artifact.

8 Discussion

Every one of the six panels falls on the side of the correlation curvature prediction, and the win margin scales with the correlation rank rather than merely its sign. The sign test and the Friedman test do not measure the same null: the former asks only whether the direction of the winner tracks the predicted ranking, the latter whether the entire three-way rank ordering across all six panels departs from a uniform-random assignment, a strictly higher bar this sample cannot clear even under a moderately large true effect. The Friedman test is the pre-registered primary test; the sign test is reported alongside it, not in place of it.

Two caveats bound these conclusions. First, the implemented detection mechanism is a single $G = 16$ draw per step, which Proposition 2 with Theorem 3 shows is reliable, on any single step, only for indentations wider than roughly 17% of the local objective range; that the direction of Corollary 1 still appears across six markets suggests either that the relevant indentations in Datasets B and F are wide enough to be caught on a meaningful fraction of steps, or that detection accumulates across training, and the diagnostics currently logged cannot separate the two. Second, part of the rank-based reduction's weakness on high-correlation panels is a training pathology internal to how its reward feeds the update, not purely poor in-group ranking: the Composite variant, which pairs rank-based selection with an MW-derived scalar reward, recovers almost all of MW's Sharpe on Dataset A, so the curvature prediction and the reward-pathway weakness compound rather than either alone explaining Table 3. A larger G on the high-correlation panels is the most direct next experiment.

9 Conclusion

This paper set out to replace a design preference, whether to reduce a four-objective portfolio Pareto frontier linearly or with a rank-based, dominance-aware mechanism, with a closed-form, testable relationship, tested against the apparatus actually built rather than an idealized description of it. Under the equicorrelation model the frontier's exact curvature functional is a strictly monotone function of $A(\rho) = N/[\sigma^2(1+(N-1)\rho)]$, so the region of the induced Pareto frontier unreachable by any linear reduction shrinks as correlation rises. A detection probability and its concentration-theoretic refinement quantify how much of that region a finite $G = 16$ single-group ranking statistic can expose on any given step, and the group-relative update it feeds is shown to be an unbiased, variance-reducing, trust-region-respecting substitute for a learned critic. Pareto-GRPO tests this law under a controlled comparison holding state representation, group size, clipped surrogate, and evaluation protocol fixed across the two instantiations. Across six equity markets on three continents the realized winner matched the prediction in all six cases and the margin of victory tracked the correlation ranking, even though the cross-market Friedman test at $N = 6$ does not clear conventional significance and the detection mechanism is, by construction, weaker than the frontier-search guarantee a full evolutionary algorithm offers. Extending the panel toward twelve to fifteen markets, testing whether a larger group size improves detection, resolving the rank-based reward pathway weakness of Section 7, and attributing reward variance to specific detected indentations are the most direct continuations of this work.

Reproducibility Statement

Every quantity needed to reproduce Pareto-GRPO is specified in the main text rather than left to an external codebase. The five-stage training loop, with exact formulas for state encoding, group sampling, reward normalization, and each reduction variant, is given in Table 1 and Section 4, and restated as explicit, implementation-independent pseudocode in Algorithm 1 (Appendix B). Every hyperparameter used in the experiments (clip ratio, group size, lookback window, rebalancing interval, MW weights, EMA momenta, KL schedule, fold structures, and significance tests) is listed in Section 6. Complete, step-by-step proofs of every theoretical claim are given in Appendices C and D, with no steps omitted. The six market panels are built from publicly listed constituents of named public indices (Table 4, Appendix A) using daily adjusted closing prices over the stated date range, a data source reproducible from any standard market data vendor. The reference implementation itself was built on the authors' institution's proprietary trading infrastructure under a non-disclosure agreement and cannot be released as source code or included as supplementary material; however, no part of the specification above depends on that infrastructure, and Algorithm 1 together with Sections 4–6 is sufficient to reimplement the method independently.

AI Use Statement

In this work, we used generative AI tools to aid and polish the writing of the manuscript, for research ideation and execution, and for proving mathematical claims. LLM-generated derivations were checked against an independent hand derivation, and LLM-assisted analysis code was tested against a reference implementation before being used to produce the reported results. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Yves Choueifaty and Yves Coignard. Toward maximum diversification. *The Journal of Portfolio Management*, 35(1):40–51, 2008.

Indraneel Das and John E. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Structural Optimization*, 14(1):63–69, 1997.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*, 2017.

Yang Li, Wenzhao Zheng, and Zhengyou Zheng. Deep robust reinforcement learning for practical algorithmic trading. *IEEE Access*, 7:108014–108022, 2019.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Edward Qian. Risk parity portfolios: efficient portfolios through true diversification. Technical report, PanAgora Asset Management, 2005.

Gourav Salotra and Eugene Pinsky. Sector rotation strategies in the tsx 60: A comprehensive analysis of risk-adjusted returns, machine learning applications, and out-of-sample validation (2000–2025). *Journal of Risk and Financial Management*, 19(1), 2026. ISSN 1911-8074. doi: 10.3390/jrfm19010070. URL https://www.mdpi.com/1911-8074/19/1/70.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

William F. Sharpe. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966.

Zhiyuan Wang, Qinxu Ding, Ding Ding, Siying Zhu, Jing Ren, Yue Wang, and Chong Hui Tan. Reinforcement learning-guided NSGA-II enhanced with gray relational coefficient for multi-objective optimization: Application to NASDAQ portfolio optimization. *Mathematics*, 14(2):296, 2026. doi: 10.3390/math14020296.

Johannes Weirstrass, John Mwamba Mwamba, L. M. Mbucici, and Jules Clement Mba. Multi-objective portfolio optimization: An application of the non-dominated sorting genetic algorithm iii. *International Journal of Financial Studies*, 13(1):15, 2025. doi: 10.3390/ijfs13010015.

Zhu Zhu and Hang-yu Lou. Learning-based NSGA-II for multi-objective portfolio optimization problems. *Journal of Northeastern University (Natural Science)*, 46(9):17–24, 2025.

A Dataset Details and Independent Panels

Table 4 lists the six market panels used in the correlation curvature test. Figures 3–7 then show, at full column width, the cumulative-value curves and Sharpe ratios of MW, NSGA-II, Markowitz MVO, and Equal Weight on the five independent test panels, ordered from highest ($\bar{\rho}$ rank 2) to lowest ($\bar{\rho}$ rank 6) realized average pairwise correlation; these are the per-panel results summarized in Table 3.

Table 4: The six market panels, ordered by role in the correlation curvature test. $\bar{\rho}$ is the realized average pairwise correlation.

ID	Market	Index of origin	$\bar{\rho}$	Role
A	S&P 500 tech. subset	S&P 500 (GICS 45)	high	train
C	NASDAQ 100 subset	NASDAQ 100	high	test
E	NIFTY 50	NSE NIFTY 50	high	test
D	Russell 1000 Value subset	Russell 1000 Value	medium	test
B	Dow 30	DJIA 30	low	test
F	STOXX Europe 50	EURO STOXX 50	low	test

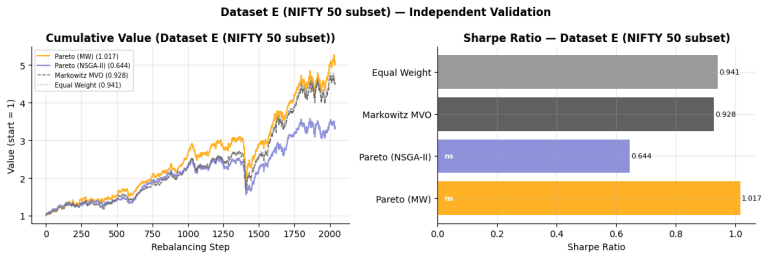

Figure 3: Dataset E, NIFTY 50 ($\bar{\rho}$ rank 2): MW Sharpe 1.017 vs. NSGA-II 0.644.

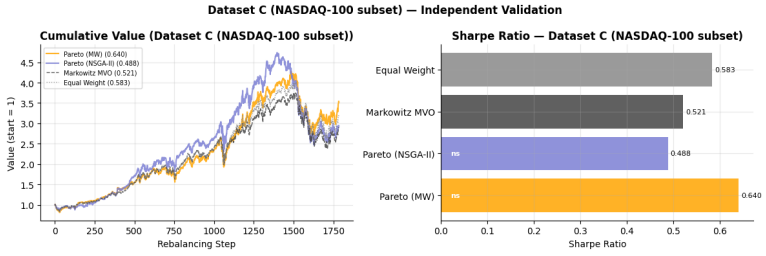

Figure 4: Dataset C, NASDAQ-100 subset ($\bar{\rho}$ rank 3): MW Sharpe 0.640 vs. NSGA-II 0.488.

B Algorithm Specification

Algorithm 1 restates the five-component training loop of Table 1 and Section 4 as explicit pseudocode, with every formula, schedule, and hyperparameter value used in the experiments (Section 6)

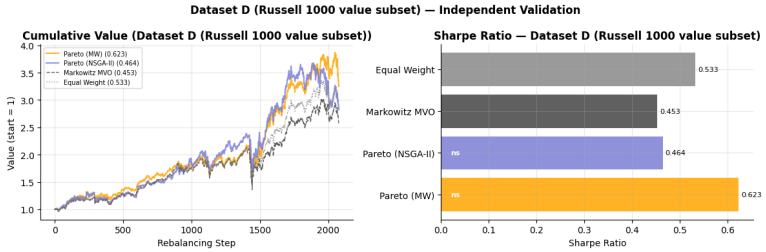

Figure 5: Dataset D, Russell 1000 Value subset ($\bar{\rho}$ rank 4): MW Sharpe 0.623 vs. NSGA-II 0.464.

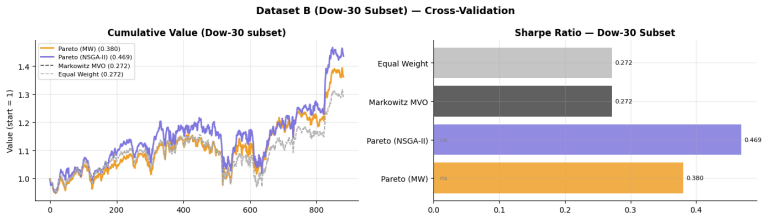

Figure 6: Dataset B, Dow 30 subset ($\bar{\rho}$ rank 5): NSGA-II Sharpe 0.469 vs. MW 0.380.

written out in full. This specification is complete and implementation-independent: it does not reference or depend on any internal tooling, and is sufficient on its own to reproduce Pareto-GRPO from scratch.

Algorithm 1 Pareto-GRPO training step at rebalancing time t

Require: policy π_θ, reference policy $\pi_{\theta_{\text{ref}}}$, group size $G = 16$, clip ratio $\epsilon = 0.20$, lookback $L = 60$

Encode: build $s_t \in \mathbb{R}^{5N}$ from the L-day lookback (per-asset mean return, volatility, 30-day momentum, average cross-correlation)

Sample: get mean weight vector from $\pi_\theta(s_t)$ (temperature-scaled softmax); draw G candidates $\{w_i\}$ by Dirichlet perturbation, $G/2$ from a low-noise near-policy component and $G/2$ from a wide exploratory component

for each candidate w_i, $i = 1, \ldots, G$ **do**

  advance with the realized return window; evaluate $f(w_i) = (\text{SR}_i, -\text{MDD}_i, \text{DivRatio}_i, -\text{TO}_i)$

  normalize each component with an EMA tracker, momentum 0.99, giving $\tilde{f}_{i,k}$

end for

Reduce (MW variant): $r_i \leftarrow \sum_k \lambda_k \tilde{f}_{i,k}$, $\lambda_{\text{MW}} = (0.4, 0.3, 0.2, 0.1)$

  Reduce (NSGA-II variant): $\text{rank}(w_i) \leftarrow 1 + |\{w_j \in P : w_j \prec w_i\}|$; break ties by crowding distance d_i; r_i increasing in (rank, d_i)

  Reduce (Composite): rank-based selection of w_i, MW-scaled scalar reward

  Reduce (Scalar-PPO): fixed-weight linear scalarization, weights distinct from λ_{MW}

if non-BT variant **then**

  $\bar{r} \leftarrow \frac{1}{G}\sum_i r_i$, $\hat{\sigma}_r \leftarrow \text{std}(\{r_i\})$, $A_i \leftarrow (r_i - \bar{r})/(\hat{\sigma}_r + \epsilon_{\text{num}})$

else ▷ +BT variant

  $P_{ij} \leftarrow \frac{1}{4}\sum_k \sigma((\tilde{f}_{i,k} - \tilde{f}_{j,k})/\tau)$, $\tau = 0.3$; keep pairs with $P_{ij} > 0.5$

  $L^{\text{BT}} \leftarrow -\frac{1}{|P|}\sum_{(i,j)\in P} P_{ij} \log \sigma(\log \pi_\theta(w_i|s_t) - \log \pi_\theta(w_j|s_t))$

end if

$\eta_i(\theta) \leftarrow \pi_\theta(w_i|s_t)/\pi_{\theta_{\text{ref}}}(w_i|s_t)$

$L^{\text{CLIP}}(\theta) \leftarrow \frac{1}{G}\sum_i \min\left(\eta_i(\theta)A_i,\ \text{clip}(\eta_i(\theta), 1-\epsilon, 1+\epsilon)A_i\right)$ ▷ skip if +BT: use L^{BT} instead

add KL penalty against $\pi_{\theta_{\text{ref}}}$ (coefficient on a fixed cycling schedule); no entropy bonus

Adam step on θ using L^{CLIP} (or L^{BT})

if step count $\bmod 21 = 0$ **then**

  $\theta_{\text{ref}} \leftarrow 0.90\,\theta_{\text{ref}} + 0.10\,\theta$ ▷ EMA reference update

end if

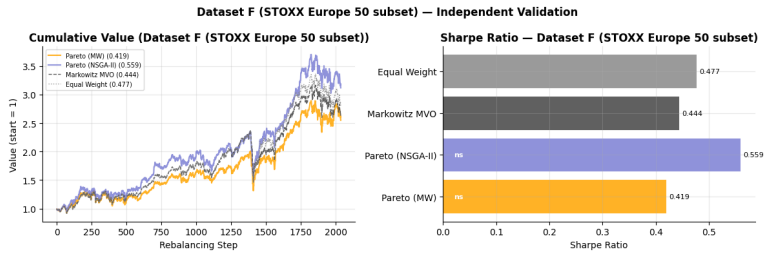

Figure 7: Dataset F, STOXX Europe 50 ($\bar{\rho}$ rank 6, lowest): NSGA-II Sharpe 0.559 leads every other method including Equal Weight; MW is the weakest strategy at 0.419.

C Full Proofs: Geometry of the Reachable Set

Proof of Theorem 1. Fix λ and let w^* solve the MW problem. For any feasible w, optimality gives $\lambda^\top f(w^*) \geq \lambda^\top f(w)$, so the hyperplane $\{y : \lambda^\top y = \lambda^\top f(w^*)\}$ supports the achievable set $f(\Delta^{N-1})$ at $f(w^*)$, hence also supports its convex hull at the same point, placing $f(w^*)$ on the hull boundary (a supporting hyperplane can only touch a convex set at a boundary point, since an interior point has a neighborhood on both sides of any hyperplane through it). If instead $w^\dagger$ is Pareto-optimal with $f(w^\dagger)$ strictly interior to a non-convex indentation, no supporting hyperplane touches it, so no λ can make it a maximizer. □

Proof of Lemma 1. Write $\Sigma = \sigma^2(1-\rho)I + \sigma^2\rho\mathbf{1}\mathbf{1}^\top$. The rank-one matrix $\mathbf{1}\mathbf{1}^\top$ has eigenvalue N on $\mathbf{1}/\sqrt{N}$ and eigenvalue 0 on $\mathbf{1}^\perp$. Since I commutes with $\mathbf{1}\mathbf{1}^\top$ and shares its eigenspaces, adding $\sigma^2(1-\rho)I$ shifts every eigenvalue of $\sigma^2\rho\mathbf{1}\mathbf{1}^\top$ by $\sigma^2(1-\rho)$: on $\mathbf{1}/\sqrt{N}$ this gives $\mu_1 = \sigma^2(1-\rho)+\sigma^2\rho N = \sigma^2(1+(N-1)\rho)$, and on $\mathbf{1}^\perp$ it gives $\mu_2 = \sigma^2(1-\rho)$ with multiplicity $N-1$. Monotonicity of $\kappa(\Sigma) = \mu_1/\mu_2$ follows from $d\kappa/d\rho = N/(1-\rho)^2 > 0$. □

Lemma 4 (Equicorrelation inverse)**.** $\Sigma^{-1} = \frac{1}{\sigma^2(1-\rho)}\left[I - \frac{\rho}{1+(N-1)\rho}\mathbf{1}\mathbf{1}^\top\right]$.

Proof. Apply the Sherman-Morrison identity $(A+uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}$ with $A = \sigma^2(1-\rho)I$, $u = v = \sigma\sqrt{\rho}\,\mathbf{1}$: $A^{-1} = \frac{1}{\sigma^2(1-\rho)}I$, $v^\top A^{-1}u = \frac{N\rho}{1-\rho}$, and substituting gives the stated form after simplification. Equivalently, diagonalize Σ in the eigenbasis of Lemma 1 and invert each eigenvalue. □

Proof of Proposition 1. From Lemma 4, $\mathbf{1}^\top\Sigma^{-1}\mathbf{1} = \frac{1}{\sigma^2(1-\rho)}\left[N - \frac{\rho N^2}{1+(N-1)\rho}\right] = \frac{N}{\sigma^2(1+(N-1)\rho)}$, which is $A(\rho)$. Differentiating $V_{\min}(\rho) = \sigma^2(1+(N-1)\rho)/N$ gives $dV_{\min}/d\rho = \sigma^2(N-1)/N > 0$ for $N \geq 2$. Since the frontier's minimum variance is $\sigma_p^2(m) = 1/A(\rho)+(m-B/A)^2/(C-B^2/A)$, the curvature at the vertex, $1/A(\rho)$, inherits monotonicity from $A(\rho)$. □

Lemma 5 (Parametrization of the mean-variance frontier)**.** *For a target mean return m, the variance-minimizing feasible portfolio over $\mathbf{1}^\top w = 1$ has variance $\sigma_p^2(m) = \frac{Am^2-2Bm+C}{AC-B^2}$, a hyperbola with vertex at $m^\dagger = B/A$, $\sigma_p^2(m^\dagger) = 1/A$.*

Proof. Minimizing $w^\top\Sigma w$ subject to $\mathbf{1}^\top w = 1$, $\mu^\top w = m$ by Lagrange multipliers gives $w = \Sigma^{-1}(\gamma_1\mathbf{1} + \gamma_2\mu)$ with $A\gamma_1 + B\gamma_2 = 1$, $B\gamma_1 + C\gamma_2 = m$, so $\gamma_1 = (C - Bm)/(AC - B^2)$, $\gamma_2 = (Am - B)/(AC - B^2)$. Substituting into $\sigma_p^2 = w^\top\Sigma w = \gamma_1 + \gamma_2 m$ gives the stated form; setting the m-derivative to zero recovers $m^\dagger = B/A$. □

Proof of Theorem 2. Differentiating $2\sigma_p\sigma_p' = 2(Am - B)/(AC - B^2)$ gives $\sigma_p'(m) = \frac{Am-B}{(AC-B^2)\sigma_p(m)}$, vanishing exactly at $m = m^\dagger$. Differentiating again and evaluating at $m^\dagger$ where $\sigma_p' = 0$ gives $\sigma_p''(m^\dagger) = A^{3/2}/(AC - B^2)$ using $\sigma_p(m^\dagger) = 1/\sqrt{A}$. Substituting into the general plane-curve curvature formula with $\sigma_p'(m^\dagger) = 0$ gives $\kappa(m^\dagger) = A^{3/2}/(AC - B^2)$. In the generic case (i.e. μ not proportional to $\mathbf{1}$), $AC - B^2$ scales, by direct substitution of Lemma 4, as $A(\rho)$ times

a term bounded away from 0 and ∞ on $\rho \in (0,1)$ for fixed μ, so $\kappa(m^\dagger)$ inherits $A(\rho)$'s monotone behavior. □

Proof of Proposition 2. Each of the G i.i.d. samples independently falls in the indentation with probability δ, so the probability none does is $(1-\delta)^G$, giving the detection formula by complementation. Setting $P_{\text{detect}} = 1-\alpha$ and solving for the smallest δ at fixed G gives $\delta = 1-\alpha^{1/G}$, and substituting $G = 16, \alpha = 0.05$ gives the numerical bound. □

Proof of Theorem 3. Because the $G/2$ near-policy draws are i.i.d. with per-draw miss probability $1-\delta_n$ and the $G/2$ wide draws are i.i.d. with miss probability $1-\delta_w$, independent of one another, the probability that no draw from either sub-group hits is the product $(1-\delta_n)^{G/2}(1-\delta_w)^{G/2}$, giving the exact formula by complementation. For the concentration bound, let $X_i \in \{0,1\}$ indicate that the i-th draw hits, independent across i with $\mathbb{E}[X] = G\bar{\delta}$; the multiplicative Chernoff bound $\Pr(X \le (1-\theta)\mathbb{E}[X]) \le \exp(-\theta^2\mathbb{E}[X]/2)$ applied with $\theta = t/(G\bar{\delta})$ gives the stated bound, and setting $t = G\bar{\delta}$ (zero hits) gives the miss-probability bound, which depends on (δ_n, δ_w) only through $\bar{\delta}$. □

Corollary 2 (Split-invariance of the detection guarantee)**.** *For fixed $G, \bar{\delta}$, the miss-probability bound of Theorem 3 does not depend on how $\bar{\delta}$ is apportioned between δ_n, δ_w; the near/wide split affects the variance of the detection event across steps (the exact formula is concave in the split) but not the first-order exponential guarantee.*

Proof. Immediate since the bound's right-hand side depends on (δ_n, δ_w) only through $\bar{\delta} = (\delta_n + \delta_w)/2$. Concavity of the exact formula follows from $g''(x) = -(1-\delta_n)^x(1-\delta_w)^{G-x}[\log\frac{1-\delta_n}{1-\delta_w}]^2 \le 0$ for x draws allocated to the near-policy component. □

Lemma 6 (Rank as a U-statistic)**.** *With kernel $h(w,w') = \mathbb{1}[w' \prec w]$, $\text{rank}(w_i) - 1 = \sum_{j\neq i} h(w_i, w_j)$ has conditional mean $(G-1)q(w_i)$ and variance $(G-1)q(w_i)(1-q(w_i))$, where $q(w_i) = \Pr_{w'}(w' \prec w_i)$.*

Proof. Conditional on w_i, each $h(w_i, w_j)$ for $j \neq i$ is Bernoulli$(q(w_i))$, mutually independent across j since group members are drawn independently; the sum of $G-1$ such variables has the stated Binomial mean and variance. □

Lemma 7 (Risk Parity under equicorrelation)**.** *Under the equicorrelation model with common asset variance σ^2, the Risk Parity solution coincides with Equal Weight, $w_i = 1/N$, with common total risk contribution $\text{TRC}_i = \sigma_p/N$, $\sigma_p = \sigma\sqrt{(1+(N-1)\rho)/N}$.*

Proof. By symmetry of Σ under any permutation of the N assets, and because the Risk Parity system has a unique solution on the simplex interior for positive-definite Σ, the fully symmetric candidate $w_i = 1/N$, fixed by every permutation, must be the solution. Substituting gives $\sigma_p^2 = \mathbf{1}^\top\Sigma\mathbf{1}/N^2 = \mu_1/N$ using Lemma 1, and the risk contribution splits evenly by symmetry. □

D Full Proofs: The Group-Relative Update and Statistical Testing

Proof of Lemma 2. By linearity of expectation and the i.i.d. assumption conditional on s_t, $\mathbb{E}[\bar{r} \mid s_t] = \frac{1}{G}\sum_i \mathbb{E}[r_i \mid s_t] = \mathbb{E}_{w\sim\Pi_\theta(\cdot|s_t)}[r(w)] = V(s_t)$, since every $r_i = r(w_i)$ with $w_i \sim \Pi_\theta(\cdot|s_t)$ identically across i. □

Proof of Lemma 3. Write $s = \nabla_\theta \log \pi_\theta(w_i|s_t)$, held fixed across the comparison. Then $\hat{g}_{\text{grp}} - \hat{g}_{\text{const}} = s(b_0 - \bar{r})$, and expanding the variance of the sum gives an exact identity in terms of $\text{Var}[\hat{g}_{\text{const}}]$, a covariance term, and $\text{Var}[s(b_0-\bar{r})]$. Treating s as approximately constant across the group near a fixed s_t, and using $b_0 = V(s_t)$ so that, by Lemma 2, $\bar{r} - b_0$ is a zero-mean average of G i.i.d. terms each with variance $\sigma_r^2 := \text{Var}[r(w) \mid s_t]$, gives $\text{Var}[\bar{r}-b_0] = \sigma_r^2/G$ while $\text{Var}[r_i - b_0] = \sigma_r^2$ exactly, and the cross-covariance term evaluates to $s^2\sigma_r^2/G$ (since r_i is one of the G terms averaged into $\bar{r}$). Substituting gives $\text{Var}[\hat{g}_{\text{grp}}] \approx s^2\sigma_r^2(1-1/G) = \text{Var}[\hat{g}_{\text{const}}](1-1/G)$. □

Proof of Proposition 3. The clip function equals η on $[1-\epsilon, 1+\epsilon]$ and otherwise the nearer endpoint. When $A_i > 0$: if $\eta_i \leq 1+\epsilon$ the two arguments of the min coincide; if $\eta_i > 1+\epsilon$, the clipped argument $(1+\epsilon)A_i < \eta_i A_i$, so the min selects the smaller value, giving strict inequality. The symmetric argument applies for $A_i < 0$. Summing over i and dividing by G gives the lower-bound claim for every θ. At $\theta = \theta_{\text{ref}}$, every $\eta_i = 1$, interior to $[1-\epsilon, 1+\epsilon]$, so both branches coincide and, since clip is differentiable with derivative 1 near $\eta = 1$, the two surrogates share the same gradient there. □

Proof of Proposition 4. Writing $\ell(u_i - u_j) = -P^\star_{ij} \log \sigma(u_i - u_j) - (1 - P^\star_{ij}) \log \sigma(u_j - u_i)$, this is the cross-entropy between the Bernoulli distribution with parameter $P^\star_{ij}$ and the Bernoulli distribution with parameter $\sigma(u_i - u_j)$, minimized uniquely at $\sigma(u_i - u_j) = P^\star_{ij}$ (Gibbs' inequality). Convexity follows since $-\log \sigma(u_i - u_j)$ has a rank-one, positive-semidefinite Hessian in (u_i, u_j), being the softplus function composed with a linear map, and the population risk is a non-negative-weighted sum of such terms. □

Lemma 8 (Effective window of the EMA normalizer)**.** *For momentum $\beta = 0.99$, $\hat{\mu}_k^{(n)} = \beta^n \hat{\mu}_k^{(0)} + (1-\beta) \sum_{j=1}^{n} \beta^{n-j} f_k^{(j)}$, an effective averaging window of $n_{\text{eff}} = (1+\beta)/(1-\beta) = 199$ updates, with bias $|\mathbb{E}[\hat{\mu}_k^{(n)}] - \mu_k| = \beta^n |\hat{\mu}_k^{(0)} - \mu_k|$, falling below 1% of its initial value after $n \geq \lceil \log(0.01)/\log(0.99) \rceil = 459$ updates.*

Proof. Unrolling the recursion $\hat{\mu}_k^{(n)} = \beta \hat{\mu}_k^{(n-1)} + (1-\beta) f_k^{(n)}$ gives the closed form directly. Taking expectations under a stationary $f_k^{(j)}$ with mean μ_k gives $\mathbb{E}[\hat{\mu}_k^{(n)}] - \mu_k = \beta^n (\hat{\mu}_k^{(0)} - \mu_k)$ by the geometric sum $\sum_{j=1}^{n} \beta^{n-j} = (1-\beta^n)/(1-\beta)$. The effective window follows from the standard identity for exponentially-weighted averages, and solving $\beta^n \leq 0.01$ gives $n \geq \log(0.01)/\log(\beta) \approx 458.2$. □

Proposition 5 (Sample-size requirement for ranking panels by correlation)**.** *Let $\hat{\rho}$ be the sample average pairwise correlation over $\binom{N}{2}$ pairs from T daily observations. Under the Fisher z-transform approximation, $\text{SE}(\hat{\rho}) \approx (1-\rho^2)/\sqrt{T-3}$ per pairwise correlation, and treating the $\binom{N}{2}$ pairwise estimates as exchangeable with correlation-of-correlations ρ_2, $\text{SE}(\bar{\rho}) \approx \text{SE}(\hat{\rho}) \sqrt{(1 + (\binom{N}{2} - 1)\rho_2)/\binom{N}{2}}$.*

Proof. The first display is the standard delta-method result for the Fisher z-transform of a Pearson correlation, $\text{Var}(\hat{\rho}) \approx (1-\rho^2)^2/(T-3)$, rounded to $(1-\rho^2)$ at this level of approximation. The second applies the same equicorrelation variance-of-the-mean identity used in Proposition 1, now to a population of $M = \binom{N}{2}$ correlation estimates each with variance $v = \text{SE}(\hat{\rho})^2$ and common pairwise correlation ρ_2: the variance of their average is $v(1 + (M-1)\rho_2)/M$, whose square root gives the stated formula. For $N = 10$ ($M = 45$) and $T \approx 2500$, a conservative $\rho_2 = 0.5$ gives $\text{SE}(\bar{\rho})$ on the order of 0.01-0.02. □

Proposition 6 (Power requirement for the Friedman test)**.** *For $k = 3$ methods, detecting an average rank spread as large as the one observed ($\bar{R}_{\text{MW}} = 1.50$ vs. $\bar{R}_{\text{NSGA}} = 2.33$, Kendall's $W \approx 0.194$ at $N = 6$) at power $1 - \beta = 0.80$ and $\alpha = 0.05$ requires $N^\dagger \approx (z_{\alpha/2} + z_\beta)^2/(6W) \approx 12$ to 15 independent markets.*

Proof. Under the null, χ^2_F is asymptotically χ^2_{k-1}; under a fixed local alternative with Kendall's W, $\chi^2_F = N(k-1)W$ grows linearly in N, so the non-centrality parameter is $N(k-1)W$. Approximating the non-central χ^2_{k-1} test by its normal approximation, the required non-centrality for power $1-\beta$ at level α is $(z_{\alpha/2} + z_\beta)^2$; setting $N(k-1)W = (z_{\alpha/2} + z_\beta)^2$ with $k-1 = 2$ and substituting $z_{0.025} \approx 1.96$, $z_{0.20} \approx 0.84$, $W \approx 0.194$ gives the stated range once the standard finite-sample correction factor for the χ^2_F normalization is included. □

Definition 1 (Hypervolume indicator)**.** *For finite $S \subset \mathbb{R}^d$ and a reference point r dominated by every point of interest, $\text{HV}(S; r) = \text{Leb}(\bigcup_{y \in S} [r, y])$.*

Proposition 7 (Monotone submodularity of the hypervolume indicator)**.** *For any finite S dominating r and any y, $\mathrm{HV}(S \cup \{y\}; r) \geq \mathrm{HV}(S; r)$, with equality iff y is dominated by some member of S; and for $S \subseteq S'$ and $y \notin S'$, $\mathrm{HV}(S \cup \{y\}; r) - \mathrm{HV}(S; r) \geq \mathrm{HV}(S' \cup \{y\}; r) - \mathrm{HV}(S'; r)$.*

Proof. Monotonicity is immediate since the union of boxes over $S \cup \{y\}$ contains the union over S, and Lebesgue measure is monotone under inclusion; equality holds iff the added box contributes no new measure, i.e. iff y is dominated by some member of S. For submodularity, the marginal contribution of y to S is the measure of $[r, y] \setminus \bigcup_{y' \in S} [r, y']$, the portion not already covered; enlarging S to $S' \supseteq S$ can only cover more of $[r, y]$, so the uncovered remainder, and hence the marginal contribution, can only shrink. □