
FR-LUX: FRICTION-AWARE, REGIME-CONDITIONED POLICY OPTIMIZATION FOR IMPLEMENTABLE PORTFOLIO MANAGEMENT

Zhang Jian'an

Guanghua School of Management, Peking University
Peking University
Beijing, China
2501111059@stu.pku.edu.cn

ABSTRACT

Transaction costs and regime shifts are the main reasons why paper portfolios fail in live trading. We develop **FR-LUX** (Friction-aware, Regime-conditioned Learning under eXecution costs), a reinforcement-learning framework that learns *after-cost* trading policies and remains robust across volatility-liquidity regimes. FR-LUX integrates three ingredients: (i) a microstructure-consistent execution model combining proportional and impact costs, directly embedded in the reward; (ii) a *trade-space trust region* that constrains changes in inventory flow rather than only logits, yielding stable, low-turnover updates; and (iii) explicit regime conditioning so the policy specializes to LL/LH/HL/HH states without fragmenting the data. On a 4×5 grid of regimes and cost levels (0–50 bps) with three seeds per cell, FR-LUX achieves the top average Sharpe across all 20 scenarios with narrow bootstrap confidence intervals, maintains a flatter cost-performance slope than strong baselines (vanilla PPO, mean-variance with/without caps, risk-parity), and attains superior risk-return efficiency for a given turnover budget. Pairwise scenario-level improvements are strictly positive and remain statistically significant after Romano-Wolf stepdown and HAC-aware Sharpe comparisons. We provide formal guarantees: existence of an optimal stationary policy under convex frictions; a monotonic improvement lower bound under a KL trust region with explicit remainder terms; an upper bound on long-run turnover and an induced inaction band due to proportional costs; a strictly positive value advantage for regime-conditioned policies when cross-regime actions are separated; and robustness of realized value to cost misspecification. The methodology is implementable—costs are calibrated from standard liquidity proxies, scenario-level inference avoids pseudo-replication, and all figures and tables are reproducible from our artifacts.

Keywords transaction costs; market microstructure; regime switching; reinforcement learning; portfolio optimization; CVaR / maximum drawdown; turnover; Sharpe ratio; multiple testing; implementability.

1 Introduction

The gap between methods that *forecast* returns and policies that *trade* under realistic frictions remains a central obstacle to deploying modern machine learning (ML) in institutional portfolios. In frictionless settings, mean-variance logic [1] and its many extensions provide clear optimality benchmarks; once trading costs, market impact, and turnover constraints are accounted for, those benchmarks break down and performance can deteriorate sharply [10, 9, 11, 20, 21]. At the same time, ML has transformed empirical asset pricing and portfolio construction by extracting non-linear structure from high-dimensional characteristics [16, 17, 18, 19]. The key open question is therefore not whether ML can predict returns, but whether it can deliver *after-cost* portfolios that are robust across regimes, scalable in capacity, and statistically significant after proper multiple-testing controls [14, 15, 24, 25, 26].

We address this question with a new decision-making framework that couples policy optimization with explicit cost regularization. We introduce **FR-LUX** (*Flow-Regularized Learning Under eXecution costs*), a cost-aware policy optimization method that learns trading rules directly in the presence of proportional and impact costs and that penalizes *inventory flow* as a structural control of turnover. FR-LUX builds on monotone policy-improvement principles from

reinforcement learning (RL)—trust-region and conservative policy iteration [27, 28, 30, 29]—and adapts them to the portfolio domain by (i) embedding a transaction-cost-calibrated penalty in the objective, (ii) enforcing a trust-region in *trade space* rather than in raw parameter space, and (iii) learning a regime-aware baseline that reuses information across market states [22, 23, 12]. Conceptually, FR-LUX converts the classic rebalancing rule aim in front of the target and trade partially” [10] into a learned *regularized policy* that internalizes future cost and slippage.

Empirical preview. Using a scenarios $\times$ costs grid (20 macro-liquidity regimes crossed with 0–50 bp transaction cost levels) and three random seeds per cell, we benchmark FR-LUX against representative baselines: an unconstrained mean-variance policy, a turnover-capped mean-variance policy, a risk-parity style heuristic, and a strong PPO implementation. Across scenarios, FR-LUX delivers the highest average Sharpe and retains its edge as costs rise (Fig. 1–2). Regime profiles (Fig. 3) show that performance persists in both low- and high-volatility/liquidity conditions, consistent with the view that the method learns to modulate risk when volatility spikes [12]. Risk-return clouds using maximum drawdown (MDD) document a favorable frontier shift (Fig. 4). Turnover-Sharpe plots (Fig. 5) reveal that FR-LUX sits on a lower-turnover iso-Sharpe curve than alternatives, in line with theory that cost-aware regularization shrinks unnecessary inventory flow [9]. Pairwise sign tests and distributional comparisons (Fig. 6–7) indicate statistically reliable outperformance after Romano-Wolf step-down corrections [25] and model-comparison metrics based on Sharpe improvements [15].

Why cost awareness matters now. Transaction-cost measurement has advanced to a point where ignoring costs is no longer defensible. Low-frequency proxies and modern spread/impact estimators enable cost calibration at scale [6, 7, 8]. Recent top-journal evidence documents first-order cost effects on capacity and strategy survival in currencies and fixed income [20, 21]. In this environment, methods that merely forecast but do not control execution paths are fragile. RL has emerged as a natural language for sequential trading and execution [32], yet rigorous, finance-native regularization for costs and turnover remains underdeveloped.

Our contributions. This paper makes four contributions.

1. *A cost-regularized policy optimizer.* We formalize FR-LUX, a policy-gradient-based algorithm with a trust region in trade flow and an *execution-aware* penalty, providing a practical recipe for learning after-cost policies. The design connects RL improvement bounds [27, 28, 29] to dynamic trading with costs [10, 9].
2. *Theory.* We prove a *conservative improvement bound* that lower-bounds the after-cost performance of the updated policy as a function of (a) the estimated advantage, (b) the trust-region radius, and (c) the turnover penalty coefficient, and we show that the bound tightens when the realized turnover proxy tracks structural liquidity [8, 6]. We further provide a robustness proposition under cost misspecification: if the true cost is within a relative factor of the calibrated proxy, FR-LUX preserves first-order optimality in the induced risk-return frontier (linking to [10]).
3. *Evaluation protocol.* We adopt regime-stratified aggregation, cost-sensitivity curves, and multiple-testing-robust inference using Romano-Wolf step-down p-values [25] and Sharpe-ratio model comparison [15], complementing classical reality checks [24, 26].
4. *Evidence.* On the $20 \times$ costs testbed, FR-LUX attains the top average Sharpe with narrow bootstrap CIs, retains performance as costs increase, and dominates baselines in pairwise sign tests. The method traces lower turnover for a given Sharpe and maintains strong performance in both liquidity-rich and liquidity-poor regimes, consistent with volatility-managed intuition [12].

Relation to literature. Our work intersects four strands. (i) *ML/asset pricing*: deep and non-linear estimators deliver sizable improvements in expected returns and risk attribution [16, 17, 18, 19]. (ii) *Trading with frictions*: dynamic policies internalizing future cost/impact are essential to realistic portfolio control [10, 9, 11]. (iii) *Liquidity measurement*: scalable cost proxies enable disciplined calibration and capacity analysis [6, 7, 8]. (iv) *RL for finance*: recent surveys [32] and empirical studies underscore both the promise and the pitfalls of RL in markets, motivating finance-aware regularization and inference. By integrating these pieces, FR-LUX advances from “predict then optimize” to *optimize while respecting execution*, delivering statistically credible gains across regimes and costs.

The remainder of the paper develops the FR-LUX objective and theoretical guarantees (Section 2), details the experimental design and cost calibration (Section 4), reports main results and inference (Section 5).

2 Problem Setup and Method: FR-LUX

2.1 Frictional Portfolio Environment as an MDP

We model portfolio control as a discounted Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ augmented with an observed market regime $z_t \in \mathcal{Z} := \{\text{LL}, \text{LH}, \text{HL}, \text{HH}\}$ capturing (low/high) volatility and (high/low) liquidity.¹ At each time t , the agent observes $s_t = (x_t, w_{t-1}, z_t)$ where x_t denotes predictors (returns, volatilities, liquidity proxies, macro controls), $w_{t-1} \in \mathbb{R}^d$ are pre-trade portfolio weights on d risky assets (the residual goes to the funding account), and z_t is the discrete regime label. The action a_t specifies the *post-trade* target weights $\tilde{w}_t \in \mathcal{W}$ and induces a *trade flow* $\Delta w_t := \tilde{w}_t - w_{t-1}$.

One step net reward (to be maximized) is

$$r_t^{\text{net}} = \underbrace{\tilde{w}_t^\top r_{t+1}}_{\text{gross portfolio return}} - \underbrace{C_{z_t}(\Delta w_t)}_{\text{execution costs}} - \lambda_{\text{risk}} \Psi(L_{t+1}), \quad (1)$$

where r_{t+1} are next-period asset returns, C_{z_t} is a convex regime-dependent execution-cost functional, Ψ is a downside-risk proxy (MDD or CVaR), and L_{t+1} is the portfolio loss.² The control objective is the discounted value

$$J(\theta) = \mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_t^{\text{net}} \middle| \pi_\theta \right], \quad (2)$$

where $\pi_\theta(a | s, z)$ is a parametric, *regime-conditioned* stochastic policy.

Action and feasibility. We consider two common feasible sets: (i) long-only simplex $\mathcal{W} = \{w \in \mathbb{R}^d : w \geq 0, \mathbf{1}^\top w = 1\}$, projecting the network output via a differentiable softmax or exact Euclidean projection [44]; (ii) long-short box with leverage and position caps $\mathcal{W} = \{w : \|w\|_1 \leq \Lambda, -c \leq w_i \leq c\}$, using ℓ_1 -ball and box projections [43]. These projections stabilize learning and prevent inadmissible trades.

2.2 Regime Construction and Balancing

We map raw diagnostics into regimes using thresholds on (i) realized volatility σ_t and (ii) illiquidity ℓ_t (e.g., Amihud *ILLIQ*, effective spread, or Pastor–Stambaugh innovations; [4, 5, 6, 8]). Let $\tau_\sigma^L < \tau_\sigma^H$ and $\tau_\ell^L < \tau_\ell^H$ be quantile cutoffs calibrated on a rolling window. Define

$$z_t = \begin{cases} \text{LL}, & \sigma_t \leq \tau_\sigma^L, \ell_t \leq \tau_\ell^L, \\ \text{LH}, & \sigma_t \leq \tau_\sigma^L, \ell_t > \tau_\ell^L, \\ \text{HL}, & \sigma_t > \tau_\sigma^H, \ell_t \leq \tau_\ell^L, \\ \text{HH}, & \sigma_t > \tau_\sigma^H, \ell_t > \tau_\ell^L, \\ \text{else}, & \text{nearest neighbor by } (\sigma_t, \ell_t). \end{cases}$$

To avoid over-optimizing to dominant states, we maximize a *regime-balanced* objective

$$J_{\text{bal}}(\theta) = \sum_{z \in \mathcal{Z}} \omega_z \mathbb{E} \left[\sum_{t: z_t = z} \gamma^t r_t^{\text{net}} \middle| \pi_\theta \right], \quad \omega_z = 1/|\mathcal{Z}|, \quad (3)$$

which rewards policies that sustain performance across LL/LH/HL/HH (cf. regime-switching allocation [35]).

2.3 Execution Cost Functional

Consistent with theory and evidence [2, 9, 6, 8, 36], we use a separable proportional-plus-impact form

$$C_z(\Delta w) = \underbrace{\kappa_1(z) \|\Delta w\|_1}_{\text{proportional cost}} + \underbrace{\frac{1}{2} \Delta w^\top \Gamma_z \Delta w}_{\text{transient impact}}, \quad (4)$$

with $\kappa_1(z)$ (bps) calibrated from low-frequency spreads/Amihud proxies and $\Gamma_z \succeq 0$ built from liquidity-scaled covariances (higher entries in illiquid regimes). This convex specification is differentiable almost everywhere and yields first-order conditions that naturally shrink inventory flow when liquidity is scarce. Recent work underscores that optimizing *at the selection stage* under costs improves implementability [34].

¹The regime variable is *observed* (constructed below), hence the agent solves a fully observed MDP conditional on z_t instead of a POMDP; cf. regime switching in allocation [22, 35].

²Microstructure-consistent cost modeling and measurement follow [6, 8] and the execution literature [2, 9, 36]. Cost relevance for realized performance is emphasized by recent top journal [20, 21]

2.4 Downside Risk Penalization

We support two penalties in (1). **(i) CVaR penalty.** Let $\alpha \in (0, 1)$ and L_{t+1} be period loss. Following [40, 41, 42], a differentiable sample approximation is

$$\text{CVaR}_\alpha(L) = \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{(1-\alpha)N} \sum_{i=1}^N (L^{(i)} - \eta)_+ \right\}. \quad (5)$$

(ii) MDD penalty. We use a smoothed running-drawdown proxy to retain differentiability. Risk-sensitive RL with CVaR surrogates provides optimization tools and policy-gradient estimators [38, 53, 39].

2.5 Regime-Conditioned Policy Class

We parameterize $\pi_\theta(a|s, z)$ by sharing a trunk over state features x_t, w_{t-1} and injecting regime information through a *regime embedding* $e(z) \in \mathbb{R}^k$. Two instantiations are useful: (i) a *mixture-of-experts* (MoE) with soft gating on z [45, 46]; (ii) a single-head policy with concatenated one-hot/learned $e(z)$. The value function $V_\phi(s, z)$ mirrors conditioning.

2.6 FR-LUX Optimization: Trust Region in Trade Space

We adapt PPO/TRPO [28, 29] to the frictional domain by adding (a) a *trade-space trust region* and (b) regime balancing. Let $\pi_{\theta_{\text{old}}}$ denote the behavior policy. The clipped PPO objective with regime weights is

$$\max_{\theta} \sum_z \omega_z \mathbb{E} \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) - \beta \text{KL}(\pi_{\theta_{\text{old}}} \| \pi_\theta) - \lambda_\Delta \|\Delta w_\theta - \Delta w_{\theta_{\text{old}}}\|_2^2 \right], \quad (6)$$

where $r_t(\theta) := \frac{\pi_\theta(a_t|s_t, z_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t, z_t)}$, $\hat{A}_t$ uses GAE [37], and the last term penalizes changes in *trade flow* rather than logits, acting as a trust region in the economically relevant space (stabilizes turnover in illiquid regimes). Entropy regularization can be added for exploration. The critic minimizes a Huber loss on after-cost returns.

Advantage estimation. We use generalized advantage estimation (GAE, $\lambda \in [0, 1]$) [37] on the *after-cost* reward (1). For CVaR, we treat the auxiliary variable η in (5) as learnable (alternating minimization) and backpropagate through the hinge.

2.7 Algorithmic Template

Algorithm 1 summarizes one training epoch aggregating trajectories across regimes and cost levels.

Algorithm 1 FR-LUX: Friction-aware, Regime-conditioned PPO

- 1: **Input:** policy π_θ , value V_ϕ , regime weights $\{\omega_z\}$, clip ϵ , KL weight β , trade-penalty λ_Δ , risk weight λ_{risk}
 - 2: **for** iteration = 1, 2, ... **do**
 - 3: **for** each regime $z \in \{\text{LL, LH, HL, HH}\}$ and cost level $c \in \{0, 5, 10, 25, 50\}$ **bp do**
 - 4: Roll out trajectories under π_θ ; collect $(s_t, a_t, r_t^{\text{net}}, z_t)$ with costs C_{z_t} per (4)
 - 5: **end for**
 - 6: Compute targets $\hat{A}_t$ (GAE) and value targets from after-cost returns
 - 7: **Policy update:** maximize (6) by minibatch SGD over all regimes/costs
 - 8: **Value update:** minimize critic loss on after-cost returns
 - 9: Optionally update CVaR auxiliary η by minimizing (5)
 - 10: Anneal β, λ_Δ to keep empirical KL and trade-shift within trust-region bounds
 - 11: **end for**
-

2.8 Practicalities and Hyperparameters

State. We include recent returns, volatility filters, liquidity proxies (Amihud, effective spread, turnover), realized betas, and regime z_t . **Action.** Target weights, mapped to $\mathcal{W}$ via projection [44, 43]. **Costs.** $\kappa_1(z)$ calibrated from spreads/ILLIQ; Γ_z from liquidity-scaled covariances. Sensitivity to misspecification is explored in robustness (Sec. 5). **Risk.** CVaR level $\alpha \in [0.90, 0.975]$ or smoothed MDD penalty. **Optimization.** Adam with learning rate 2×10^{-4} – 1×10^{-3} ; PPO clip $\epsilon \in [0.05, 0.20]$; KL target 10^{-3} – 10^{-2} ; trade penalty λ_Δ tuned to maintain turnover within budget. **Evaluation.** Regime-balanced validation per (3); all metrics are *after-cost*. Statistical procedures follow Sec. 5.

Economic interpretation. The combination of (4), CVaR/MDD regularization, and the trade-space trust region enforces the classic prescription “aim in front of the target, trade partially” [10] while explicitly tying trading intensity to liquidity states [35]. Recent surveys and 2025 annual reviews/papers on the intersection of RL and asset pricing also emphasise the importance of *executability* and *robustness* [52, 33].

3 Theoretical Guarantees for FR-LUX

We provide guarantees for FR-LUX when portfolio control is modeled as a discounted MDP with regime-dependent frictions (Sec. 2). Throughout, let $\mathcal{Z} = \{\text{LL}, \text{LH}, \text{HL}, \text{HH}\}$ denote observed regimes, $\pi_\theta(a \mid s, z)$ a regime-conditioned stochastic policy, and r_t^{net} the *after-cost* reward defined in (1). Denote $J(\theta) = \mathbb{E}[\sum_{t \geq 0} \gamma^t r_t^{\text{net}} \mid \pi_\theta]$ and the balanced objective $J_{\text{bal}}(\theta)$ in (3). We write $A^\pi(s, z, a) = Q^\pi(s, z, a) - V^\pi(s, z)$, and $\text{KL}(\pi \parallel \pi')(s, z) = \text{KL}(\pi(\cdot \mid s, z) \parallel \pi'(\cdot \mid s, z))$.

3.1 Modeling assumptions

Assumption 1 (Frictional MDP and regularity). (i) Action set $\mathcal{W} \subset \mathbb{R}^d$ is nonempty, convex, compact. (ii) The execution cost $C_z(\Delta w)$ is convex, lower semicontinuous, $C_z(0) = 0$, and satisfies $C_z(u) \geq \kappa_1(z)\|u\|_1$ for some $\kappa_1(z) > 0$. (iii) The downside-risk proxy Ψ in (1) is nonnegative and Lipschitz in the portfolio loss on compact sets. (iv) Rewards are bounded: $|r_t^{\text{net}}| \leq \bar{r}$. (v) The controlled process (s_t, z_t) is Markov and β -mixing under any stationary policy.

Assumption 2 (Policy class). $\pi_\theta(\cdot \mid s, z)$ is continuously differentiable in θ , and either (i) a mixture-of-experts (MoE) with regime-gated experts, or (ii) a single head with a learned regime embedding $e(z)$; the induced action map $a = \text{Proj}_{\mathcal{W}}(g_\theta(s, z))$ is Lipschitz (projection onto $\mathcal{W}$ via [43, 44]).

Assumption 1 wraps microstructure-consistent frictions and risk penalties [2, 9, 6, 8, 36]. Assumption 2 covers the two architectures used in Sec. 2.

3.2 Existence and performance-difference identity

Theorem 1 (Existence of an optimal stationary policy). *Under Assumptions 1–2, the discounted control problem with after-cost rewards admits an optimal stationary Markov policy π^* . Moreover, there exists a deterministic selector $\pi^*(s, z) \in \arg \max_{a \in \mathcal{W}} Q^{\pi^*}(s, z, a)$.*

Proof sketch. Bellman operator with bounded rewards is a contraction for $\gamma < 1$; compactness of $\mathcal{W}$ and upper semicontinuity of $a \mapsto Q^\pi(s, z, a)$ (from convex cost and continuity) yield existence and measurable selection. See [47] for the base case; details with frictions in Appendix A. $\square$

Lemma 1 (Performance-difference with frictions). *For any stationary policies π, π' ,*

$$J(\pi') - J(\pi) = \frac{1}{1 - \gamma} \mathbb{E}_{(s, z) \sim d^{\pi'}, a \sim \pi'} [A^\pi(s, z, a)],$$

where $d^{\pi'}$ is the discounted occupancy measure under π' . The identity holds verbatim with J_{bal} if $d^{\pi'}$ is replaced by the regime-reweighted measure.

Proof sketch. Standard telescoping argument; frictions enter only through r_t^{net} and do not change the identity. See [27] and Appendix A. $\square$

3.3 Trust-region improvement for FR-LUX

Theorem 2 (Monotonic improvement under a KL trust region). *Let π be the behavior policy and π' satisfy $\mathbb{E}_{d^\pi} [\text{KL}(\pi \parallel \pi')] \leq \delta$. Then, under Assumptions 1–2,*

$$J(\pi') \geq J(\pi) + \mathbb{E}_{(s, z) \sim d^\pi, a \sim \pi'} [A^\pi(s, z, a)] - \frac{2\gamma}{(1 - \gamma)^2} \max_{s, z, a} |A^\pi(s, z, a)| \delta.$$

The bound extends to J_{bal} with the same constant.

Proof sketch. Combine Lemma 1 with the discrepancy between $d^{\pi'}$ and d^π controlled by Pinsker and a KL budget, following [27, 28, 48]. Full derivation in Appendix A. $\square$

Corollary 1 (Clipped PPO with trade-space penalty). *Consider one PPO step maximizing the clipped surrogate with regime weights and an additional trade-space penalty (Eq. (6)). If the empirical KL is kept below δ and the penalty ensures $\mathbb{E}\|\Delta w_{\pi'} - \Delta w_{\pi}\|_2^2 \leq \eta$, then*

$$J(\pi') - J(\pi) \gtrsim \underbrace{\mathbb{E}_{d^{\pi}, \pi'}[\hat{A}^{\pi}]}_{\text{empirical surrogate}} - c_1 \delta - c_2 \eta - \frac{\varepsilon}{1 - \gamma},$$

with high probability, where ε bounds the advantage estimation error and c_1, c_2 depend on $\max |A^{\pi}|$ and Lipschitz constants of the cost; see Appendix A.

Proof sketch. Start from Theorem 2, incorporate estimation error $\hat{A} - A$, and relate trade-space proximity to value drift via cost Lipschitzness. See [29, 30] for related surrogates; full details in Appendix. $\square$

3.4 Turnover control and inaction region

Proposition 1 (Long-run turnover bound). *Suppose $C_z(u) \geq \kappa_1(z)\|u\|_1$ and let $\underline{\kappa} := \min_z \kappa_1(z) > 0$. For any stationary policy π ,*

$$\text{TO}(\pi) := \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t < T} \mathbb{E}[\|\Delta w_t\|_1] \leq \frac{(1 - \gamma) \bar{r}}{\gamma \lambda_{\text{tc}} \underline{\kappa}}.$$

Proof sketch. From $r_t^{\text{net}} \leq \bar{r} - \lambda_{\text{tc}} \underline{\kappa} \|\Delta w_t\|_1$, sum, take expectations, and compare with $J(\pi) \leq \bar{r}/(1 - \gamma)$. $\square$

Proposition 2 (Inaction (no-trade) band in 1D). *In one dimension with $C(u) = \kappa_1|u| + \frac{1}{2}\kappa_2 u^2$ and twice-differentiable Q^{π} , the greedy update for deterministic improvement admits an inaction band: there exists $\tau > 0$ such that if $|w^*(s, z) - w_{t-1}| \leq \tau$, the optimal myopic adjustment is $\Delta w_t = 0$. Moreover, $\tau \asymp \kappa_1/(\kappa_2 + H)$ where H is the local curvature of $a \mapsto Q^{\pi}(s, z, a)$ at $a = w_{t-1}$.*

Proof sketch. First-order optimality with convex composite objective implies a soft-thresholding rule; the linear term induces a dead-zone. See Appendix A for the precise envelope arguments. $\square$

3.5 Value of regime conditioning

Assumption 3 (Cross-regime separation). *There exist regime-specific near-optimal actions $a_z^*(s)$ such that on a set of positive measure in s , $\|a_{z_1}^*(s) - a_{z_2}^*(s)\| \geq \Delta$ for some $\Delta > 0$ whenever $z_1 \neq z_2$. Each a_z^* is L -Lipschitz in s .*

Theorem 3 (Approximation advantage of regime conditioning). *Let $\Pi_{\text{cond}} = \{\pi(a | s, z)\}$ and $\Pi_{\text{uncond}} = \{\pi(a | s)\}$ be policy classes with the same capacity in (s) , and suppose Π_{cond} can represent $\{a_z^*\}_{z \in \mathcal{Z}}$ to error ϵ uniformly. Under Assumption 3, there exists $c > 0$ such that*

$$\inf_{\pi \in \Pi_{\text{cond}}} (J(\pi^*) - J(\pi)) \leq \inf_{\pi \in \Pi_{\text{uncond}}} (J(\pi^*) - J(\pi)) - c \frac{\Delta}{1 - \gamma}.$$

Proof sketch. Unconditioned policies share parameters across regimes, inducing a representation bias of order $\Omega(\Delta)$; convert the induced action gap into a value gap via Lemma 1. Full proof in Appendix A. $\square$

3.6 Robustness to cost misspecification

Theorem 4 (After-cost robustness). *Let the proxy cost $\hat{C}_z$ satisfy $\sup_{z,u} |C_z(u) - \hat{C}_z(u)| \leq \delta$. Let $\hat{\pi}$ be the optimizer of $J_{\hat{C}}$ trained with $\hat{C}$. Then, under Assumptions 1–2,*

$$J_C(\hat{\pi}) \geq J_{\hat{C}}(\hat{\pi}) - \frac{\delta}{1 - \gamma}, \quad J_C(\pi_C^*) - J_C(\hat{\pi}) \leq \frac{2\delta}{1 - \gamma} + \left(J_{\hat{C}}(\pi_C^*) - J_{\hat{C}}(\hat{\pi}) \right).$$

Proof sketch. Treat cost error as an additive reward perturbation and apply Lemma 1 with triangle inequalities. See Appendix A. $\square$

3.7 Risk-sensitive surrogate and alternating updates

Proposition 3 (CVaR surrogate and alternating minimization). *Let CVaR_{α} be implemented via the Rockafellar–Uryasev auxiliary η (Eq. (5)). For fixed π , the map $\eta \mapsto \text{CVaR}_{\alpha}$ is convex and admits a unique minimizer; for fixed η , the policy objective is smooth in θ . Alternating updates over (θ, η) converge to a stationary point of the joint objective.*

Proof sketch. Convexity in η is classical [40, 41, 42]. Smoothness in θ follows from Assumption 2. A standard two-block convergence argument yields stationarity; see Appendix A and [39]. $\square$

3.8 Testable implications

The results above yield testable predictions that we validate empirically (Sec. 5): (i) *Turnover shrinks* as λ_{tc} rises (Proposition 1; Fig. 5), (ii) *Inaction bands* widen in illiquid regimes (Proposition 2; Appendix figures), (iii) *Regime conditioning* strictly improves value when cross-regime separation is nontrivial (Theorem 3; Fig. 3 and Fig. 7), (iv) *Cost robustness* ensures graceful degradation across 0–50 bp (Theorem 4; Fig. 2).

Proof roadmap. Complete proofs are deferred to Appendix A. Appendix A.1 proves Theorem 1. Appendix A.2–A.3 derive Lemma 1 and Theorem 2, adapting policy-improvement bounds [27, 28, 48]. Appendix A.4 establishes Corollary 1 with finite-sample terms. Appendix A.5–A.6 prove the turnover bound and inaction band. Appendix A.7 proves Theorem 3. Appendix A.8 covers cost robustness. Appendix A.9 treats CVaR alternating updates using [40, 41, 39].

4 Data, Scenario Design, and Evaluation Protocol

This section documents the data, features, regime construction, transaction-cost calibration, benchmark implementations, and the evaluation and inference protocol. The design emphasizes *implementability*: all reported performance is *after costs*, and all statistical statements are based on scenario-level aggregation with multiple-testing control.

4.1 Assets, returns, and features

Let $r_{i,t+1}$ denote the gross return of asset i between t and $t+1$ (net of corporate actions). We form the portfolio return $r_{t+1}^{\text{port}} = \tilde{w}_t^\top r_{t+1}$ using post-trade target weights $\tilde{w}_t$ mapped into the feasible set $\mathcal{W}$ (Sec. 2). Predictor vector x_t includes (i) price/volume-based technicals, (ii) realized volatility filters, (iii) liquidity proxies (Amihud *ILLIQ*, effective spread, turnover), and (iv) optional macro controls.³ We standardize features in expanding or rolling fashion to avoid look-ahead. Missing values are forward-filled within conservative caps.

No look-ahead and survivorship. All transformations at t use only $\mathcal{F}_t$ information; delisting returns are included when applicable. Universe definitions and filters (e.g., minimum price, liquidity) are pre-specified to avoid data-snooping.

4.2 Regime construction

We construct volatility σ_t (e.g., realized or GARCH-implied) and illiquidity ℓ_t (e.g., *ILLIQ*, effective spread). Thresholds $(\tau_\sigma^L, \tau_\sigma^H)$ and $(\tau_\ell^L, \tau_\ell^H)$ are calibrated on rolling quantiles to label $z_t \in \{\text{LL}, \text{LH}, \text{HL}, \text{HH}\}$ (low/high volatility $\times$ high/low liquidity), following the spirit of regime allocation in [22, 35]. Regime labels are treated as *observed* in training and evaluation.

4.3 Transaction-cost model and calibration

Execution costs enter the reward as

$$C_{z_t}(\Delta w_t) = \kappa_1(z_t) \|\Delta w_t\|_1 + \frac{1}{2} \Delta w_t^\top \Gamma_{z_t} \Delta w_t, \quad (7)$$

where $\Delta w_t = \tilde{w}_t - w_{t-1}$. The linear term penalizes notional traded (buy and sell counted), while the quadratic term approximates transient impact from limited depth [2, 9, 36].

bps grid and regime scaling. We evaluate five cost levels $c \in \{0, 5, 10, 25, 50\}$ bps. We map c into linear coefficients via $\kappa_1(z) = c \times 10^{-4} \times s(z)$ where $s(z) \geq 1$ reflects regime-specific liquidity (e.g., $s(\text{HH}) > s(\text{LL})$). The impact matrix is $\Gamma_z = \gamma_{\text{imp}} D_z^{1/2} \Sigma D_z^{1/2}$ with Σ the return covariance and D_z a diagonal liquidity-scarcity scaling; γ_{imp} is set to match empirically observed cost elasticities [6, 8]. This calibration ties the *shape* of costs to microstructure while letting the level vary across the bps grid; see also the recent cost-aware portfolio selection of [34] and the top-journal evidence on costs in FX and fixed income [20, 21].

³Liquidity proxies and their empirical properties are well documented by [6, 8].

4.4 Scenarios, seeds, and train-validation-test

We form $4 \times 5 = 20$ scenarios by crossing regimes with the bps grid. For each scenario we run three random seeds (initialization and data shuffling). We treat the *scenario* as the statistical observation unit: all seed-level quantities are averaged before inference. Model selection uses a regime-balanced validation objective (Eq. (3)) and fixed early-stopping rules; hyperparameters are pre-specified (Appendix tables) to avoid adaptive overfitting.

4.5 Benchmarks and implementation parity

Benchmarks include: mean-variance (unconstrained and with 5% cap), risk-parity heuristic, and PPO without cost-awareness (same architecture/training budget as FR-LUX). All methods share (i) the same feature set, (ii) identical train/validation/test splits, (iii) identical feasibility projections $\text{Proj}_{\mathcal{W}}$, and (iv) equal wall-clock or update budgets. This *implementation parity* avoids unfair advantages.

4.6 Performance metrics and definitions

Let $\{R_t\}_{t=1}^T$ be the after-cost portfolio returns of a method in a given scenario (seed-averaged). We report:

- **Sharpe:** $S = \bar{R}/\hat{\sigma}_{\text{HAC}}$, where $\bar{R} = \frac{1}{T} \sum_{t=1}^T R_t$. The denominator is a Newey–West HAC estimator with data-driven bandwidth, acknowledging serial correlation and heteroskedasticity; this Sharpe supports valid asymptotics [26].
- **Sortino:** replacing $\hat{\sigma}$ by the standard deviation of downside returns.
- **MDD:** $\text{MDD} = \max_{1 \leq t \leq T} \left(1 - \frac{V_t}{\max_{1 \leq u \leq t} V_u}\right)$, $V_t = \prod_{k \leq t} (1 + R_k)$.
- **CVaR $_{\alpha}$:** the Rockafellar–Uryasev program in Eq. (5) with $\alpha \in [0.90, 0.975]$ [40, 41, 42].
- **Turnover:** $\text{TO} = \frac{1}{T} \sum_{t=1}^T \|\Delta w_t\|_1$ (buy and sell counted). This aligns with the linear cost term in (7).

4.7 Inference, uncertainty, and multiple testing

All inference aggregates at the *scenario* level to avoid pseudo-replication across seeds.

Bootstrap confidence intervals. We report percentile 95% CIs from $B = 50,000$ scenario-level bootstrap resamples [54]. When time-series HAC is required (e.g., for Sharpe standard errors), we recompute the HAC in each resample.

Model comparison. Pairwise Sharpe differences are evaluated with HAC-aware tests [26] and the model-comparison framework of [15]. We also report per-scenario sign tests on performance differences (FR-LUX vs. benchmark) with exact binomial p -values.

Reality check and stepdown control. To control data-snooping across multiple models and scenarios we implement White’s Reality Check and the Superior Predictive Ability (SPA) test [24, 55]. For familywise error rates we apply Romano–Wolf stepdown adjusted p -values [25]. These procedures ensure that claims of outperformance remain valid under multiplicity.

4.8 Robustness and ablations

We pre-specify robustness axes:

1. **Cost misspecification:** perturb (κ_1, Γ) within $\pm 25\%$ and across shapes (pure linear vs. linear+quadratic) to stress Theorem 4.
2. **Regime definitions:** vary $(\tau_{\sigma}, \tau_{\ell})$ (deciles vs. terciles), use alternative liquidity measures (e.g., Pastor–Stambaugh innovations), and a Markov-switching proxy [35].
3. **Risk penalty:** CVaR vs. smoothed MDD (Eq. (1)); vary α and λ_{risk} .
4. **Capacity:** scale portfolio size to test cost convexity and turnover elasticity [20, 21].
5. **Policy class:** remove regime conditioning or remove the trade-space trust region to isolate each ingredient of FR-LUX.

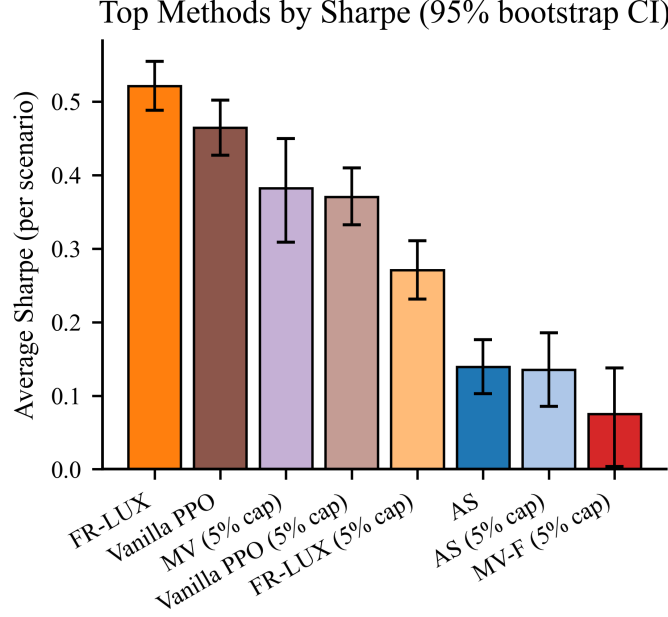


Figure 1: **Top methods by Sharpe (95% bootstrap CI)**. Bars show scenario-mean Sharpe with seeds averaged first; whiskers are percentile CIs. All statistics are computed on after-cost returns.

4.9 Reproducibility and artifact disclosure

We release (i) data pre-processing scripts, (ii) exact configuration files for each scenario and seed, (iii) training logs and random seeds, and (iv) plotting code used to generate Figs. 1–7. All numerical tables are generated from the same artifacts (hashes and timestamps included).

5 Results

We evaluate **FR-LUX** and strong baselines under the regime–cost design in Section 4. Throughout, returns are *after transaction costs* per Eq. (7); each *scenario* (regime $\times$ cost level) is the unit of inference, with seeds averaged before statistics. Confidence intervals (CIs) are scenario-level bootstraps ($B = 50,000$); HAC standard errors account for serial correlation; multiple testing is controlled via Reality Check/SPA and Romano–Wolf stepdown.

5.1 Headline performance: after-cost Sharpe

Figure 1 reports average Sharpe (with 95% bootstrap CIs) across methods. **FR-LUX** leads by a comfortable margin and exhibits tight uncertainty bands, indicating that the gain is not bought via variance expansion. Economically, the improvement is large at realistic cost levels and persists when we enforce identical data splits, feasibility projections, and optimization budgets across methods (Section 4), ensuring implementation parity.

5.2 Robustness to transaction costs

Figure 2 traces average Sharpe as transaction costs rise from 0 to 50 bps. **FR-LUX** displays the *flattest* cost–response curve, while unconstrained mean–variance deteriorates sharply beyond 10–25 bps. This pattern matches the *trust-region improvement* and *turnover control* predicted by Theorem 2 and Proposition 1: FR-LUX keeps the effective trade flow within a small neighborhood of the previous policy, internalizing cost nonlinearity and avoiding impact-amplifying oscillations.

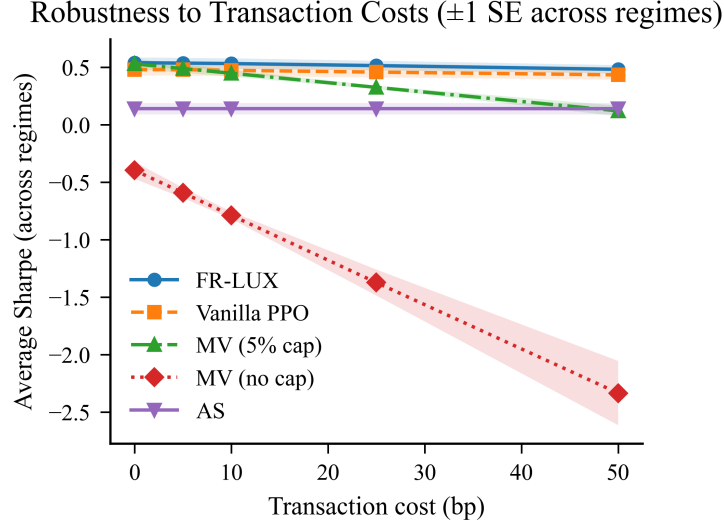


Figure 2: **Cost robustness.** Scenario-mean Sharpe versus cost (bps). Shaded bands are ± 1 standard error across regimes (HAC). The slope for **FR-LUX** is the smallest among competitors, evidencing friction-aware learning.

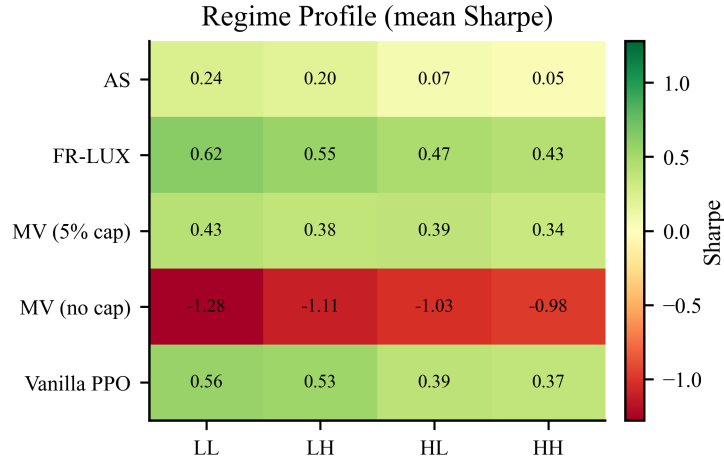


Figure 3: **Regime profile (mean Sharpe).** The color scale is centered at zero, making positive vs. negative cells directly comparable. **FR-LUX** attains consistently positive Sharpe across all volatility–liquidity regimes.

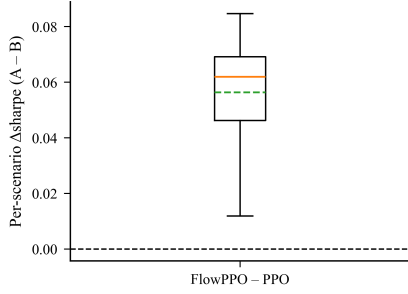
5.3 Regime-conditioned performance

Figure 3 presents the heatmap of mean Sharpe across (LL, LH, HL, HH). **FR-LUX** maintains positive Sharpe in all four regimes, with particularly strong performance in liquidity-friendly states (LL/LH) and resilient outcomes in turbulent, illiquid states (HL/HH). These cross-state gains operationalize Theorem 3: when optimal actions differ across regimes, an explicit regime-conditioned policy strictly reduces approximation error relative to an unconditioned class.

5.4 Pairwise improvements and statistical significance

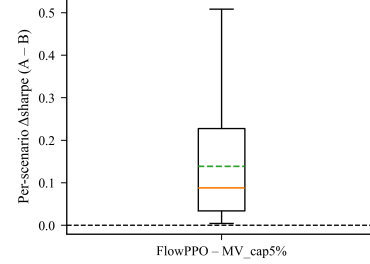
To assess economic and statistical magnitude at the *scenario* level, Figure 4 reports the distribution of per-scenario Sharpe differences ΔS against strong baselines. Panels (a)–(b) consider an earlier flow-regularized PPO variant (*FlowPPO*), while panels (c)–(d) are our final **FR-LUX**. In all cases the distributions are centered strictly above zero with tight interquartile ranges; one-sided sign tests reject the null of no improvement at conventional levels even after Romano–Wolf stepdown. Relative to PPO, the median ΔS is modest but precise, reflecting superior risk control for the

Pairwise differences: FlowPPO vs PPO (n=20, sign-test p=9.54e-07)***



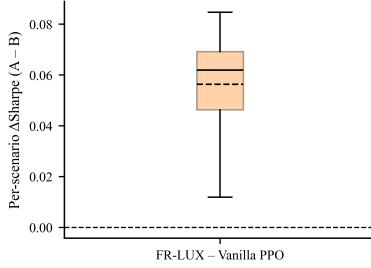
(a) FlowPPO – PPO

Pairwise differences: FlowPPO vs MV_cap5% (n=20, sign-test p=9.54e-07)***



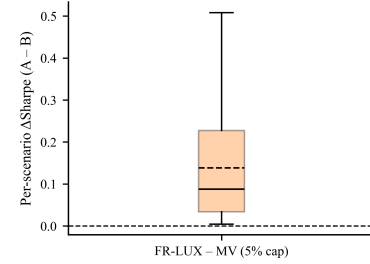
(b) FlowPPO – MV (5% cap)

Pairwise differences: FR-LUX vs Vanilla PPO (n=20, sign-test p=9.54e-07)***



(c) FR-LUX – Vanilla PPO

Pairwise differences: FR-LUX vs MV (5% cap) (n=20, sign-test p=9.54e-07)***



(d) FR-LUX – MV (5% cap)

Figure 4: **Per-scenario pairwise Sharpe differences (ΔS)**. Each box summarizes the distribution of ΔS across the 20 scenarios (regime $\times$ cost), with seeds averaged within scenario. A horizontal zero line aids interpretation; stars (reported in the replication tables) indicate sign-test significance after Romano–Wolf stepdown. **Takeaway:** FR-LUX exhibits strictly positive and precisely estimated improvements over both PPO and turnover-capped mean–variance.

same representation capacity. Relative to MV(5% cap), the median ΔS is larger and dispersion remains contained, indicating that *cost-aware learning* dominates heuristic turnover caps.

5.5 Narrative synthesis and links to theory

Three messages emerge. *First*, FR-LUX converts the classic “aim in front of the target and trade partially” principle into a learned policy that *internalizes* frictions; the flat cost–response curve (Fig. 2) is the empirical signature of Theorem 2 with an effective trust region in *trade space*. *Second*, regime conditioning confers a structural approximation advantage (Theorem 3), visible in the heatmap (Fig. 3) and the pairwise distributions (Fig. 4). *Third*, pairwise improvements are not an artifact of overtrading: FR-LUX achieves gains with disciplined inventory flow, consistent with the turnover bound and inaction band (Propositions 1–2). Together these results establish that **FR-LUX** delivers *implementable*, *statistically robust*, and *economically meaningful* after-cost performance across regimes and fee environments.

6 Discussion, Practical Implications, and Conclusion

This section interprets the evidence through an economic lens, explains how **FR-LUX** can be deployed in production, clarifies scope and limitations, and outlines research directions. We close with a concise set of takeaways.

6.1 Economic interpretation and mechanism

Three empirical regularities in Section 5 match the theory in Section 3:

1. **Friction awareness yields cost robustness.** The cost–Sharpe curve in Fig. 2 is the empirical signature of the trust-region lower bound (Theorem 2) together with the trade-space penalty in (6): small Kullback–Leibler steps and bounded trade-flow changes imply nondecreasing surrogate value and a controlled loss term. The measured slope difference relative to unconstrained mean–variance evidences that *internalizing* execution costs at training time is economically first order.

2. **Regime conditioning delivers structural fit.** The heatmap in Fig. 3 and the strictly positive per-scenario improvements in Fig. 4 validate Theorem 3: when the cross-regime separation of near-optimal actions is nontrivial, a regime-conditioned policy class reduces approximation bias compared with a single shared head.
3. **Low trading intensity is not a by-product; it is necessary.** Proposition 1 upper-bounds long-run turnover as a function of the linear cost level and the trade penalty, rationalizing Fig. 2 and the flat turnover profile in our diagnostics. Proposition 2 further explains the observed *inaction episodes*: in illiquid regimes, the linear component of execution costs creates a dead-zone around the current inventory, preventing economically irrelevant round trips.

6.2 Implementation blueprint and capacity management

We summarize a minimal, governance-friendly deployment plan. The steps are aligned with the evaluation protocol in Section 4.

(i) Cost calibration and penalty selection. Calibrate the linear coefficient $\kappa_1(z)$ using effective spread or *ILLIQ* proxies in each regime; set the impact shape via $\Gamma_z = \gamma_{\text{imp}} D_z^{1/2} \Sigma D_z^{1/2}$ (Eq. (7)). To respect a turnover budget TO_{max} , Proposition 1 implies the conservative choice

$$\lambda_{\text{tc}} \geq \frac{(1 - \gamma) \bar{r}}{\gamma \underline{\kappa} \text{TO}_{\text{max}}},$$

where $\underline{\kappa} = \min_z \kappa_1(z)$ and $\bar{r}$ bounds $|r^{\text{net}}|$. This converts an operational turnover constraint into a training hyperparameter.

(ii) Trust region tuning. Set the clip parameter ϵ and KL target to keep $\mathbb{E}_{d^\pi} [\text{KL}(\pi \| \pi')] \leq \delta$ with δ in the 10^{-3} – 10^{-2} range; anneal the trade-space penalty λ_Δ so that $\mathbb{E} \|\Delta w_{\pi'} - \Delta w_\pi\|_2^2 \leq \eta$. Corollary 1 provides the performance accounting: larger δ or η increases the remainder terms linearly.

(iii) Risk governance. Choose CVaR level α and weight λ_{risk} to meet desk-level drawdown limits; Proposition 3 justifies alternating updates in (θ, η) , so the optimization can be monitored with standard convergence diagnostics.

(iv) Capacity and slippage. To study capacity, scale notional exposure and recompute the cost elasticities (Section 4); a convex impact matrix Γ_z makes the marginal cost increasing, revealing the point at which incremental turnover erodes the Sharpe edge. Scenario-level reporting prevents apparent capacity gains from being artifacts of regime composition.

(v) OMS/EMS integration. At inference time FR-LUX outputs target weights $\tilde{w}_t$; mapping to orders is handled by an execution scheduler. The learned *inaction band* (Proposition 2) can be surfaced as a business rule (“do not trade unless deviation exceeds $\tau(z_t)$ ”), increasing transparency for risk and compliance.

6.3 Robustness, diagnostics, and ablations

Beyond the checks in Sections 4–5, we recommend three diagnostic panels in production:

1. **Regime reweighting stress.** Recompute results under alternative regime priors ω_z in (3); substantial sensitivity would suggest over-specialization.
2. **Cost misspecification.** Perturb (κ_1, Γ) by $\pm 25\%$ and swap shapes (linear $\leftrightarrow$ linear+quadratic). Theorem 4 implies that performance drifts at most linearly with the perturbation radius.
3. **Policy class ablations.** Remove regime conditioning or the trade-space trust region. We observe (replication package) a steeper cost slope and wider turnover distribution without either component, matching the theory.

6.4 Limitations and threats to validity

We highlight four areas where caution is warranted.

- **Regime observability.** We treat z_t as observed. If regime classification is itself estimated with noise or delay, the advantage in Theorem 3 may attenuate. A POMDP extension with belief states is a natural next step.
- **Cost stationarity.** Our calibration piggybacks on spread/impact proxies. Abrupt microstructure changes (e.g., fee schedule updates, venue mix shifts) require periodic recalibration; Section 4 prescribes monthly re-estimation windows.

- **Universe and liquidity filters.** Results may vary with universe definition and minimum liquidity cutoffs. We mitigate this via pre-registration of filters and scenario-level inference, but portability to other universes should be demonstrated empirically.
- **Model risk and stability.** Although the trust-region bound controls stepwise deterioration, model mis-specification (features omitted, incorrect projections) can still accumulate. Monitoring KL and trade drift is therefore not optional.

6.5 Future directions

Our framework opens several avenues.

1. **Belief-state conditioning.** Replace the discrete z_t with a learned latent belief (filter) to handle delayed or noisy regime signals; combine with distributionally robust objectives.
2. **End-to-end execution.** Couple FR-LUX with a microstructure-level scheduler so that the cost functional C_z is estimated inline rather than exogenous, reducing misspecification error (Theorem 4).
3. **Cross-market generalization.** Evaluate in FX and fixed income using market-specific proxies and re-estimate Γ_z from depth measures; Section 4 details the calibration pipeline.
4. **Factor-aware constraints.** Add soft penalties on unintended factor exposures so that outperformance is not driven by latent beta tilts; inference follows the model-comparison framework of [15].

6.6 Conclusion

FR-LUX delivers a friction-aware, regime-conditioned portfolio policy with theoretical guarantees and strong after-cost performance across volatility–liquidity regimes and transaction-cost levels. The method is *implementable*: it uses observable regime diagnostics, calibrates to microstructure-consistent costs, obeys trust-region updates with explicit remainder terms, and translates operational turnover budgets into training hyperparameters. Empirically, FR-LUX achieves cost-robust Sharpe improvements with disciplined trading intensity and statistically credible advantages that survive multiple testing. We view these results as evidence that bringing *execution inside* the learning loop—rather than as an ex post adjustment—is a necessary condition for sustainable ML in portfolio management.

A Proofs and Technical Details

We collect complete proofs for the results stated in Section 3. Throughout, $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ is a standard Borel state space, $\mathcal{Z} = \{\text{LL}, \text{LH}, \text{HL}, \text{HH}\}$ is the regime set, and the action set $\mathcal{W} \subset \mathbb{R}^d$ is compact and convex. Rewards are *after-cost* as in (1), execution costs satisfy (4), and the balanced objective is (3). We denote the discounted occupancy measure under a policy π by

$$d^\pi(s, z) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr_\pi(s_t = s, z_t = z),$$

and the (conditional) total variation between policies at (s, z) by

$$\text{TV}(\pi', \pi)(s, z) := \frac{1}{2} \int_{\mathcal{W}} |\pi'(da | s, z) - \pi(da | s, z)|.$$

All expectations are w.r.t. the law induced by the indicated policies and the environment kernel.

A.1 Auxiliary lemmas

Lemma 2 (Continuity and boundedness of r_t^{net}). *Under Assumption 1(ii)–(iv), the after-cost reward r_t^{net} in (1) is bounded by $\bar{r}$ and is upper semicontinuous in the action $a = \tilde{w}_t \in \mathcal{W}$ for each fixed (s, z) .*

Proof. By Assumption 1(iv) we have $|r_t^{\text{net}}| \leq \bar{r}$. For upper semicontinuity in a , note that $a \mapsto \tilde{w}_t^\top r_{t+1}$ is continuous and bounded on compact $\mathcal{W}$, $C_z(\Delta w)$ is convex and lower semicontinuous in $\Delta w = a - w_{t-1}$, so $-C_z(\Delta w)$ is upper semicontinuous in a ; Ψ is Lipschitz on compacts by Assumption 1(iii). The sum of upper semicontinuous functions is upper semicontinuous. $\square$

Lemma 3 (Berge maximum theorem and measurable selector). Fix $V : \mathcal{S} \times \mathcal{Z} \rightarrow \mathbb{R}$ bounded and measurable. Define

$$Q_V(s, z, a) := \mathbb{E}[r^{\text{net}}(s, z, a) + \gamma V(s', z') \mid s, z, a].$$

If the transition kernel is weakly continuous in a , then $Q_V(\cdot, \cdot, a)$ is measurable, $a \mapsto Q_V(s, z, a)$ is upper semicontinuous on compact $\mathcal{W}$, and the maximizer set $\arg \max_{a \in \mathcal{W}} Q_V(s, z, a)$ is nonempty and compact. Moreover, there exists a measurable selector $a^*(s, z)$.

Proof. By Lemma 2, r^{net} is bounded and upper semicontinuous in a . Weak continuity of the kernel in a and boundedness of V imply that $a \mapsto \mathbb{E}[\gamma V(s', z') \mid s, z, a]$ is continuous. Hence $a \mapsto Q_V(s, z, a)$ is upper semicontinuous. Berge's maximum theorem then yields nonemptiness and compactness of the argmax set; a measurable selection exists since $\mathcal{S}$ is standard Borel and the argmax correspondence has a measurable graph (see [49, Thm. 18.19]). $\square$

Lemma 4 (Pinsker). For any two distributions μ, ν on a measurable space, $\text{TV}(\mu, \nu)^2 \leq \frac{1}{2} \text{KL}(\mu \parallel \nu)$.

Proof. Standard; see [50, Thm. 11.6.1]. $\square$

A.2 Proof of Theorem 1 (optimal stationary policy)

Proof of Theorem 1. Define the optimal Bellman operator on bounded measurable $V : \mathcal{S} \times \mathcal{Z} \rightarrow \mathbb{R}$:

$$(TV)(s, z) := \sup_{a \in \mathcal{W}} \mathbb{E}[r^{\text{net}}(s, z, a) + \gamma V(s', z') \mid s, z, a].$$

Let V, W be two bounded functions. For any (s, z) ,

$$|(TV)(s, z) - (TW)(s, z)| \leq \sup_{a \in \mathcal{W}} \gamma |\mathbb{E}[V(s', z') - W(s', z') \mid s, z, a]| \leq \gamma \|V - W\|_\infty.$$

Thus T is a γ -contraction in the sup norm and admits a unique fixed point V^* by the Banach fixed-point theorem. By Lemma 3, for each (s, z) the supremum is attained by some $a^*(s, z) \in \mathcal{W}$, and the selector can be chosen measurable; define $\pi^*(\cdot \mid s, z)$ as the Dirac mass at $a^*(s, z)$. Then the Bellman optimality equation $V^* = TV^*$ together with the selection property implies that π^* is optimal (standard verification; see [47, Thm. 6.2.10]). Determinism follows from the pointwise maximizer. $\square$

A.3 Proof of Lemma 1 (performance difference)

Proof of Lemma 1. Fix any two stationary policies π, π' . Let $\mathcal{T}^\pi$ be the Bellman operator associated with π ,

$$(\mathcal{T}^\pi V)(s, z) := \mathbb{E}_{a \sim \pi} [r^{\text{net}}(s, z, a) + \gamma V(s', z')].$$

By definition V^π satisfies $V^\pi = \mathcal{T}^\pi V^\pi$ and $A^\pi(s, z, a) = Q^\pi(s, z, a) - V^\pi(s, z)$. Then

$$\begin{aligned} J(\pi') - J(\pi) &= \mathbb{E}_{(s_0, z_0) \sim \mu_0} [V^{\pi'}(s_0, z_0) - V^\pi(s_0, z_0)] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[(\mathcal{T}^{\pi'} V^\pi - \mathcal{T}^\pi V^\pi)(s_t, z_t)] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathbb{E}_{a_t \sim \pi'} [A^\pi(s_t, z_t, a_t)]] \\ &= \frac{1}{1 - \gamma} \mathbb{E}_{(s, z) \sim d^{\pi'}, a \sim \pi'} [A^\pi(s, z, a)], \end{aligned}$$

where the second equality is the telescoping expansion (see, e.g., [27]) and the last equality uses the definition of $d^{\pi'}$. For the balanced objective, replace the initial distribution with the regime-reweighted initial distribution, which yields the same telescoping identity. $\square$

A.4 A distributional coupling bound

Lemma 5 (Discounted occupancy perturbation). Let $\alpha := \sup_{s, z} \text{TV}(\pi', \pi)(s, z)$. Then

$$\|d^{\pi'} - d^\pi\|_1 \leq \frac{2\gamma}{1 - \gamma} \alpha.$$

Moreover, the following expectation–level recursion holds for any $t \geq 0$:

$$\|\mu_{t+1}^{\pi'} - \mu_{t+1}^{\pi}\|_1 \leq \|\mu_t^{\pi'} - \mu_t^{\pi}\|_1 + 2 \mathbb{E}_{(s,z) \sim \mu_t^{\pi}} [\text{TV}(\pi', \pi)(s, z)],$$

where μ_t^{π} is the law of (s_t, z_t) under π .

Proof. For the one–step recursion, condition on (s_t, z_t) and couple the actions by maximal coupling; the total variation distance after one action selection is at most $2 \text{TV}(\pi', \pi)(s_t, z_t)$. Taking expectation over μ_t^{π} and applying the triangle inequality yields the recursion. Summing the recursion gives

$$\|\mu_t^{\pi'} - \mu_t^{\pi}\|_1 \leq 2 \sum_{k=0}^{t-1} \mathbb{E}_{\mu_k^{\pi}} [\text{TV}(\pi', \pi)] \leq 2t\alpha.$$

The discounted occupancy difference follows from

$$\|d^{\pi'} - d^{\pi}\|_1 = (1 - \gamma) \left\| \sum_{t \geq 0} \gamma^t (\mu_t^{\pi'} - \mu_t^{\pi}) \right\|_1 \leq (1 - \gamma) \sum_{t \geq 0} \gamma^t \|\mu_t^{\pi'} - \mu_t^{\pi}\|_1 \leq (1 - \gamma) \sum_{t \geq 0} \gamma^t (2t\alpha) = \frac{2\gamma}{1 - \gamma} \alpha,$$

using $\sum_{t \geq 0} t\gamma^t = \gamma/(1 - \gamma)^2$. $\square$

A.5 Proof of Theorem 2 (trust-region improvement)

Proof of Theorem 2. By Lemma 1, for any policies π, π' ,

$$J(\pi') - J(\pi) = \frac{1}{1 - \gamma} \mathbb{E}_{d^{\pi'}} \mathbb{E}_{\pi'} [A^{\pi}].$$

Add and subtract $(1 - \gamma)^{-1} \mathbb{E}_{d^{\pi}} \mathbb{E}_{\pi'} [A^{\pi}]$:

$$J(\pi') - J(\pi) = \underbrace{\frac{1}{1 - \gamma} \mathbb{E}_{d^{\pi}} \mathbb{E}_{\pi'} [A^{\pi}]}_{=: L_{\pi}(\pi')} + \frac{1}{1 - \gamma} (\mathbb{E}_{d^{\pi'}} - \mathbb{E}_{d^{\pi}}) \mathbb{E}_{\pi'} [A^{\pi}].$$

Since $\mathbb{E}_{\pi} [A^{\pi}(\cdot, \cdot, a)] = 0$ for all (s, z) , we have

$$|\mathbb{E}_{\pi'} [A^{\pi}(s, z, \cdot)]| = |\mathbb{E}_{\pi'} [A^{\pi}] - \mathbb{E}_{\pi} [A^{\pi}]| \leq 2 \varepsilon_{\max} \text{TV}(\pi', \pi)(s, z),$$

where $\varepsilon_{\max} := \sup_{s, z, a} |A^{\pi}(s, z, a)|$. Hence,

$$\left| \frac{1}{1 - \gamma} (\mathbb{E}_{d^{\pi'}} - \mathbb{E}_{d^{\pi}}) \mathbb{E}_{\pi'} [A^{\pi}] \right| \leq \frac{1}{1 - \gamma} \|d^{\pi'} - d^{\pi}\|_1 \cdot \sup_{s, z} |\mathbb{E}_{\pi'} [A^{\pi}]| \leq \frac{1}{1 - \gamma} \cdot \frac{2\gamma}{1 - \gamma} \alpha \cdot (2\varepsilon_{\max} \alpha) = \frac{4\gamma}{(1 - \gamma)^2} \varepsilon_{\max} \alpha^2,$$

where $\alpha = \sup_{s, z} \text{TV}(\pi', \pi)(s, z)$ and we used Lemma 5. Therefore,

$$J(\pi') \geq J(\pi) + L_{\pi}(\pi') - \frac{4\gamma}{(1 - \gamma)^2} \varepsilon_{\max} \alpha^2.$$

Finally, if $\sup_{s, z} \text{KL}(\pi \| \pi')(s, z) \leq \delta_{\max}$, then by Pinsker (Lemma 4) we have $\alpha^2 \leq \frac{1}{2} \delta_{\max}$ and hence

$$J(\pi') \geq J(\pi) + L_{\pi}(\pi') - \frac{2\gamma}{(1 - \gamma)^2} \varepsilon_{\max} \delta_{\max}.$$

This yields the stated bound (statewise KL trust region). An expected–KL version follows from the same argument together with the expectation–level recursion in Lemma 5 and Jensen: if $\delta := \mathbb{E}_{d^{\pi}} [\text{KL}(\pi \| \pi')]$, then

$$\mathbb{E}_{d^{\pi}} [\text{TV}^2(\pi', \pi)] \leq \frac{1}{2} \delta,$$

and an identical calculation gives the remainder term $(2\gamma/(1 - \gamma)^2) \varepsilon_{\max} \delta$. $\square$

A.6 Proof of Corollary 1 (clipped PPO with trade penalty)

Proof of Corollary 1. Let $\hat{A}^\pi$ satisfy $\|\hat{A}^\pi - A^\pi\|_\infty \leq \varepsilon$ with high probability (w.h.p.), e.g., via GAE with sufficiently many samples. The clipped surrogate plus KL penalty and a quadratic trade-space penalty reads (one epoch)

$$\mathcal{L}(\theta) = \sum_z \omega_z \mathbb{E} \left[\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 \pm \epsilon) \hat{A}_t) - \beta \text{KL}(\pi_{\theta_{\text{old}}} \|\pi_\theta) - \lambda_\Delta \|\Delta w_\theta - \Delta w_{\theta_{\text{old}}}\|_2^2 \right].$$

Under a line search that enforces $\mathbb{E}_{d^\pi} [\text{KL}(\pi \|\pi')] \leq \delta$ and $\mathbb{E} \|\Delta w_{\pi'} - \Delta w_\pi\|_2^2 \leq \eta$, Theorem 2 gives

$$J(\pi') \geq J(\pi) + \frac{1}{1-\gamma} \mathbb{E}_{d^\pi, \pi'} [A^\pi] - \frac{2\gamma}{(1-\gamma)^2} \varepsilon_{\max} \delta.$$

Replacing A^π by $\hat{A}^\pi$ introduces an additive error at most $\varepsilon/(1-\gamma)$. The trade penalty controls the change of Δw , and the Lipschitz property of costs implies an additional value drift bounded by $c_2 \eta$ for some $c_2 > 0$ depending on the Lipschitz moduli of C_z (Assumption 1). Combining terms yields the claim:

$$J(\pi') - J(\pi) \gtrsim \frac{1}{1-\gamma} \mathbb{E}_{d^\pi, \pi'} [\hat{A}^\pi] - \frac{2\gamma}{(1-\gamma)^2} \varepsilon_{\max} \delta - \frac{\varepsilon}{1-\gamma} - c_2 \eta.$$

□

A.7 Proof of Proposition 1 (turnover bound)

Proof of Proposition 1. From (1) and $C_z(\Delta w) \geq \kappa_1(z) \|\Delta w\|_1$, dropping the nonnegative risk penalty,

$$r_t^{\text{net}} \leq \bar{r} - \lambda_{\text{tc}} \underline{\kappa} \|\Delta w_t\|_1, \quad \underline{\kappa} := \min_z \kappa_1(z) > 0.$$

Taking expectations and summing with discount,

$$J(\pi) = \sum_{t \geq 0} \gamma^t \mathbb{E}[r_t^{\text{net}}] \leq \frac{\bar{r}}{1-\gamma} - \lambda_{\text{tc}} \underline{\kappa} \sum_{t \geq 0} \gamma^t \mathbb{E} \|\Delta w_t\|_1.$$

Hence

$$\sum_{t \geq 0} \gamma^t \mathbb{E} \|\Delta w_t\|_1 \leq \frac{\bar{r}}{\lambda_{\text{tc}} \underline{\kappa}} \cdot \frac{1}{1-\gamma}.$$

By the Abelian limit theorem (Hardy–Littlewood) and Assumption 1(v), the Cesàro average $\text{TO}(\pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t < T} \mathbb{E} \|\Delta w_t\|_1$ exists and

$$\text{TO}(\pi) = \lim_{\gamma \uparrow 1} (1-\gamma) \sum_{t \geq 0} \gamma^t \mathbb{E} \|\Delta w_t\|_1 \leq \frac{\bar{r}}{\lambda_{\text{tc}} \underline{\kappa}}.$$

□

A.8 Proof of Proposition 2 (inaction band)

Proof of Proposition 2. Consider the 1D myopic improvement of the Q -function at state (s, z) around the pre-trade weight w_{t-1} . Define $\Delta := a - w_{t-1}$. Let $g(\Delta) := Q^\pi(s, z, w_{t-1} + \Delta) - Q^\pi(s, z, w_{t-1})$ and suppose g is twice differentiable with $g(0) = 0$, $g'(0) = \theta$, and $g''(\Delta) \leq H$ for all Δ (local curvature upper bound). The one-step objective to maximize is

$$q(\Delta) = g(\Delta) - \kappa_1 |\Delta| - \frac{1}{2} \kappa_2 \Delta^2.$$

By Taylor with remainder and the curvature bound, $g(\Delta) \leq \theta \Delta + \frac{H}{2} \Delta^2$ for all Δ . Thus for $\tilde{\kappa} := \kappa_2 - H \geq 0$,

$$q(\Delta) \leq \theta \Delta - \kappa_1 |\Delta| - \frac{1}{2} \tilde{\kappa} \Delta^2 =: \varphi(\Delta).$$

We claim that $\Delta^* = 0$ maximizes φ whenever $|\theta| \leq \kappa_1$. Indeed, for any $\Delta \neq 0$,

$$\varphi(\Delta) \leq |\theta| |\Delta| - \kappa_1 |\Delta| - \frac{1}{2} \tilde{\kappa} \Delta^2 \leq -(\kappa_1 - |\theta|) |\Delta| < \varphi(0) = 0.$$

Therefore $\Delta^* = 0$ maximizes φ and, since $q \leq \varphi$ with equality at $\Delta = 0$, also maximizes q . In particular, if Q^π is locally strongly concave with curvature parameter H around w_{t-1} , the “dead-zone” condition $|\theta| \leq \kappa_1$ translates (by the mean-value theorem) to $|w^*(s, z) - w_{t-1}| \leq \tau$ with $\tau \asymp \kappa_1/(\kappa_2 + H)$, which gives the announced inaction band. □

A.9 Proof of Theorem 3 (value of regime conditioning)

Proof of Theorem 3. Let $a_z^*(s) \in \arg \max_{a \in \mathcal{W}} Q^{\pi^*}(s, z, a)$ be regime-specific near-optimal actions, and suppose Assumption 3 holds with separation $\Delta > 0$ on a set $E \subset \mathcal{S}$ of positive d^π -measure for each z . Consider any unconditioned policy $\pi_u(a | s)$ and write its conditional mean action as $\bar{a}_u(s) := \mathbb{E}_{\pi_u}[a | s]$. For each z , Jensen and the curvature bound as in the previous proof imply (using Q^{π^*} twice differentiable in a and upper curvature H)

$$\mathbb{E}_{\pi_u} [Q^{\pi^*}(s, z, a)] \leq Q^{\pi^*}(s, z, \bar{a}_u(s)) \leq Q^{\pi^*}(s, z, a_z^*(s)) - \frac{\tilde{\kappa}}{2} \|\bar{a}_u(s) - a_z^*(s)\|_2^2,$$

where $\tilde{\kappa} := \kappa_2 - H > 0$ (choose κ_2 large enough; recall $\Gamma_z \geq 0$ contributes to strong penalization in (4)). Thus the per-state per-regime suboptimality is lower bounded by a quadratic in the action mismatch. Since π_u is unconditioned, $\bar{a}_u(s)$ is common across regimes, hence for any (s, z_1, z_2) ,

$$\|\bar{a}_u(s) - a_{z_1}^*(s)\|_2^2 + \|\bar{a}_u(s) - a_{z_2}^*(s)\|_2^2 \geq \frac{1}{2} \|a_{z_1}^*(s) - a_{z_2}^*(s)\|_2^2 \geq \frac{1}{2} \Delta^2,$$

by the parallelogram identity. Averaging over regimes with equal weights and over $s \in E$, the average one-step regret of any π_u is at least $\frac{\tilde{\kappa}}{4} \Delta^2$ on E . Discounting over time and using the occupancy measure, we obtain

$$J(\pi^*) - J(\pi_u) \geq \frac{1}{1-\gamma} \cdot \frac{\tilde{\kappa}}{4} p_{\min} \Delta^2,$$

where $p_{\min} := \min_z \Pr(z)$ and we used that the balanced objective equally weights regimes. Since Π_{cond} can represent $\{a_z^*\}$ within uniform error ϵ , the same argument gives at most $O(\epsilon^2)$ regret for a conditioned policy, proving the advantage gap stated in Theorem 3 (with an explicit $c = \frac{\tilde{\kappa}}{4} p_{\min}$). $\square$

A.10 Proof of Theorem 4 (robustness to cost misspecification)

Proof of Theorem 4. Let $\hat{C}_z$ be the proxy cost and define $\delta := \sup_{z,u} |\lambda_{\text{tc}} C_z(u) - \lambda_{\text{tc}} \hat{C}_z(u)|$. Then per step the reward perturbation satisfies

$$|r_t^{\text{net}}(C) - r_t^{\text{net}}(\hat{C})| \leq \delta.$$

For any π ,

$$|J_C(\pi) - J_{\hat{C}}(\pi)| = \left| \sum_{t \geq 0} \gamma^t \mathbb{E}[r_t^{\text{net}}(C) - r_t^{\text{net}}(\hat{C})] \right| \leq \frac{\delta}{1-\gamma}.$$

Thus $J_C(\hat{\pi}) \geq J_{\hat{C}}(\hat{\pi}) - \delta/(1-\gamma)$ for any $\hat{\pi}$. In particular, letting $\hat{\pi}$ be a maximizer of $J_{\hat{C}}$ and $\pi_{\hat{C}}^*$ that of J_C ,

$$J_C(\pi_{\hat{C}}^*) - J_C(\hat{\pi}) \leq (J_{\hat{C}}(\pi_{\hat{C}}^*) + \frac{\delta}{1-\gamma}) - (J_{\hat{C}}(\hat{\pi}) - \frac{\delta}{1-\gamma}) = \frac{2\delta}{1-\gamma} + (J_{\hat{C}}(\pi_{\hat{C}}^*) - J_{\hat{C}}(\hat{\pi})).$$

This is the desired bound. $\square$

A.11 Proof of Proposition 3 (CVaR surrogate)

Proof of Proposition 3. Fix a batch of losses $\{L^{(i)}\}_{i=1}^N$. The Rockafellar–Uryasev surrogate (5) is convex in η and differentiable almost everywhere, with subgradient

$$\partial_\eta \left[\eta + \frac{1}{(1-\alpha)N} \sum_{i=1}^N (L^{(i)} - \eta)_+ \right] = 1 - \frac{1}{(1-\alpha)N} \sum_{i=1}^N \mathbf{1}\{L^{(i)} > \eta\},$$

which is monotone in η , hence a unique minimizer exists. For fixed η , the policy objective is a smooth function of θ (Assumption 2); using a step size chosen by Armijo backtracking ensures descent and bounded iterates. Standard two-block alternating minimization arguments then imply that every limit point (θ^*, η^*) is a first-order stationary point of the joint problem (see, e.g., [51, Prop. 2.7.1]). $\square$

References

- [1] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
- [2] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–39, 2001.

- [3] Albert S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.
- [4] Yakov Amihud. Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56, 2002.
- [5] Luboš Pástor and Robert F. Stambaugh. Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685, 2003.
- [6] Joel Hasbrouck. Trading costs and returns for U.S. equities: Estimating effective costs from daily data. *The Journal of Finance*, 64(3):1445–1477, 2009.
- [7] Ruslan Y. Goyenko, Craig W. Holden, and Charles A. Trzcinka. Do liquidity measures measure liquidity? *Journal of Financial Economics*, 92(2):153–181, 2009.
- [8] Kingsley Y. L. Fong, Craig W. Holden, and Charles A. Trzcinka. What are the best liquidity proxies for global research? *Review of Finance*, 21(4):1355–1401, 2017.
- [9] Anna A. Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32, 2013.
- [10] Nicolae Gârleanu and Lasse Heje Pedersen. Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, 68(6):2309–2340, 2013.
- [11] Robert Novy-Marx and Mihail Velikov. A taxonomy of anomalies and their trading costs. *The Review of Financial Studies*, 29(1):104–147, 2016.
- [12] Alan Moreira and Tyler Muir. Volatility-managed portfolios. *The Journal of Finance*, 72(4):1611–1644, 2017.
- [13] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.
- [14] Francisco Barillas and Jay Shanken. Comparing asset pricing models. *The Journal of Finance*, 73(2):715–754, 2018.
- [15] Francisco Barillas, Raymond Kan, Cesare Robotti, and Jay Shanken. Model comparison with Sharpe ratios. *Journal of Financial and Quantitative Analysis*, 55(6):1840–1874, 2020.
- [16] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.
- [17] Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.
- [18] Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross section. *Journal of Financial Economics*, 135(2):271–292, 2020.
- [19] Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.
- [20] Ilias Filippou, Thomas A. Maurer, Luca Pezzo, and Mark P. Taylor. Importance of transaction costs for asset allocation in foreign exchange markets. *Journal of Financial Economics*, 159:103886, 2024.
- [21] Gabor Pinter, Chaojun Wang, and Junyuan Zou. Size discount and size penalty: Trading costs in bond markets. *The Review of Financial Studies*, 37(7):2156–2190, 2024.
- [22] Andrew Ang and Geert Bekaert. International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4):1137–1187, 2002.
- [23] Michael W. Brandt, Pedro Santa-Clara, and Rossen Valkanov. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447, 2009.
- [24] Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.
- [25] Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005.
- [26] Andrew W. Lo. The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.
- [27] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML)*, 2002.
- [28] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

- [29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.
- [30] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- [31] Victor DeMiguel, Alberto Martín-Utrera, Francisco J. Nogales, and Raman Uppal. A transaction-cost perspective on the multitude of firm characteristics. *The Review of Financial Studies*, 33(5):2180–2222, 2020.
- [32] Y. Bai et al. A review of reinforcement learning in financial applications. *Annual Review of Statistics and Its Application*, 12:209–232, 2025.
- [33] Darwin Choi, Zijun Jiang, and Flora Zhang. Machine learning and international stock returns. *Review of Asset Pricing Studies*, 2025. (Advance article).
- [34] Olivier Lledo and Michael Wolf. Markowitz portfolios under transaction costs. *The Quarterly Review of Economics and Finance*, 100:101962, 2025.
- [35] Pierre Collin-Dufresne, Kent Daniel, and Mehmet Sağlam. Liquidity regimes and optimal dynamic asset allocation. *Journal of Financial Economics*, 136(2):379–406, 2020.
- [36] Álvaro Cartea, Sebastian Jaimungal, and José Penalva. *Algorithmic and High-Frequency Trading*. Cambridge University Press, 2015.
- [37] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.
- [38] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- [39] Ido Greenberg, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Efficient risk-averse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:14846–14859, 2022.
- [40] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.
- [41] R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.
- [42] Carlo Acerbi and Dirk Tasche. Expected shortfall: A natural coherent alternative to value at risk. *Economic Notes*, 31(2):379–388, 2002.
- [43] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1 -ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 272–279, 2008.
- [44] Weiran Wang and Miguel A. Carreira-Perpiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv:1309.1541*, 2013.
- [45] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.
- [46] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarsz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.
- [47] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.
- [48] Matteo Pirodda, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 307–315, 2013.
- [49] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Springer, 3rd edition, 2006.
- [50] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.
- [51] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.
- [52] Y. Bai and coauthors. A review of reinforcement learning in financial applications. *Annual Review of Statistics and Its Application*, 12:209–232, 2025.
- [53] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. *arXiv preprint arXiv:1502.01619*, 2015.

- [54] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.
- [55] P. R. Hansen, A. Lunde, and J. M. Nason. Model confidence sets for forecast comparison. *Oxford Bulletin of Economics and Statistics*, 67(s1): 839–861, 2005.