KNOWING WHEN NOT TO ACT: LATENT NO-ACTION REGION RECOVERY HIDDEN IN NEURAL CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear action costs induce structured inaction regions in which the optimal control is exactly zero. Neural policies can nevertheless achieve near-saturated objective values while blurring this regime structure into near-zero actions, leaving users unable to see where, or why, the controller should wait. Post-hoc magnitude thresholding can enforce zeros, but its induced geometry depends on an arbitrary cutoff, need not match the true switching region, and may degrade performance by suppressing beneficial interventions. We instead recover the latent action/inaction regime directly from a frozen neural controller, without learning a global value function, explicitly parameterizing a free boundary, or thresholding action magnitudes. Local continuation gradients yield marginal-value ratios whose known action-cost determines the regime labels, so each decision comes with an economic reason and a robustness margin; a Yosida map then realizes the recovered hold regime as exact zero while preserving the switching structure. Fixed-time control refinement re-evaluates this action at virtual states to propose a boundary-directed net action, executed once under the original cost model. Across linear-action-cost benchmarks, including high-dimensional dynamic portfolio choice, we show that structurally correct inaction cannot generally be inferred from raw action magnitude alone, and that recovering the correct switching regime can improve both interpretability and control performance.

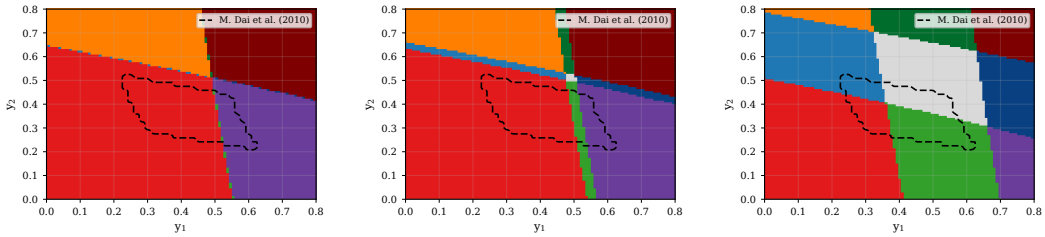
1 INTRODUCTION

Linear action costs create distinct intervention and inaction regimes: inside a no-action region the optimal control is exactly zero. Proportional transaction-cost portfolio choice is a canonical example: an investor waits inside a no-trade region and intervenes only at its boundary Magill & Constantinides (1976); Constantinides (1986); Davis & Norman (1990); Dumas & Luciano (1991); Shreve & Soner (1994). The resulting finite-variation control and free-boundary geometry are well characterized in low dimension, but classical solvers scale poorly with the number of assets Liu & Loewenstein (2002); Dai & Yi (2009); Belak & Sass (2019); Liu (2004); Muthuraman & Kumar (2006); Chen & Dai (2013).

Neural policy optimization scales more naturally, yet high objective need not identify the correct inactive set. A small nonzero action can represent either beneficial nudge or involuntary leakage from a true hold region. This is also an interpretability problem: a user deploying a trading, thrust, or release controller needs to know *where* it will deliberately do nothing and *why*, and a policy that acts a little everywhere answers neither question. Magnitude thresholding enforces zeros, but different cutoffs produce different geometries (Figure 1) and may suppress useful actions.

We recover *where inaction is structurally warranted* from the neural controller itself and then deploy that information to improve control. We freeze a directly trained neural policy and differentiate its Monte Carlo continuations by feedback back-propagation through time (BPTT). The resulting marginal-value ratio compares one more dollar in an asset with one more dollar of cash; known transaction-cost turns it into a buy/hold/sell label, and a closed-form Yosida map realizes the hold label as an exact-zero *regularized control*, without fitting a global value function or free boundary (Stages I–III). Because this map returns a finite trading rate rather than the immediate displacement sought in singular control, Stage IV, *fixed-time control refinement*, repeats the query at virtual portfolios and executes one net trade in the original market. Recovery asks which action the regularized continuation

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.



(a) $\tau = 10^{-3}$ (b) $\tau = 10^{-2}$ (c) $\tau = 10^{-1}$

Figure 1: Magnitude-thresholded inactive sets (gray) for a directly optimized policy: coordinate i is inactive when $|u_i| < \tau$. Geometry varies with the arbitrary cutoff. Dashed black: held-out DP/QVI reference of M. Dai et al. (2010).

prescribes; refinement asks how far that action should move the portfolio when physical time is held fixed. Our contributions are:

- **Regime recovery from a frozen policy:** a continuation-gradient detector with a threshold-free, exact-zero realization governed by a switching margin (Propositions 1–2, Corollary B.7).
- **An interpretable decision map:** each label states whether a trade's marginal gain covers its cost, with a robustness margin; the maps expose how correlation reshapes joint inaction and audit trained policies against held-out DP labels (Section 4).
- **From regime to performance:** one-shot improvement of J_η, a conditional oracle rate separation (Theorems 3–4), and higher held-out mean utility than the regularized neural policy for $N \leq 50$, with transfer beyond finance.

Relation to neural control methods. Neural PDE solvers and PINNs approximate value equations, while deep hedging and simulation-based control optimize decisions through sampled trajectories E et al. (2017); Han et al. (2018); Raissi et al. (2019); Buehler et al. (2019); Fahim & Tsai (2018); we use a trained feedback as a local continuation model whose componentwise realization rule is explicit, with state dependence and covariance supplied by the joint continuation (Appendix B.1).

2 The Model and Yosida Action Representation

Singular portfolio problem. Cash earns r; risky asset i has drift μ_i, volatility σ_i, Brownian correlation ρ_{ij}, and proportional sale cost $\alpha_i \in (0, 1)$. With cash X_t, nonnegative risky-dollar positions Y_t, and $\gamma > 1$, define liquidation wealth and CRRA utility by

$$L_t = X_t + \sum_i (1 - \alpha_i) Y_t^i, \qquad U(c) = \frac{c^{1-\gamma} - 1}{1 - \gamma}. \tag{1}$$

Nondecreasing purchases B^i and sales S^i generate

$$dX_t = r X_t \, dt - \sum_i dB_t^i + \sum_i (1 - \alpha_i) dS_t^i, \tag{2}$$

$$dY_t^i = \mu_i Y_t^i \, dt + \sigma_i Y_t^i \, dW_t^i + dB_t^i - dS_t^i, \tag{3}$$

where $d\langle W^i, W^j \rangle_t = \rho_{ij} \, dt$. The singular value is $V_0(t, z) = \sup \mathbb{E}_{t,z} U(L_T)$ over admissible finite-variation strategies (Appendix E.2).

Switching geometry and regularization. At smooth points with $V_{0,x} > 0$, set

$$R_i^0 = \frac{V_{0,y_i}}{V_{0,x}}, \qquad c_\alpha(u) = \sum_i (u_i^+ - (1 - \alpha_i) u_i^-), \qquad \mathcal{C}_\alpha = \prod_i [1 - \alpha_i, 1]. \tag{4}$$

Buying or selling one dollar has marginal gain $V_{0,x}(R_i^0 - 1)$ or $V_{0,x}(1 - \alpha_i - R_i^0)$, respectively, and the singular switching relation is

$$R^0 \in \partial c_\alpha(u) \quad \Longleftrightarrow \quad u \in N_{\mathcal{C}_\alpha}(R^0), \tag{5}$$

where $N_{\mathcal{C}_\alpha}$ is the Euclidean normal cone: an interior ratio forces hold, while the upper and lower faces permit buy and sell directions (Stieltjes complementarity: Appendix B.2). In intervention regions the exact singular ratio is pinned to a face, so it identifies the active direction but not the required displacement. Regularization instead encodes intensity through violations of that face; our fixed-time refinement uses the resulting finite rate to construct an immediate candidate displacement, without claiming an exact inverse of the regularization. With $A_t = L_t u_t$, let

$$dX_t = \left[rX_t - \sum_i (A_t^i)^+ + \sum_i (1-\alpha_i)(A_t^i)^- - \frac{\eta}{2L_t}\|A_t\|^2 \right] dt, \tag{6}$$

$$dY_t^i = \mu_i Y_t^i\, dt + \sigma_i Y_t^i\, dW_t^i + A_t^i\, dt, \qquad \eta > 0. \tag{7}$$

Write $J_\eta(v) = \mathbb{E}U(L_T^{\eta,v})$ for a fixed-policy value under these dynamics, $V_\eta = \sup_v J_\eta(v)$, and $J_0(v)$ for the original-model value. The normalized local Hamiltonian and its ratio are

$$h_\eta(u;R) = R^\top u - c_\alpha(u) - \frac{\eta}{2}\|u\|^2, \qquad R = V_{\eta,y}/V_{\eta,x}. \tag{8}$$

Unlike R^0, this ratio can leave the hold wedge, and its violation encodes a needed intensity.

Yosida action. The Yosida approximation of $N_{\mathcal{C}_\alpha}$ is

$$\mathcal{P}_\eta(R) = \eta^{-1}(R - \mathrm{Proj}_{\mathcal{C}_\alpha} R), \qquad [\mathcal{P}_\eta(R)]_i = \eta^{-1}\big([R_i - 1]_+ - [1 - \alpha_i - R_i]_+\big). \tag{9}$$

Consequently,

$$\mathcal{P}_\eta(R) = 0 \quad \Longleftrightarrow \quad R \in \mathcal{C}_\alpha. \tag{10}$$

Proposition 1 (Exact-zero realization and regime robustness)**.** *For every R and $\eta > 0$,*

$$\arg\max_u h_\eta(u;R) = \mathcal{P}_\eta(R), \qquad \max_u h_\eta(u;R) = \frac{\mathrm{dist}^2(R,\mathcal{C}_\alpha)}{2\eta}. \tag{11}$$

Two ratios $R, \widetilde{R}$ have the same coordinate-i buy/hold/sell label whenever

$$|\widetilde{R}_i - R_i| < \mathrm{dist}(R_i, \{1-\alpha_i, 1\}). \tag{12}$$

In particular, a strict-hold label yields exact zero for both ratios.

Proof. See Appendix B.3 and Appendix B.8. □

The margin condition separates regime recovery from action regression: inside a strict hold region, estimation error moves the ratio without changing the exact-zero action, whereas near a switching face it can change the label; by (54), any small nonzero action at a strict-hold score is strictly worse to first order, so near-zero leakage differs structurally from inaction. $\mathcal{P}_\eta$ maximizes the *η-regularized local Hamiltonian*; it is neither a magnitude threshold nor a direct solver for J_0. Given the true regularized ratio it recovers the regularized control, and given an estimated frozen-continuation ratio it is a local regularized action update. It preserves the ratio-space wedge for every η, but the state-space boundary depends on the supplied continuation, and wedge preservation alone gives no global free-boundary convergence (Appendix B.3, Appendix D.9).

Reading the recovered regime. The labels are statements about marginal values, not action sizes. R_i prices one more dollar held in asset i in units of cash; since a purchase costs one dollar and a sale returns $1-\alpha_i$, coordinate i is a buy if $R_i > 1$, a sell if $R_i < 1-\alpha_i$, and a hold otherwise, because no marginal trade then recovers its own cost. Each decision thus has a reason in the units of the cost model and a margin $m_i = \mathrm{dist}(R_i, \{1-\alpha_i, 1\})$ bounding the ratio error it tolerates (12); labels near a face are the borderline decisions worth scrutinizing. A thresholded policy provides neither, since its zeros depend on a cutoff in action units.

Regime accuracy and rate accuracy. The realization map also satisfies

$$\|\mathcal{P}_\eta(\widetilde{R}) - \mathcal{P}_\eta(R)\| \le \eta^{-1}\|\widetilde{R} - R\|. \tag{13}$$

A label can thus remain correct throughout a strict-margin neighborhood while the nonzero rate stays sensitive to ratio error. Smaller η amplifies this sensitivity, and retraining at another η changes the continuation itself, so small η alone ensures neither accurate actions nor a correct state-space boundary (Appendix F.3).

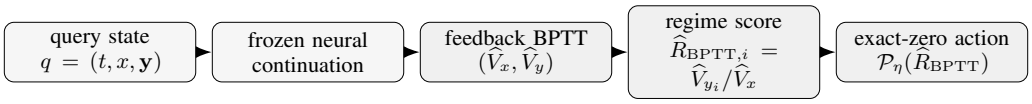


Figure 2: The neural policy is used as a differentiable continuation model, not as the final action rule. Local continuation sensitivities yield a regime score; transaction-cost α_i detect buy/hold/sell; the realization layer maps hold to exact zero.

3 METHODOLOGY

Stages I–II: train, freeze, and differentiate. In stage I, direct policy optimization (DPO) trains a feasible neural feedback u_θ on Monte Carlo terminal utility under the declared η-regularized simulator. Freeze the supplied reference $\bar{u}$ and retain its state-to-action derivatives in stage II; any transformation of the trained feedback must be included in this continuation definition (Appendix A.1). At a query $q = (t, x, \mathbf{y})$, differentiate M_{MC} continuation payoffs by feedback BPTT to obtain pathwise sensitivities $\widehat{\lambda}^{(j)}$. Use the ratio of mean derivatives:

$$\widehat{\lambda} = \frac{1}{M_{\mathrm{MC}}} \sum_{j=1}^{M_{\mathrm{MC}}} \widehat{\lambda}^{(j)}, \qquad \widehat{R}_{\mathrm{BPTT},i} = \widehat{\lambda}_{y_i}/\widehat{\lambda}_x. \tag{14}$$

For the untransformed reference $\bar{u} = u_{\theta^\star}$, the target is:

Proposition 2 (Frozen-continuation gradient identity)**.** *Under the differentiability and domination conditions of Assumption B.3, let $\Lambda_K^{\mathrm{pw},\Delta} = D_z G_\Delta(Z_K^\Delta)$. Then*

$$\Lambda_k^{\mathrm{pw},\Delta} = \left[D_z F_k^\Delta + D_u F_k^\Delta D_z u_{\theta^\star,k}(Z_k^\Delta)\right]^\top \Lambda_{k+1}^{\mathrm{pw},\Delta}, \qquad \mathbb{E}[\Lambda_k^{\mathrm{pw},\Delta} \mid Z_k^\Delta = z] = \nabla V_{\eta,k}^{\theta^\star,\Delta}(z). \tag{15}$$

Thus feedback BPTT targets the gradient of the declared discrete continuation value generated by the frozen policy $u_{\theta^\star}$.

Proof. See Appendix B.5. □

Freezing fixes the network's parameters, not its response to a perturbed state: the factor $D_u F_k^\Delta D_z u_{\theta^\star,k}$ retains that response along future continuations, whereas holding future actions fixed would differentiate a different, open-loop quantity. The identity concerns the declared simulator, including its domain rule and terminal payoff; automatic differentiation does not remove their modeling effects.

What continuation accuracy requires. Where the ratios exist, write $R^{\bar{u},\Delta}$ and $R^{\bar{u}}$ for the frozen discrete and continuous continuation ratios. Relative to the optimal regularized ratio R_η, their pointwise errors separate as

$$\|\widehat{R}_{\mathrm{BPTT}}^\Delta - R_\eta\| \le \underbrace{\|\widehat{R}_{\mathrm{BPTT}}^\Delta - R^{\bar{u},\Delta}\|}_{\text{acquisition}} + \underbrace{\|R^{\bar{u},\Delta} - R^{\bar{u}}\|}_{\text{simulator bias}} + \underbrace{\|R^{\bar{u}} - R_\eta\|}_{\text{reference bias}}. \tag{16}$$

More continuation paths address only the acquisition term, not discretization, payoff, or domain-treatment bias or a suboptimal reference. Theorem B.6 links Stage-I utility to ratio accuracy under separate quadratic-growth and ratio-stability conditions, so a nearly saturated training objective alone does not certify an accurate switching boundary; the ablations in Appendix F.1 vary training effort and sampling budget and are not estimates of a global policy gap.

Stage III: recover the regularized control. On a successful query with positive cash marginal, the recovered action is

$$u(q) = \mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}}(q)). \tag{17}$$

This is already an action, not just a label; it maximizes the local η-regularized Hamiltonian and so targets J_η, not the instantaneous net trade sought under J_0; Appendix F.7 provides an empirical J_η perturbation diagnostic, which cannot establish stationarity or global optimality. Failed queries require an explicitly supplied admissible fallback (Appendix A). Structural recovery only requires estimation error to remain below the switching margin: Corollary B.7 gives a fixed-query median-of-means guarantee with $O(\bar{m}^{-2}\log(N/\delta))$ rollouts for a positive bias-adjusted margin $\bar{m}$ when marginal and variance constants are fixed.

Amortized ratio distillation. To avoid recomputing BPTT continuations at every query, a feed-forward student $R_\theta(q)$ regresses teacher R_{BPTT} on trajectories disjoint from held-out evaluation paths and, at deployment, replaces online ratio estimation while retaining the same Yosida map (Appendix A.2).

Theorem 3 (One-shot policy improvement)**.** *Let $V_\eta^{\theta^\star} = J_\eta(u_{\theta^\star})$ and $R_{\theta^\star} = V_{\eta,y}^{\theta^\star}/V_{\eta,x}^{\theta^\star}$. Under the regularity, admissibility, positive cash-marginal, and complete-feedback conditions in Appendix B.6,*

$$J_\eta(\mathcal{P}_\eta(R_{\theta^\star})) \geq J_\eta(u_{\theta^\star}). \tag{18}$$

For a complete admissible estimated-ratio feedback, it suffices that

$$\|\widehat{R}_{\mathrm{BPTT}} - R_{\theta^\star}\| \leq \eta \|u_{\theta^\star} - \mathcal{P}_\eta(R_{\theta^\star})\| \tag{19}$$

throughout its reachable state–time region to obtain

$$J_\eta(\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})) \geq J_\eta(u_{\theta^\star}). \tag{20}$$

Proof. See Appendix B.6. □

This fixed-η comparison concerns J_η, not the original-model objective J_0, and requires the stated error bound; it neither identifies discrete BPTT gradients with optimal continuous-time gradients nor implies unconditional improvement in J_0.

Stage IV: fixed-time control refinement. Singular control may call for an immediate trade toward the no-trade boundary rather than the finite rate $\mathcal{P}_\eta(\widehat{R})$. We refine the recovered control into a candidate net displacement by repeating Stages II–III at *fixed physical time* t_k (Figure 9). Starting from $q_k = (t_k, z_k)$ and $\widetilde{z}_k^{(0)} = z_k$, re-query $\widehat{R}_k^{(m)} := \widehat{R}(t_k, \widetilde{z}_k^{(m)})$ and update

$$a_k^{(m)} = \Pi_m\Big(\delta_{\mathrm{FR}} L(\widetilde{z}_k^{(m)}) \mathcal{P}_\eta(\widehat{R}_k^{(m)})\Big), \qquad \widetilde{z}_k^{(m+1)} = \Gamma_0(\widetilde{z}_k^{(m)}, a_k^{(m)}), \tag{21}$$

where $\Gamma_0((x,y),a) = (x - \mathbf{1}^\top a, y + a)$ is a cost-free virtual transfer and Π_m enforces virtual and net-order feasibility. For fixed depth d or adaptive stopping ($\star$), the refined controls are the *cumulative orders*

$$\begin{aligned} \mathcal{P}_\eta^{[d]}(\widehat{R})(q_k) &= \sum_{m < M_k^{[d]}} a_k^{(m)}, \qquad M_k^{[d]} \leq d, \\ \mathcal{P}_\eta^{[\star]}(\widehat{R})(q_k) &= \sum_{m < M_k^{[\star]}} a_k^{(m)}, \quad M_k^{[\star]} \leq M_{\max}. \end{aligned} \tag{22}$$

Writing A_k for the selected output, only one net order is executed:

$$z_k^+ = \Gamma_\alpha(z_k, A_k) = \big(x_k - c_\alpha(A_k),\, y_k + A_k\big). \tag{23}$$

Virtual updates neither advance market time nor incur transaction costs; only the final order is charged. Hold, stall, query failure, or a depth cap can stop refinement before boundary entry, and the fee-adjusted z_k^+ need not lie in the estimated hold region even if the virtual endpoint does (Appendix A.4, Appendix A.5).

Guarantees for refinement. Theorem C.3 bounds the smallest ratio-wedge residual along the *virtual* path of the learned continuation, and Theorem C.4 shows that J_0 improvement must also cover the executed fee, the market-step residual, and the baseline correction. Virtual progress, or entering an estimated hold region, therefore certifies neither improved J_0, optimal accumulated trades, nor convergence to the singular optimum. To isolate what accumulated execution can buy, set $R_\eta = V_{\eta,y}/V_{\eta,x}$ and let $\pi_\eta^\star$ be an assumed admissible exact-ratio strategy whose charged net trade and continuation satisfy certificate R4 (Appendix D.9); finite Stage IV does not construct it. Define

$$e_{\mathrm{reg}} := V_0 - V_\eta, \qquad e_{\mathrm{one}} := V_0 - J_0(\mathcal{P}_\eta(R_\eta)), \qquad e_\star := V_0 - J_0(\pi_\eta^\star). \tag{24}$$

Theorem 4 (Conditional oracle refinement advantage)**.** *Under R1–R3 in Appendix D.9, fix $N < \infty$ and an interior initial state (t,z) with portfolio ratio $\theta = ((1-\alpha_i)y_i/L)_{i=1}^N \in \operatorname{int}\mathcal{K}_\theta$ and $Q_0(t,z) > 0$, where $Q_0 = -(V_{0,t} + \mathcal{L}^0 V_0)$. For sufficiently small η,*

$$D_\eta := J_0(\mathcal{P}_\eta(R_\eta)) - V_\eta > 0, \qquad e_{\mathrm{reg}} = e_{\mathrm{one}} + D_\eta, \qquad e_{\mathrm{reg}} > e_{\mathrm{one}} \geq 0. \tag{25}$$

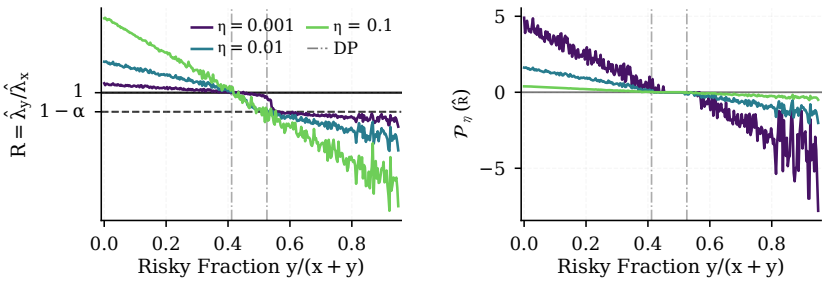


Figure 3: Single-asset marginal-value ratio and induced Yosida action. The interval $[1-\alpha, 1]$ is the exact ratio-space hold wedge. Whenever the recovered ratio lies inside this interval, the projected rate is identically zero; outside it, the sign of the rate determines the buy or sell regime.

If the certified refinement condition R4 holds with residual $\kappa_\eta = O_N(\eta^{2/3})$, then eventually

$$e_{\mathrm{reg}} > e_{\mathrm{one}} > e_\star \geq 0, \qquad e_\star = O_N(\eta^{2/3}). \tag{26}$$

The detailed square-root comparison for e_{reg} and e_{one}, the constants, the residual definition, and the proof are in Appendix D.9.9.

At this strict intervention state both e_{reg} and e_{one} are $\Theta_N(\sqrt{\eta})$: removing leakage improves the gap but not its order, whereas certified accumulated execution attains the faster upper order $O_N(\eta^{2/3})$. This isolates the value of certified accumulated execution from estimator, simulator, and finite-step errors; it is not a rate for learned fixed-time refinement, and constants may depend on dimension and the initial state.

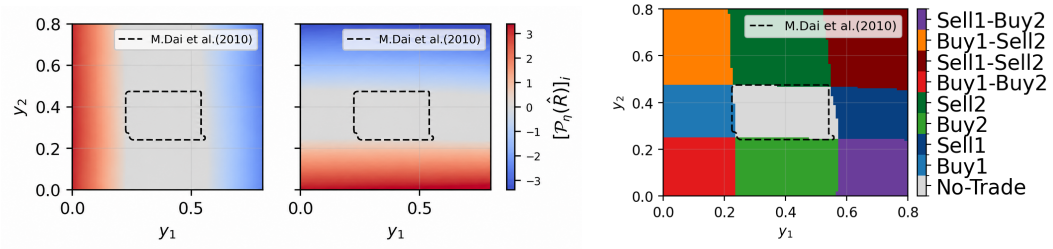
4 Experiments

We evaluate recovered switching geometry and the original objective J_0; DP/QVI boundaries are held out from training and continuation estimation, and further diagnostics are in Appendix F.

One asset: exact hold recovery. Figure 3 shows the R_{BPTT} and induced Yosida action $\mathcal{P}_\eta$. The recovered ratio enters $[1-\alpha, 1]$ over a flat zero-action interval aligned with the held-out DP boundary; outside this interval, the action sign identifies buy or sell. This directly visualizes Proposition 1: a smooth learned continuation can produce an exactly inactive action interval. The exact-zero property follows from the ratio-space map, whereas agreement with the held-out DP boundary separately assesses the accuracy of the recovered state-space region. Boundary-discrepancy grids over α, η, warm start, train seeds, and M_{MC} are in Appendix F.

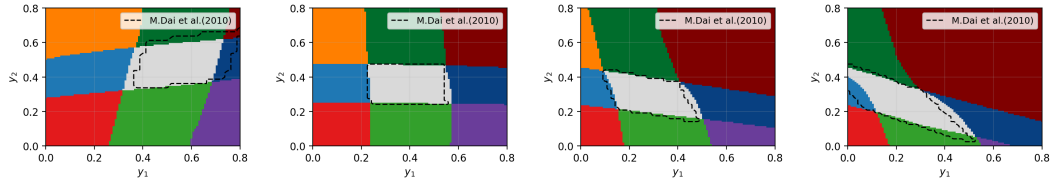
Multiple assets: joint switching geometry. Figure 4 shows $\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})$ for two assets $i = 1, 2$ and induced buy/hold/sell regime at $\rho = 0$, with the central hold region aligned with the DP/QVI reference. Varying correlation rotates and slants that region, and the recovered hold region follows the corresponding tilt of the held-out reference (Figure 5): positive correlation slants it into a band along which one risky position substitutes for the other, and negative correlation tilts it the other way. The Yosida map is coordinatewise and unchanged across correlations, so this joint geometry does not come from the realization layer: feedback BPTT differentiates the joint continuation, so each marginal-value ratio already carries the full joint state and the cross-asset covariance before the componentwise map is applied, so the partition does not arise from independent scalar bands: covariance information survives continuation estimation and appears in the recovered geometry. Figure 6 extends the recovered partition to the full $3^3 = 27$ buy/hold/sell combinations for three assets at zero pairwise correlation (full correlation sweep: Appendix F).

Reading the recovered maps. These partitions answer questions that raw or thresholded actions cannot: *where to wait* (the gray sets, where no marginal trade pays for its cost under the learned continuation, with no cutoff to tune; cf. Figure 1), *which trades co-occur* (Figure 6 separates states that rebalance one asset while holding the others from those requiring joint rebalancing), and *how far to trust a label* (the margin m_i flags coordinates near a switching face).



(a) Projected continuous controls. (b) Induced buy/hold/sell regions.

Figure 4: Two-asset projected controls and induced regimes at $\rho = 0$. Dashed lines: held-out DP/QVI no-trade reference.



(a) $\rho = -0.3$ (b) $\rho = 0$ (c) $\rho = 0.3$ (d) $\rho = 0.6$

Figure 5: Correlation changes the recovered partition through joint continuation marginals; the coordinatewise Yosida map is unchanged.

Original objective and selective intervention. Figure 7 compares held-out original-model J_0 and executed trading activity for $N \in \{5, 10, 20, 50\}$. Relative to the nearly always active $u_{\theta,\eta}$, the adaptive BPTT refinement controller intervenes during 52–74% of the horizon and attains higher sample-mean J_0 in each dimension. Less activity is not always sufficient: the adaptive PINN refinement controller trades less frequently but has lower mean utility. The result illustrates selective intervention rather than an optimal trading-duration rule or an isolated causal effect of regime recovery. Because this CRRA utility model has no tractable DP/QVI boundary reference at these dimensions, Figure 7 itself does not measure high-dimensional free-boundary accuracy; we test regime recovery directly below. Baselines, uncertainty conventions, and execution diagnostics are in Appendix F.15.

High-dimensional regime recovery against a DP oracle. To measure regime recovery itself at $N = 50$, we use a CARA market with independent returns whose Bellman criterion factorizes across coordinates, giving high-resolution held-out DP labels (Appendix F.14). Table 1 also audits trained policies: both direct policies trade on every coordinate (hold recall 0.0000), magnitude masks exchange action recall for hold recall as the cutoff varies, and $\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})$ recovers both. In battery storage, interior hold IoU against grid DP is 0.0093 for the direct policy and 0.8743 for the online detector (Table 11).

Transfer beyond finance. Stochastic minimum-fuel spacecraft rendezvous uses Hill–Clohessy–Wiltshire dynamics and symmetric ℓ_1 fuel cost (Appendix F.8); its exact no-thrust set $\|B^\top \lambda\|_\infty \le \alpha$ has the same reading: thrust only when the marginal value of a velocity change exceeds its fuel price. Figure 8 shows that deeper mapping front-loads intervention, reduces docking distance sooner, and lowers the final cost. The adaptive map spends more fuel initially and subsequently exhibits slower fuel accumulation. Lower final cost can therefore coexist with a stronger initial intervention; refinement affects the timing and extent of intervention as well as the terminal objective. The one-sided two-reservoir benchmark (Table 2; Appendix F.9) provides a complementary test: adaptive refinement approaches the numerical DP reference. Both are minimization problems.

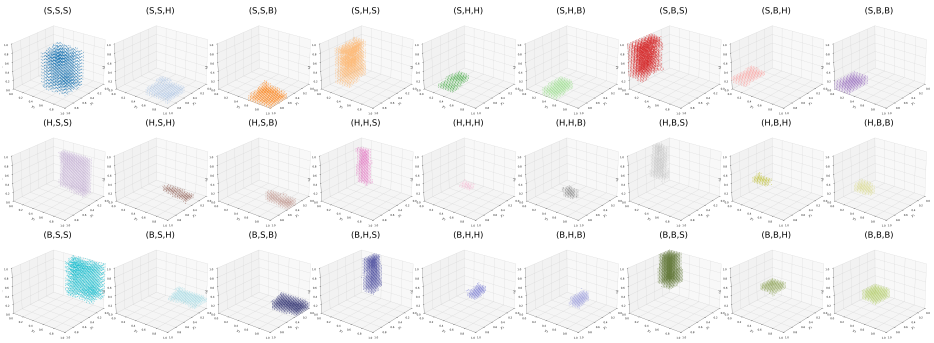


Figure 6: The induced $3^3 = 27$ buy/hold/sell regime partition for three assets at zero pairwise correlation. The corresponding actions and 3-asset correlation sweep are shown in Figure 11 in the Appendix.

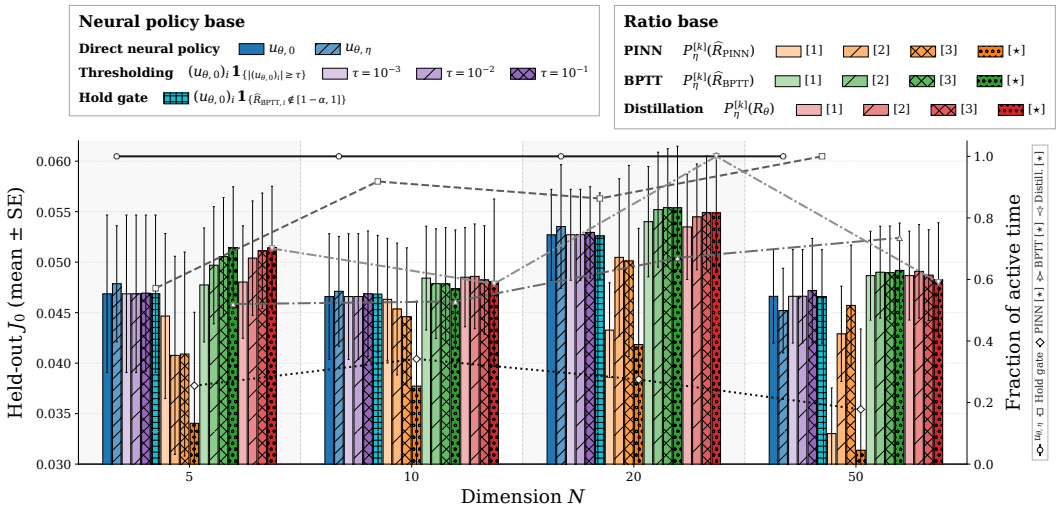


Figure 7: **Held-out objective and executed activity.** For $N \in \{5, 10, 20, 50\}$, 18 bars report J_0 means with outer-path standard errors (left axis); five curves report the fraction of the horizon with a nonzero executed net trade (right axis). $[\star]$ denotes adaptive refinement. Activity uncertainty is not displayed.

5 Discussion and Limitations

A neural policy can contain useful switching information although its actions blur the inactive set, and recovering it yields labels with an economic reason and a margin. Selective intervention can coexist with higher observed mean utility, whereas less trading alone does not ensure it, and deeper mapping need not help. The transfer examples cover symmetric linear fuel costs and one-sided singular releases, providing evidence across different intervention geometries.

The guarantees concern different objects (supplied ratios, the declared discrete continuation, J_η, virtual refinement paths, or exact ratios under R1–R4), and none establishes unconditional improvement or global free-boundary convergence. Monte Carlo error, discretization, domain treatment, and refinement caps remain material; the objective–depth curves measure aggregate performance rather than local refinement conditions and, for the coupled CRRA model with $N \geq 5$, are indirect evidence about geometry, while the separable $N = 50$ CARA oracle gives no Hausdorff, IoU, or free-boundary guarantee for the correlated model. Masking ablations are coordinatewise rather than exhaustive; R_θ has no independent guarantee; accuracy, latency, and termination depend on the ratio estimator; and empirical J_0 gains do not validate the oracle assumptions.

Table 1: **Direct high-dimensional regime recovery against the held-out separable CARA DP oracle at $N = 50$ over five training seeds.** All learned methods are global 50-output networks; the DP decomposition is used only to generate evaluation labels. The same frozen set of 1,000 states (50,000 coordinate labels) is used for every seed. Each metric is reported as mean $\pm$ sample standard deviation $[\min, \max]$ across seeds.

Group	Method	Regime acc.	Action recall	Hold recall
Raw policy	$u_{\theta,0}$	0.5136 ± 0.0072 $[0.5041, 0.5218]$	1.0000 ± 0.0000 $[1.0000, 1.0000]$	0.0000 ± 0.0000 $[0.0000, 0.0000]$
	$u_{\theta,\eta}$	0.5118 ± 0.0039 $[0.5057, 0.5144]$	1.0000 ± 0.0000 $[1.0000, 1.0000]$	0.0000 ± 0.0000 $[0.0000, 0.0000]$
Coord. mask	$\tau = 10^{-3}$ on $u_{\theta,0}$	0.6716 ± 0.0048 $[0.6639, 0.6765]$	0.8574 ± 0.0300 $[0.8176, 0.8908]$	0.4661 ± 0.0435 $[0.4223, 0.5268]$
	$\tau = 10^{-2}$ on $u_{\theta,0}$	0.4669 ± 0.0003 $[0.4665, 0.4673]$	0.0007 ± 0.0006 $[0.0000, 0.0015]$	1.0000 ± 0.0000 $[1.0000, 1.0000]$
	$\tau = 10^{-1}$ on $u_{\theta,0}$	0.4665 ± 0.0000 $[0.4665, 0.4665]$	0.0000 ± 0.0000 $[0.0000, 0.0000]$	1.0000 ± 0.0000 $[1.0000, 1.0000]$
BPTT ratio	Hold gate on $u_{\theta,0}$	0.8507 ± 0.0054 $[0.8436, 0.8575]$	0.8979 ± 0.0006 $[0.8971, 0.8984]$	0.8301 ± 0.0012 $[0.8288, 0.8312]$
	$\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})$	$\mathbf{0.8663 \pm 0.0007}$ $[\mathbf{0.8653}, \mathbf{0.8668}]$	$\mathbf{0.8979 \pm 0.0006}$ $[\mathbf{0.8971}, \mathbf{0.8984}]$	$\mathbf{0.8301 \pm 0.0012}$ $[\mathbf{0.8288}, \mathbf{0.8312}]$
	$\mathcal{P}_\eta(\widehat{R}_\theta)$	0.7651 ± 0.0092 $[0.7517, 0.7772]$	0.8322 ± 0.0127 $[0.8167, 0.8468]$	0.6883 ± 0.0102 $[0.6774, 0.6976]$
PINN ratio	$\mathcal{P}_\eta(\widehat{R}_{\mathrm{PINN}})$	0.4673 ± 0.0004 $[0.4667, 0.4678]$	0.0015 ± 0.0007 $[0.0005, 0.0025]$	1.0000 ± 0.0000 $[1.0000, 1.0000]$

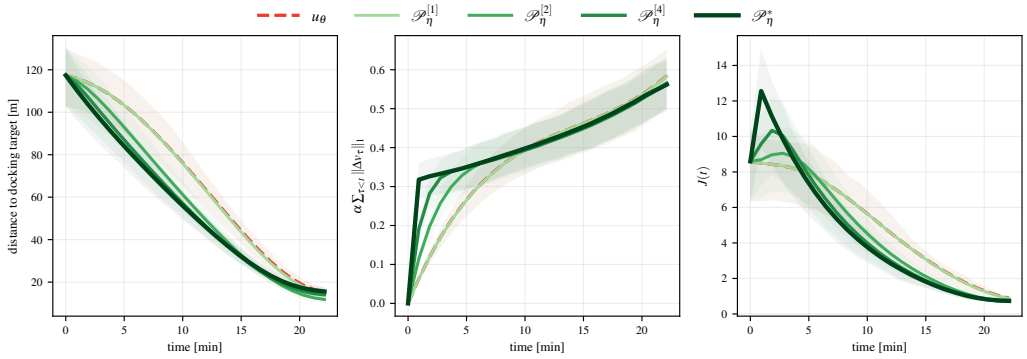


Figure 8: From left to right: distance to the docking target, cumulative fuel cost, and the objective trace. The warm policy u_θ is compared with finite fixed-time projection depths $\mathcal{P}_\eta^{[1]}$, $\mathcal{P}_\eta^{[2]}$, and $\mathcal{P}_\eta^{[4]}$, and with the adaptive $\mathcal{P}_\eta^*$.

Table 2: Two-reservoir results (mean $\pm$ standard error). Adaptive refinement approaches the same-problem grid DP reference. Details: Appendix F.9.

Method	$J_0 \downarrow$	$J_0 - J_0^{\mathrm{DP}} \downarrow$	Gap (%) $\downarrow$	Release cost $\downarrow$	Final distance $\downarrow$
DP $\eta = 0$	0.3915 ± 0.0050	0.0000 ± 0.0000	0.00	0.3553 ± 0.0043	0.1071 ± 0.0042
u_θ	0.5319 ± 0.0083	0.1404 ± 0.0036	35.87	0.3095 ± 0.0044	0.1297 ± 0.0037
$\mathcal{P}_\eta^{[1]}(\widehat{R}_{BPTT})$	0.5328 ± 0.0083	0.1413 ± 0.0036	36.10	0.3090 ± 0.0044	0.1300 ± 0.0036
$\mathcal{P}_\eta^{[2]}(\widehat{R}_{BPTT})$	0.4580 ± 0.0065	0.0665 ± 0.0018	16.99	0.3361 ± 0.0046	0.1062 ± 0.0038
$\mathcal{P}_\eta^{[4]}(\widehat{R}_{BPTT})$	0.4182 ± 0.0056	0.0267 ± 0.0009	6.82	0.3515 ± 0.0046	0.0966 ± 0.0038
$\mathcal{P}_\eta^*(\widehat{R}_{BPTT})$	0.3912 ± 0.0048	-0.0002 ± 0.0005	-0.06	0.3653 ± 0.0044	0.0905 ± 0.0042

Reproducibility Statement

Assumptions and proofs are given in Appendices B–E. Appendix A defines the frozen-continuation target, simulator and payoff conventions, and refinement algorithm. Appendix F reports numerical settings, comparison protocols, and supplementary results.

AI Use Statement

Generative AI tools were used to assist with language editing, document restructuring, LaTeX drafting, code drafting and editing, feedback on experimental design and methodology, checking mathematical claims and parts of theoretical derivations, translation, and interpretation and presentation of experimental results. All AI-assisted code, derivations, experiments, citations, and text were independently reviewed and verified by the authors. Generative AI was not used as a component of the controller, simulator, or evaluation protocol. The authors take responsibility for the final content of this work.

References

Albert Altarovici, Martin Reppen, and H. Mete Soner. Optimal consumption and investment with fixed and proportional transaction costs. *SIAM Journal on Control and Optimization*, 55(3): 1673–1710, 2017.

Pierluigi Balduzzi and Anthony W. Lynch. Transaction costs and predictability: Some utility cost calculations. *Journal of Financial Economics*, 52(1):47–78, 1999.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, Cham, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5.

Christoph Belak and Joern Sass. Finite-horizon optimal investment with transaction costs: Construction of the optimal strategies. *Finance and Stochastics*, 23(4):861–888, 2019.

Maxim Bichuch. Asymptotic analysis for optimal investment in finite time with transaction costs. *SIAM Journal on Financial Mathematics*, 3(1):433–458, 2012.

Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019. doi: 10.1080/14697688.2019.1571683.

Luciano Campi and Mark P. Owen. Multivariate utility maximization with proportional transaction costs. *Finance and Stochastics*, 15(3):461–499, 2011. doi: 10.1007/s00780-010-0125-9.

Jun Cao, Dan Harrold, Zhong Fan, Thomas Morstyn, David Healey, and Kang Li. Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model. *IEEE Transactions on Smart Grid*, 11(5):4513–4521, 2020. doi: 10.1109/TSG.2020.2986333.

René Carmona and Michael Ludkovski. Valuation of energy storage: An optimal switching approach. *Quantitative Finance*, 10(4):359–374, 2010. doi: 10.1080/14697680902946514.

Xinfu Chen and Min Dai. Characterization of optimal strategy for multiasset investment and consumption with transaction costs. *SIAM Journal on Financial Mathematics*, 4(1):857–883, 2013.

Jin Hyuk Choi, Mihai Sîrbu, and Gordan Žitković. Shadow prices and well-posedness in the problem of optimal investment and consumption with transaction costs. *SIAM Journal on Control and Optimization*, 51(6):4414–4449, 2013. doi: 10.1137/120881373.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, Philadelphia, 1990.

George M. Constantinides. Capital market equilibrium with transaction costs. *Journal of Political Economy*, 94(4):842–862, 1986.

Jakša Cvitanić and Ioannis Karatzas. Hedging and portfolio optimization under transaction costs: A martingale approach. *Mathematical Finance*, 6(2):133–165, 1996.

Jakša Cvitanić and Hui Wang. On optimal terminal wealth under transaction costs. *Journal of Mathematical Economics*, 35(2):223–231, 2001.

Min Dai and Fahuai Yi. Finite-horizon optimal investment with transaction costs: A parabolic double obstacle problem. *Journal of Differential Equations*, 246(4):1445–1469, 2009.

Min Dai, Lishang Jiang, Ping Li, and Fahuai Yi. Finite horizon optimal investment and consumption with transaction costs. *SIAM Journal on Control and Optimization*, 48(2):1134–1154, 2009.

Mark H. A. Davis and Andrew R. Norman. Portfolio selection with transaction costs. *Mathematics of Operations Research*, 15(4):676–713, 1990.

Bernard Dumas and Elisa Luciano. An exact solution to a dynamic portfolio choice problem under transactions costs. *The Journal of Finance*, 46(2):577–595, 1991.

Weinan E, Jiequn Han, and Arnulf Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

J. F. Eastham and K. J. Hastings. Optimal impulse control of portfolios. *Mathematics of Operations Research*, 13(4):588–605, 1988.

Arash Fahim and Wei-Yin Tsai. A numerical scheme for a singular control problem: Investment-consumption under proportional transaction costs. *Journal of Computational and Applied Mathematics*, 333:170–184, 2018.

Paolo Guasoni and Johannes Muhle-Karbe. Portfolio choice with transaction costs: A user's guide. In Vicky Henderson and Ronnie Sircar (eds.), *Paris-Princeton Lectures on Mathematical Finance 2013*, pp. 169–201. Springer, Cham, 2013. doi: 10.1007/978-3-319-00413-6_3.

Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

Ulrich G. Haussmann and Wulin Suo. Singular optimal stochastic controls II: Dynamic programming. *SIAM Journal on Control and Optimization*, 33(3):937–959, 1995a. doi: 10.1137/S0363012993250529.

Ulrich G. Haussmann and Wulin Suo. Singular optimal stochastic controls I: Existence. *SIAM Journal on Control and Optimization*, 33(3):916–936, 1995b. doi: 10.1137/S0363012993250256.

Karel Janeček and Steven E. Shreve. Asymptotic analysis for optimal investment and consumption with transaction costs. *Finance and Stochastics*, 8(2):181–206, 2004.

Jan Kallsen and Johannes Muhle-Karbe. On using shadow prices in portfolio optimization with transaction costs. *The Annals of Applied Probability*, 20(4):1341–1358, 2010.

Ralf Korn. Portfolio optimisation with strictly positive transaction costs and impulse control. *Finance and Stochastics*, 2(2):85–114, 1998.

Hong Liu. Optimal consumption and investment with transaction costs and multiple risky assets. *The Journal of Finance*, 59(1):289–338, 2004.

Hong Liu and Mark Loewenstein. Optimal portfolio selection with transaction costs and finite horizons. *The Review of Financial Studies*, 15(3):805–835, 2002.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Anthony W. Lynch and Sinan Tan. Multiple risky assets, transaction costs, and return predictability: Allocation rules and implications for U.S. investors. *Journal of Financial and Quantitative Analysis*, 45(4):1015–1053, 2010.

Michael J. P. Magill and George M. Constantinides. Portfolio selection with transactions costs. *Journal of Economic Theory*, 13(2):245–263, 1976.

Johannes Muhle-Karbe, Martin Reppen, and H. Mete Soner. A primer on portfolio choice with small transaction costs. *Annual Review of Financial Economics*, 9:301–331, 2017. doi: 10.1146/annurev-financial-110716-032445.

Kumar Muthuraman and Sunil Kumar. Multidimensional portfolio optimization with proportional transaction costs. *Mathematical Finance*, 16(2):301–335, 2006.

Bernt Øksendal and Agnès Sulem. Optimal consumption and portfolio with both fixed and proportional transaction costs. *SIAM Journal on Control and Optimization*, 40(6):1765–1790, 2002.

Dylan Possamaï, H. Mete Soner, and Nizar Touzi. Homogenization and asymptotics for small transaction costs: The multidimensional case. *Communications in Partial Differential Equations*, 40(11):2005–2046, 2015.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Nicola Secomandi. Optimal commodity trading with a capacitated storage asset. *Management Science*, 56(3):449–467, 2010. doi: 10.1287/mnsc.1090.1049.

Steven E. Shreve and H. Mete Soner. Optimal investment and consumption with transaction costs. *The Annals of Applied Probability*, 4(3):609–692, 1994.

H. Mete Soner and Nizar Touzi. Homogenization and asymptotics for small transaction costs. *SIAM Journal on Control and Optimization*, 51(4):2893–2921, 2013.

Peter M. van de Ven, Nidhi Hegde, Laurent Massoulié, and Theodoros Salonidis. Optimal control of end-user energy storage. *IEEE Transactions on Smart Grid*, 4(2):789–797, 2013. doi: 10.1109/TSG.2012.2232943.

A Algorithm and numerical conventions

A.1 Reference continuation and execution convention

Condition on all fitted parameters and freeze a declared reference feedback $\bar{u}$, including any fitted continuation refinement. At grid time t_k, let $W_k(z) = V_{\eta,k}^{\bar{u},\Delta}(z)$ under the declared regularized transition F_η^Δ and payoff G_Δ, including domain treatment; for an unmodified Stage-I policy, $\bar{u} = u_{\theta^\star}$. The finite-refinement results require derivatives of this actual frozen continuation, not of an optimal value. Refinement queries the same field at successive hypothetical, cost-free states and executes only its final net displacement with proportional costs. Its input specification is $(\eta, \Delta, \delta_{\mathrm{FR}}, M_{\max}, \tau_{\mathrm{FR}}, \ell_{\min}, \Pi_{\mathcal{F}}, u^{\mathrm{fb}})$, together with the continuation budget and randomness convention; the finite-grid performance identity uses the actual post-refinement market transition $F_k^{0,\Delta}$, including its domain rule, and the same terminal payoff G_Δ. It compares two explicit discrete objectives and does not remove simulator bias; the portfolio configurations below do not establish every smoothness, error, or executed-step hypothesis of the finite-refinement theorems.

For the high-dimensional objective experiment, training, continuation estimation, virtual correction, final execution, and all baselines use the same long-only/no-borrowing closed simplex. Cash and individual risky holdings are allowed to equal zero. The implementation projects proposed finite rates onto the feasible set before transition; simultaneous sales may finance purchases. All policy objectives in Figure 7 are evaluated with $\eta_{\mathrm{eval}} = 0$. The regularized simulator is used to train $u_{\theta,\eta}$ and to define its frozen continuation, while the held-out objective uses the original proportional-cost market. An emergency absorbing exit is triggered only at the declared liquidation floor; zero cash or a zero individual risky holding is not itself an exit. These implementation choices are part of the numerical target differentiated by BPTT.

Notation. $R^0 = V_{0,y}/V_{0,x}$ and $R_\eta = V_{\eta,y}/V_{\eta,x}$ denote optimal singular and regularized ratios. $R^{\bar{u}}$ and $R^{\bar{u},\Delta}$ refer to the frozen continuous and discrete continuations; a hat denotes the sample-mean BPTT estimate and a tilde the median-of-means variant. The superscript Δ always identifies the declared simulator target, and b_Δ is its discrepancy from the continuous ratio. Market steps use k, fixed-time queries use m, and M_{MC} counts continuation paths. The partial sum $A_k^{(m)}$ is the virtual signed order after m completed updates, while $\mathcal{P}_\eta^{[d]}(\widehat{R})$ and $\mathcal{P}_\eta^{[\star]}(\widehat{R})$ denote the capped and adaptive *net-order* controllers (not iterates of the pointwise Yosida map). The separate $\pi_\eta^\star$ is a certified original-market strategy, not the adaptive learned controller or the normalized rate $\mathcal{P}_\eta(R)$.

A.2 Amortized BPTT-ratio distillation

The online BPTT controller recomputes a Monte Carlo continuation derivative at each queried state. To separate the value of the recovered ratio field from this online differentiation cost, we also train a feed-forward student ratio network $R_\theta(t, x, y)$. The teacher labels are the same feedback-BPTT ratios used by the corresponding BPTT refinement trajectory. For each reported mapping depth, teacher queries are collected from an independent rollout bank, including intermediate fixed-time states and the terminal audit state. Invalid teacher ratios are archived for diagnostics but excluded from supervised optimization; valid active states and valid endpoint states form the learning set.

Teacher trajectories are disjoint from the held-out paths used to estimate J_0. The supervised train/validation split is performed by outer teacher-path identifier rather than by individual query, preventing queries from the same trajectory from appearing on both sides of the split. The student minimizes a mean-squared ratio loss after scaling deviations from the wedge midpoint by η,

$$\mathcal{L}_{\mathrm{distill}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(q,\widehat{R})\in\mathcal{D}} \left\| \frac{R_\theta(q) - (1-\alpha/2)\mathbf{1}}{\eta} - \frac{\widehat{R}_{\mathrm{BPTT}}(q) - (1-\alpha/2)\mathbf{1}}{\eta} \right\|_2^2 . \tag{27}$$

Validation loss selects the retained student checkpoint. At deployment, R_θ is queried directly and passed through the same feasible Yosida realization and refinement code as the online BPTT ratio; no BPTT rollout or teacher lookup is performed online. The student results therefore test amortized retention of the teacher's ratio geometry, not an independent recovery or improvement guarantee.

A.3 DISCRETIZATION AND ALGORITHM

On $0 = t_0 < \cdots < t_K = T$, let $u_k = u_\theta(t_k, X_k, Y_k)$ and $L_k = X_k + \sum_i (1-\alpha_i) Y_k^i$. Euler–Maruyama gives

$$X_{k+1} = X_k + \left[r X_k - L_k \sum_i (u_k^i)^+ + L_k \sum_i (1-\alpha_i)(u_k^i)^- - \frac{\eta}{2} L_k \left\| u_k \right\|^2 \right] \Delta t, \tag{28}$$

$$Y_{k+1}^i = Y_k^i + \mu_i Y_k^i \Delta t + \sigma_i Y_k^i \Delta W_k^i + L_k u_k^i \Delta t. \tag{29}$$

These are the interior Euler formulas. Stage I maximizes the sample mean of the declared payoff $G_\Delta(Z_K^\Delta)$ by ordinary feedback BPTT; $G_\Delta = U \circ L$ when the payoff is unmodified and positive. After freezing $u_{\theta^\star}$, Stage II differentiates independent discrete continuation payoffs with respect to each query state and averages the derivatives. Stage III applies (17) once. The fallback is part of the

Algorithm 1 Unmodified-reference policy with an explicit failed-query branch

Require: Declared simulator F^Δ, payoff G_Δ, and a supplied complete admissible fallback u^{fb} for deployment.

1: Sample initial states and warm-start u_θ by stochastic gradient optimization of Monte Carlo terminal utility under the η-regularized simulator.
2: Freeze $u_{\theta^\star}$, retaining state-to-action derivatives.
3: **for** each deployment/query state q **do**
4: Simulate M_{MC} continuations under $u_{\theta^\star}$.
5: Feedback-differentiate $G_\Delta(Z_K^\Delta)$ with respect to q.
6: Estimate $\widehat{\nabla V}^{\,\Delta}$ by sample means, or by the median-of-means variant of Corollary B.7.
7: **if** the sensitivities are finite and $\widehat{V}_x^\Delta > 0$ **then**
8: Form $\widehat{R}_{\mathrm{BPTT}}^\Delta = \widehat{V}_y^\Delta / \widehat{V}_x^\Delta$.
9: Output $\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}}^\Delta(q))$ and a successful-query flag.
10: **else**
11: Record a failed query and output the prescribed fallback $u^{\mathrm{fb}}(q)$.
12: **end if**
13: **end for**

deployment specification, not an action recovered by the Yosida formula; its choice and frequency must be reported, and a global performance comparison includes the actual fallback actions. No denominator clipping is used in the theoretical projection.

Numerical domain and payoff. Raw Euler steps contain the Gaussian term $\sigma_i Y_k^i \sqrt{\Delta t}\, \xi_k^i$, so feasibility before a market step does not imply feasibility after it; for $\gamma \geq 2$ and a terminal density positive near zero, the unmodified terminal step even makes $\mathbb{E}[(L_T)^{1-\gamma} \mathbf{1}_{\{0 < L_T < \varepsilon\}}]$ diverge, which finite samples cannot reveal. The differentiated transition and payoff must therefore include the actual state-domain treatment, stopping rule (with its exit payoff), or payoff floor, each of which changes both the objective and its derivative; their contribution to the discrete-to-continuous ratio discrepancy is included in b_Δ.

A.4 FIXED-TIME CONTROL REFINEMENT: VIRTUAL REFINEMENT AND SINGLE-ORDER EXECUTION

This subsection specifies Stage IV. Three objects are kept distinct: (i) the *virtual query state* $\widetilde{z}_k^{(m)}$, which evolves only inside the refinement loop; (ii) the *cumulative net order* $A_k^{(m)}$, the start-to-current virtual displacement; and (iii) the *physical post-trade state* Z_k^+, obtained only after the final net order is charged transaction costs. Intermediate virtual moves are never charged, and the virtual endpoint is never treated as an executed portfolio.

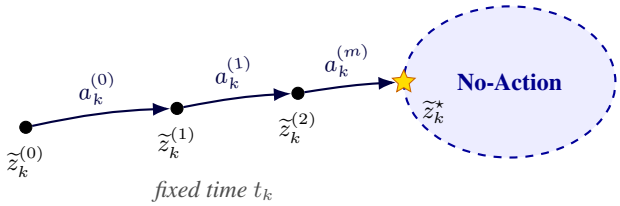


Figure 9: Virtual states $\widetilde{z}_k^{(m)}$ (black points) and increments $a_k^{(m)}$ (arrows) lead toward the terminal candidate $\widetilde{z}_k^{\star} = \widetilde{z}_k^{(M_k^{[\star]})}$ (star); contact is schematic, not guaranteed.

Two state maps and the exact fee convention. Fix a calendar time t_k and a physical predecision portfolio $z_k = (x_k, y_k)$, and define

$$\begin{aligned} \Gamma_0((x,y),a) &= (x - \mathbf{1}^\top a, y + a), \\ \Gamma_\alpha((x,y),A) &= (x - c_\alpha(A), y + A), \\ c_\alpha(A) &= \mathbf{1}^\top A + \sum_i \alpha_i A_i^-. \end{aligned} \tag{30}$$

Thus the final order alone incurs the cash deduction $\sum_i \alpha_i A_i^-$ relative to the cost-free transfer. Its liquidation wealth is $L(\Gamma_\alpha(z_k, A)) = L(z_k) - \sum_i \alpha_i A_i^+$, given the sale-side convention $L = x + \sum_i (1-\alpha_i) y_i$ and the initial long-only holdings. Intermediate buy/sell reversals are netted before execution.

Virtual and actual feasibility. Fix $\delta_{\mathrm{FR}} > 0$, tolerance $\tau_{\mathrm{FR}} \geq 0$, buffer $\ell_{\min} > 0$, and cap $M_{\max}$. Require each virtual state to have $\widetilde{y} \geq 0$ and $L(\widetilde{z}) \geq \ell_{\min}$, and every proposed *cumulative net* order A to satisfy

$$\mathcal{F}_{\mathrm{net}}(z_k) = \{A : y_k + A \geq 0, \quad L(z_k) - \sum_i \alpha_i A_i^+ \geq \ell_{\min}\}. \tag{31}$$

When $A^{(m)}$ is the already accumulated virtual order, the proposed increment a must keep $\Gamma_0(\widetilde{z}_k^{(m)}, a)$ virtually feasible and $A^{(m)} + a \in \mathcal{F}_{\mathrm{net}}(z_k)$. The feasible-increment operator Π_m is an explicitly specified measurable selection satisfying these two conditions and $\Pi_m(0) = 0$; radial clipping along a projected direction is an admissible example. The theoretical local bound below applies directly when the resulting increment has the form $a = d_m L(\widetilde{z}_k^{(m)}) u_k^{(m)}$.

Finite fixed-time refinement operator. Let $\widehat{R}^\Delta(t, z)$ denote the deployed ratio source. For the main online controller this is $\widehat{R}^\Delta_{\mathrm{BPTT}}$; the same refinement operator can be fed by the PINN or distilled ratio source used in the corresponding baselines. Initialize $\widetilde{z}_k^{(0)} = z_k$ and $A_k^{(0)} = 0$. Conditional on a valid active query at iteration $m < M_{\max}$, define

$$u_k^{(m)} = \mathcal{P}_\eta\left(\widehat{R}^\Delta(t_k, \widetilde{z}_k^{(m)})\right), \tag{32}$$

$$\widetilde{a}_k^{(m)} = \delta_{\mathrm{FR}} L(\widetilde{z}_k^{(m)}) u_k^{(m)}, \tag{33}$$

$$a_k^{(m)} = \Pi_m\left(\widetilde{a}_k^{(m)}\right), \tag{34}$$

$$\begin{aligned} A_k^{(m+1)} &= A_k^{(m)} + a_k^{(m)}, \\ \widetilde{z}_k^{(m+1)} &= \Gamma_0\left(\widetilde{z}_k^{(m)}, a_k^{(m)}\right). \end{aligned} \tag{35}$$

Equations (32)–(35) repeat Stage III at fixed physical time t_k, with a new Stage-II ratio query at every virtual state. No transaction cost, interest, drift, noise, or regularization leakage is applied between virtual iterates, so they are not steps of the physical simulator; only the final cumulative order $A_k^{(M_k)}$ reaches the market. A fallback that changes the virtual state is recorded as an actual virtual increment in $A_k^{(m)}$.

Stopping and physical execution. The theoretical stopping tolerance is applied to the recovered feasible action; the implementation-specific hold, stall, query-failure, and depth-limit tests are stated separately in Appendix A.5. The first successful queried hold index is

$$m_{\mathrm{hold}} = \inf\{m < M_{\max} : \|u_k^{(m)}\|_\infty \le \tau_{\mathrm{FR}}\}. \tag{36}$$

Set $M_k = m_{\mathrm{hold}}$ when finite, and otherwise use the actual number of completed virtual steps under the cap or failure/stall rule. The cumulative *virtual* risky-dollar displacement is

$$\begin{aligned} A_k^{(m)} &= \sum_{j<m} a_k^{(j)}, \qquad A_k^{(0)} = 0, \\ \mathcal{P}_\eta^{[d]}(\widehat{R})(q_k) &= A_k^{(M_k^{[d]})}, \qquad M_k^{[d]} \le d, \\ \mathcal{P}_\eta^{[\star]}(\widehat{R})(q_k) &= A_k^{(M_k^{[\star]})}, \qquad M_k^{[\star]} \le M_{\max}. \end{aligned} \tag{37}$$

For either controller, M_k denotes its actual number of completed virtual steps and its selected output is $A_k = A_k^{(M_k)}$. In particular, early hold, stall, or failed-query termination can give $M_k^{[d]} < d$; $[\star]$ is an adaptive finite-cap implementation, not an infinite-depth limit. The final physical state, used in the original-model market step, is

$$Z_k^+ = \Gamma_\alpha(z_k, A_k), \qquad Z_{k+1} = F_k^{0,\Delta}(Z_k^+, \xi_k). \tag{38}$$

The queried terminal virtual state is $\widetilde{Z}_{k,M_k} = \Gamma_0(z_k, A_k^{(M_k)})$, and in general $\widetilde{Z}_{k,M_k} \neq Z_k^+$: a zero-tolerance stop certifies wedge entry only at the virtual endpoint, not after the fee deduction, and a capped endpoint need not have been queried. The endpoint audit is therefore separate from the virtual stopping test.

Exact fee and switching residual. Writing $A = A_k$ and $f_k = \sum_i \alpha_i A_i^-$ gives

$$Z_k^+ = (\widetilde{X}_{k,M_k} - f_k, \widetilde{Y}_{k,M_k}), \qquad E_k^W := W_k(\widetilde{Z}_{k,M_k}) - W_k(Z_k^+). \tag{39}$$

If W_k is differentiable along the feasible cash segment, then

$$E_k^W = \int_0^{f_k} W_{k,x}(\widetilde{X}_{k,M_k} - s, \widetilde{Y}_{k,M_k})\, ds. \tag{40}$$

In particular $0 \le E_k^W \le f_k \overline{\lambda}_x$ when $0 \le W_{k,x} \le \overline{\lambda}_x$ throughout that segment. If R^W is K_R-Lipschitz there, the actual endpoint obeys

$$\mathrm{dist}_\infty(R^W(Z_k^+), \mathcal{C}_\alpha) \le \mathrm{dist}_\infty(R^W(\widetilde{Z}_{k,M_k}), \mathcal{C}_\alpha) + K_R f_k. \tag{41}$$

The bound requires both endpoint ratios to be defined; a terminal virtual hold does not automatically imply an actual hold. Performance statements are in Appendix C.2. The online policy is adapted when its continuation-query randomness is part of the controller's filtration and independent of subsequent market innovations. Feasibility of Z_k^+ does not establish solvency over the subsequent market interval; the complete fallback, domain, and market conventions remain part of admissibility.

A.5 Implemented refinement stopping and endpoint audit

The finite Stage-IV fixed-time refinement uses the same stopping logic for BPTT, PINN, and student-ratio implementations. Only the first event below is a successful numerical hold, and none is by itself a certificate of the continuous-time singular boundary.

Hold. At virtual state $\widetilde{z}_k^{(m)}$, a valid ratio query produces a feasible projected action $u_k^{(m)}$. The path is marked converged when

$$\|u_k^{(m)}\|_\infty \le \tau_{\mathrm{hold}}, \qquad \tau_{\mathrm{hold}} = 10^{-8}. \tag{42}$$

This criterion is applied after the same feasibility projection used by the executed controller. We additionally log whether the raw recovered ratio lies componentwise in $\mathcal{C}_\alpha$; a projected constraint-induced zero is therefore separable from literal raw-wedge membership.

Stall. If the path has not reached hold, one feasible cost-free virtual update produces a candidate $\widetilde{z}_{k,\mathrm{cand}}^{(m+1)}$. The update is declared immobile and the path is stopped when

$$\max\left\{|\widetilde{X}_{k,\mathrm{cand}}^{(m+1)} - \widetilde{X}_k^{(m)}|, \|\widetilde{Y}_{k,\mathrm{cand}}^{(m+1)} - \widetilde{Y}_k^{(m)}\|_\infty\right\} \le \tau_{\mathrm{pos}}, \qquad \tau_{\mathrm{pos}} = 10^{-4} \tag{43}$$

for the reported objective experiment. Stall is recorded separately from hold rather than being re-labeled as successful no-trade recovery.

Query failure. A ratio query is invalid when its source-specific validity test fails (for example, nonfinite recovered marginals or an invalid cash marginal for the BPTT ratio). A newly invalid live path is stopped and retained in the evaluation logs; it is not removed from the objective sample. Query failure is reported separately from stall, even though the implementation marks the failed row as no longer active in the refinement loop.

Depth limit and adaptive refinement. For $\mathcal{P}_\eta^{[k]}$, at most k successful virtual moves are attempted; the main objective experiment uses $k \in \{1, 2, 3\}$. The adaptive $\mathcal{P}_\eta^{[\star]}$ controller applies the same three stopping tests but uses $M_{\max} = 500$ as a safety cap. A path that ends without hold, stall, or query failure is labeled *depth limit*. Thus the empirical termination partition is

$$\{\text{hold},\ \text{stall},\ \text{depth limit},\ \text{query failure}\}.$$

Final endpoint audit and one-order execution. After the refinement loop, the terminal virtual state is queried once more for all paths, including finite-depth paths. This endpoint audit is included in the recorded ratio-query count and is used to distinguish numerical hold, raw-wedge membership, and constraint-induced hold at the terminal virtual state. The controller then converts the initial-to-final virtual displacement into one feasible net order and executes that order once under the original $\eta_{\mathrm{eval}} = 0$ transaction-cost simulator. Intermediate virtual moves incur neither market evolution nor transaction fees. Figure 15(b) reports the resulting termination frequencies; these diagnose the implemented controller and do not verify the hypotheses of Theorems C.3–C.4.

A.6 Implementation-to-ideal ratio bias

Remark A.1 (Discrete versus continuous ratio targets). Whenever $\partial_x V_{\eta,k}^{\theta^\star,\Delta} > 0$, define

$$R_{\theta^\star}^{\Delta} := \frac{\nabla_y V_{\eta,k}^{\theta^\star,\Delta}}{\partial_x V_{\eta,k}^{\theta^\star,\Delta}}, \qquad b_\Delta(q) := \left\| R_{\theta^\star}^{\Delta}(q) - R_{\theta^\star}(q) \right\|_\infty .$$

Proposition 2 is exact for the discrete target $R_{\theta^\star}^{\Delta}$, whereas Theorem 3 concerns the continuous target $R_{\theta^\star}$. Time discretization, deliberate smoothing of $(\cdot)^\pm$, and any declared boundary or payoff modification contribute to b_Δ. A derivative selection at a positive-probability kink that is not the gradient of a well-defined modified value is outside this statement. We do not claim $b_\Delta \to 0$ without a separate model-specific consistency result.

B Core recovery theory

B.1 Related work and regime-recovery positioning

Finite-horizon proportional-cost portfolio problems lead to free-boundary, double-obstacle, or gradient-constraint formulations Liu & Loewenstein (2002); Dai & Yi (2009); Dai et al. (2009); Belak & Sass (2019). Multi-asset no-trade regions depend on covariance and corner structure rather than independent scalar bands Liu (2004); Muthuraman & Kumar (2006); Chen & Dai (2013). Martingale, dual, shadow-price, and small-cost asymptotic approaches provide complementary characterizations Cvitanić & Karatzas (1996); Cvitanić & Wang (2001); Campi & Owen (2011); Kallsen & Muhle-Karbe (2010); Guasoni & Muhle-Karbe (2013); Muhle-Karbe et al. (2017); Janeček & Shreve (2004); Bichuch (2012); Soner & Touzi (2013); Possamaï et al. (2015); Choi et al. (2013); we use a consistent price system only for well-posedness. Fixed-cost or mixed fixed/proportional formulations instead lead to impulse control or quasi-variational inequalities Eastham & Hastings (1988); Korn (1998);

Øksendal & Sulem (2002); Altarovici et al. (2017). Empirical portfolio studies emphasize utility, turnover, and allocation effects Balduzzi & Lynch (1999); Lynch & Tan (2010), while neural PDE, PINN, deep-hedging/RL, and simulation-based methods offer alternative numerical routes E et al. (2017); Han et al. (2018); Raissi et al. (2019); Buehler et al. (2019); Fahim & Tsai (2018). Our target is different: we do not fit a global value function merely to extract its gradient, and we do not parameterize the free boundary. We use a frozen feedback continuation to obtain local marginal-value scores, interpret those scores as a buy/hold/sell detector, and apply a structure-preserving realization map. The distinction is consequential because a small nonzero action inside a strict hold region is a categorical regime error, not merely a small regression residual. The Moreau–Yosida interpretation explains why the realization layer preserves the entire zero-action wedge rather than smoothing it away Bauschke & Combettes (2017); Clarke (1990).

Energy-storage control beyond finance. Stochastic storage and arbitrage have been studied via optimal switching and capacitated-storage control, threshold battery policies, and deep RL with degradation-aware models Carmona & Ludkovski (2010); Secomandi (2010); van de Ven et al. (2013); Cao et al. (2020). Our deliberately low-dimensional benchmark instead isolates an asymmetric charge/hold/discharge wedge to test transfer and amortization of the same continuation-gradient regime detector and Yosida realization layer, not to propose a new state-of-the-art battery controller.

B.2 Interior gradient constraints and complementarity

On $\mathcal{D} \Subset (0,T) \times \mathcal{S}^\circ$, define

$$\mathcal{L}^0\varphi = rx\varphi_x + \sum_i \mu_i y_i \varphi_{y_i} + \frac{1}{2}\sum_{i,j} \rho_{ij}\sigma_i\sigma_j y_i y_j \varphi_{y_i y_j}. \tag{44}$$

The intervention directions are $g_i^+ = -e_x + e_{y_i}$ and $g_i^- = (1-\alpha_i)e_x - e_{y_i}$.

Meaning of the three HJB alternatives. The value cannot be increased by waiting briefly, buying immediately, or selling immediately. These three comparisons produce the diffusion term and the two gradient constraints below. Where both trade inequalities are strict, waiting is the active alternative and satisfies the diffusion PDE.

Theorem B.1 (Interior transaction-cost HJB)**.** *Assume the DPP and comparison for the state-constraint problem. On $\mathcal{D}$,*

$$\max\Big\{V_t + \mathcal{L}^0 V, \max_i(V_{y_i} - V_x), \max_i((1-\alpha_i)V_x - V_{y_i})\Big\} = 0. \tag{45}$$

If $V_x > 0$, the gradient constraints imply

$$1 - \alpha_i \le V_{y_i}/V_x \le 1 \qquad \textit{throughout the interior domain.}$$

At points where every inequality is strict, the diffusion equation holds with equality and no singular intervention is locally active. Equality identifies possible intervention faces and does not by itself characterize the state-space no-trade region.

Proof sketch. The DPP with no intervention over a short interval gives $V_t + \mathcal{L}^0 V \le 0$. Immediate purchases and sales cannot increase value, so $DV \cdot g_i^+ = V_{y_i} - V_x \le 0$ and $DV \cdot g_i^- = (1-\alpha_i)V_x - V_{y_i} \le 0$. Where all intervention inequalities are strict, the local DPP gives equality in the diffusion equation. These alternatives yield (45); for nonsmooth values, the same comparisons with touching test functions give the viscosity formulation. Dividing the gradient constraints by $V_x > 0$ gives the interior ratio wedge. Equality at a face permits intervention and therefore does not alone identify the state-space continuation set. □

Convex-analytic complementarity. In normalized coordinates the buy/hold/sell classification is the normal-cone relation $u \in N_{\mathcal{C}_\alpha}(R)$ on the ratio wedge $\mathcal{C}_\alpha = \prod_i [1-\alpha_i, 1]$; its single-valued Yosida regularization is the map of Proposition 1.

B.3 Realization part of Proposition 1

Proof. Coordinate calculation. The local trading criterion separates over assets. For one coordinate and scalar action a,

$$h_{\eta,i}(a; R_i) = \begin{cases} (R_i - 1)a - \eta a^2/2, & a \geq 0, \\ (R_i - (1 - \alpha_i))a - \eta a^2/2, & a \leq 0. \end{cases}$$

If $R_i > 1$, the positive branch is maximized at $(R_i - 1)/\eta$ and the negative branch cannot improve on zero. If $R_i < 1 - \alpha_i$, the negative branch is maximized at $(R_i - (1 - \alpha_i))/\eta$. Inside the closed wedge, each nonzero sign has a nonpositive linear contribution and a strictly negative quadratic contribution, so zero is uniquely optimal. Completing the square gives the maximum squared distance to the interval divided by 2η. Summing proves (9)–(11). *Convex-analytic interpretation.* The exchange cost is the support function of $\mathcal{C}_\alpha$, whose conjugate is $I_{\mathcal{C}_\alpha}$. The conjugate of $c_\alpha + (\eta/2)\|\cdot\|^2$ is the Moreau envelope $\mathrm{dist}^2(\cdot, \mathcal{C}_\alpha)/(2\eta)$, with gradient $\eta^{-1}(I - \mathrm{Proj}_{\mathcal{C}_\alpha})$ Bauschke & Combettes (2017). The operator resolvent is $\mathrm{Proj}_{\mathcal{C}_\alpha}$; its scaled residual is the Yosida action map. □

Unregularized local criterion. For $\kappa \geq 0$, write $h_\kappa(u; R) = R^\top u - c_\alpha(u) - \kappa\|u\|^2/2$.

Lemma B.2 (What the unregularized local problem does and does not determine)**.** *Let* $\mathcal{C}_\alpha = \prod_i [1 - \alpha_i, 1]$. *Then*

$$\sup_{u \in \mathbb{R}^N} h_0(u; R) = \begin{cases} 0, & R \in \mathcal{C}_\alpha, \\ +\infty, & R \notin \mathcal{C}_\alpha. \end{cases} \tag{46}$$

Inside the strict wedge, zero is the unique finite-rate maximizer; on a wedge face, maximizing rates can be nonunique and unbounded in size. For $g = \mathcal{P}_\eta(R)$,

$$h_0(g; R) = \eta\|g\|^2 = \frac{\mathrm{dist}(R, \mathcal{C}_\alpha)^2}{\eta}, \qquad h_\eta(g; R) = \frac{\eta}{2}\|g\|^2. \tag{47}$$

Consequently, the removed quadratic term is exactly one half of the unregularized local-criterion gain at this projected action. For a fixed ratio outside the wedge, its magnitude grows as $1/\eta$.

Proof. Since c_α is the support function of $\mathcal{C}_\alpha$, $R^\top u \leq c_\alpha(u)$ for $R \in \mathcal{C}_\alpha$, with equality at zero. If $R_i > 1$, choosing $u = se_i$, $s \uparrow \infty$, makes h_0 diverge; if $R_i < 1 - \alpha_i$, use $u = -se_i$. Strict interior membership makes every nonzero coordinate contribute strictly negatively. At a face, a nonzero normal direction has zero linear contribution at every positive scale. For the second assertion, optimality gives $R - \eta g \in \partial c_\alpha(g)$. A support-function subgradient p satisfies $p^\top g = c_\alpha(g)$: apply its subgradient inequality at 0 and $2g$ and use positive homogeneity. Taking the inner product with g gives $R^\top g - c_\alpha(g) = \eta\|g\|^2$. Subtract the quadratic term and use $\eta g = R - \mathrm{Proj}_{\mathcal{C}_\alpha} R$ to obtain the formulas. □

Graph approximation does not fix the input ratio. The projection characterization gives

$$\bar{q}_\eta = \mathrm{Proj}_{\mathcal{C}_\alpha}(q_\eta), \qquad q_\eta = \bar{q}_\eta + \eta u_\eta, \qquad u_\eta = \mathcal{P}_\eta(q_\eta) \in N_{\mathcal{C}_\alpha}(\bar{q}_\eta). \tag{48}$$

For any $u \in N_{\mathcal{C}_\alpha}(R)$, choosing $q_\eta = R + \eta u$ therefore yields $\mathcal{P}_\eta(q_\eta) = u$. Conversely, if $q_\eta \to R$ and $\mathcal{P}_\eta(q_\eta) \to u$, closedness of the normal-cone graph gives $u \in N_{\mathcal{C}_\alpha}(R)$. This is graph convergence: at a fixed $R \in \mathcal{C}_\alpha$, $\mathcal{P}_\eta(R) = 0$ for every η.

Why a smooth exchange surrogate can lose the wedge. For a differentiable mollified exchange cost $c_{\alpha,\tau}$, an unconstrained optimum at zero requires $R - c'_{\alpha,\tau}(0) = 0$. Thus zero can be optimal only at a single ratio, in contrast to the full interval preserved by the nonsmooth exchange cost.

B.4 Discrete differentiability and the BPTT identity

Assumption B.3 (Discrete differentiation and domination)**.** *Fix a query time* t_k *and state* z *in the interior positive-liquidation band. The transition* F_j^Δ *denotes the transition actually differentiated in the implementation, including any explicitly specified smooth exchange surrogate and boundary treatment. Let* G_Δ *be the actual terminal payoff, including any modification. Assume* G_Δ *is locally Lipschitz on the reached terminal band and differentiable with conditional probability one at the terminal states. For every* $j \geq k$, *assume one of the following two alternatives:*

(a) $u_{\theta^\star,j}$ and F_j^Δ are C^1 on a neighborhood of the reachable localized band; or

(b) they are locally Lipschitz and differentiable outside null sets, and the frozen trajectory hits the union of their nondifferentiability sets with conditional probability zero.

In either case, assume the declared terminal objective is integrable and, in a common neighborhood of z, the pathwise terminal-payoff map is Lipschitz with a Lipschitz constant bounded by an integrable random variable. This dominates both its difference quotients and its Jacobian-product sensitivities, permitting differentiation under expectation. A sufficient localized setting has terminal wealth in a compact positive interval and bounded local payoff derivatives; this is an assumption on the simulator's boundary rule, not a property of Gaussian Euler evolution.

B.5 Proof of Proposition 2

Proof. Differentiate one path. Fix the innovations. The terminal payoff is the finite composition

$$z \longmapsto G_\Delta\big(Z_K^\Delta(Z_k^\Delta = z, \xi_k, \dots, \xi_{K-1})\big).$$

Perturbing a state changes the next state directly through $D_z F_j^\Delta$ and indirectly through $D_u F_j^\Delta D_z u_{\theta^\star,j}$. The chain rule, with terminal derivative $D_z G_\Delta$, gives (15). *Then average the derivatives.* The neighborhood domination in Assumption B.3 permits differentiation under expectation, so

$$D_z V_{\eta,k}^{\theta^\star,\Delta}(z) = \mathbb{E}[\Lambda_k^{\mathrm{pw},\Delta} \mid Z_k^\Delta = z].$$

This identifies a derivative of the declared discrete continuation value. It does not replace that value by an optimal or continuous-time value. □

Remark B.4 (Smooth targets, domination, and positive-probability kinks). Under these assumptions Proposition 2 is exact for the declared discrete continuation value; smoothing, boundary or payoff modification, and time discretization enter only its discrepancy b_Δ from the continuous target, and a value-level smoothing estimate would not by itself control this gradient-ratio bias. A bare automatic-differentiation selection at a positive-probability kink need not equal a classical value gradient and is not covered. On a finite-horizon localized band, bounded policy Jacobians and transition-Jacobian bounds $\|DF_j^\Delta\| \le C(1 + \|\xi_j\|)$ suffice for domination under Gaussian increments.

B.6 Proof of Theorem 3

Conditions for the whole-horizon comparison. Fix the initial pair (t, z). Assume $V_\eta^{\theta^\star} \in C^{1,2}$ satisfies the fixed-policy PDE

$$\partial_t V_\eta^{\theta^\star} + \mathcal{L}^{u_{\theta^\star}} V_\eta^{\theta^\star} = 0$$

on a domain covering the candidate trajectories, has terminal value $U(L)$, and satisfies $\partial_x V_\eta^{\theta^\star} > 0$. All feedbacks compared below are admissible. The local martingale terms in Itô's formula are true martingales, and the local-criterion differences are integrable. Alternatively, a localization argument must supply the correct exit continuation values and justify the limit to T.

Proof. Step 1: accumulate local gains along the candidate trajectory. Apply Itô's formula to $V_\eta^{\theta^\star}(s, Z_s^v)$ while the state follows an admissible candidate feedback v. Since the diffusion coefficient is unaffected by the trading rate, subtracting the fixed-policy PDE leaves only the difference of local trading criteria:

$$J_\eta(v) - J_\eta(u_{\theta^\star}) = \mathbb{E}^v \int_t^T \Big[\mathcal{G}_\eta(v; DV_\eta^{\theta^\star}) - \mathcal{G}_\eta(u_{\theta^\star}; DV_\eta^{\theta^\star})\Big](s, Z_s^v)\, ds. \tag{49}$$

The occupancy in this identity belongs to the candidate policy. *Step 2: exact Yosida correction.* For fixed R, $h_\eta(\cdot; R)$ is η-strongly concave and its unique maximizer is $\mathcal{P}_\eta(R)$. Hence

$$h_\eta(\mathcal{P}_\eta(R); R) - h_\eta(a; R) \ge \frac{\eta}{2}\|a - \mathcal{P}_\eta(R)\|^2.$$

Set $R = R_{\theta^\star}$ and multiply by $L\,\partial_x V_\eta^{\theta^\star} > 0$. Using (49) with $v = \mathcal{P}_\eta(R_{\theta^\star})$ gives (18); optimality of $u_{\theta^\star}$ is not required. *Step 3: subtract the loss caused by BPTT ratio error.* At a fixed state set

$$R = R_{\theta^\star}, \qquad g = \mathcal{P}_\eta(R_{\theta^\star}), \qquad b = \mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}}).$$

Strong concavity for the criterion indexed by $\widehat{R}_{\mathrm{BPTT}}$ and the linear dependence on the ratio give

$$\begin{aligned} h_\eta(g;R) - h_\eta(b;R) &= h_\eta(g;\widehat{R}_{\mathrm{BPTT}}) - h_\eta(b;\widehat{R}_{\mathrm{BPTT}}) + (R-\widehat{R}_{\mathrm{BPTT}})^\top(g-b) \\ &\le -\frac{\eta}{2}\|g-b\|^2 + (R-\widehat{R}_{\mathrm{BPTT}})^\top(g-b) \\ &\le \frac{\|R-\widehat{R}_{\mathrm{BPTT}}\|^2}{2\eta}. \end{aligned}$$

Combining this with the exact greedy gain yields

$$h_\eta(\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}});R_{\theta^\star}) - h_\eta(u_{\theta^\star};R_{\theta^\star}) \ge \frac{\eta}{2}\left\|u_{\theta^\star}-\mathcal{P}_\eta(R_{\theta^\star})\right\|^2 - \frac{1}{2\eta}\left\|\widehat{R}_{\mathrm{BPTT}}-R_{\theta^\star}\right\|^2.$$

Multiplying by $L\,\partial_x V_\eta^{\theta^\star}$ and applying (49) with $v=\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})$ gives

$$\begin{aligned} &J_\eta\Big(\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})\Big) - J_\eta(u_{\theta^\star}) \\ &\quad\ge \mathbb{E}^{\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})}\int_t^T L_s\,\partial_x V_\eta^{\theta^\star}(s,Z_s)\left[\frac{\eta}{2}\left\|u_{\theta^\star}-\mathcal{P}_\eta(R_{\theta^\star})\right\|^2 - \frac{1}{2\eta}\left\|\widehat{R}_{\mathrm{BPTT}}-R_{\theta^\star}\right\|^2\right]ds. \end{aligned} \tag{50}$$

The sufficient condition (19) makes the integrand nonnegative and proves (20). □

B.7 Occupancy-weighted strong growth

A policy's utility detects its actions only at visited states, so strong concavity first controls an occupancy-weighted action error; a separate coverage inequality then yields a fixed norm.

Lemma B.5 (A sufficient occupancy-weighted growth mechanism)**.** *Let $u_\eta^\star$ be an admissible feedback on $\mathcal{D}$ whose classical fixed-policy value $V_\eta^\star$ satisfies the assumptions of Theorem 3, and suppose*

$$u_\eta^\star(t,z) = \mathcal{P}_\eta(R_\eta^\star(t,z)) \qquad \textit{throughout } \mathcal{D}.$$

Then every admissible feedback v with trajectories in $\mathcal{D}$ satisfies

$$J_\eta(u_\eta^\star) - J_\eta(v) \ge \frac{\eta}{2}\mathbb{E}^v\int_t^T L_s V_{\eta,x}^\star(s,Z_s^v)\left\|v(s,Z_s^v)-u_\eta^\star(s,Z_s^v)\right\|^2 ds. \tag{51}$$

If $L \ge \underline{L} > 0$ and $V_{\eta,x}^\star \ge \underline{\lambda} > 0$ on $\mathcal{D}$, the right-hand side is at least

$$\frac{\eta\underline{L}\underline{\lambda}}{2}\mathbb{E}^v\int_t^T \left\|v-u_\eta^\star\right\|^2 ds.$$

Consequently, the fixed-reference-norm quadratic growth (56) *follows with $\kappa_{\mathrm{qg}}(\eta) \ge (\eta\underline{L}\underline{\lambda}c_{\mathrm{occ}})/2$ whenever the additional coverage inequality*

$$\mathbb{E}^v\int_t^T \left\|v-u_\eta^\star\right\|^2 ds \ge c_{\mathrm{occ}}\left\|v-u_\eta^\star\right\|_{\mathfrak{U}}^2 \tag{52}$$

holds uniformly over the policy neighborhood.

Proof. Apply the performance-difference identity with reference feedback $u_\eta^\star$. Pointwise self-greediness and η-strong concavity of $h_\eta(\cdot;R_\eta^\star)$ give

$$\mathcal{G}_\eta(u_\eta^\star;DV_\eta^\star) - \mathcal{G}_\eta(v;DV_\eta^\star) \ge \frac{\eta}{2}LV_{\eta,x}^\star\left\|v-u_\eta^\star\right\|^2.$$

Integrating under the candidate-policy occupancy proves (51); the compact-band and coverage conclusions follow immediately. □

Because the occupancy measure depends on v, the lemma supplies the fixed norm of Theorem B.6 only under (52); it does not verify the ratio-stability constant $C_R(\eta,\nu)$.

B.8 Robustness part of Proposition 1

Let $R, \widetilde{R}$ satisfy the margin condition (12). The interval between R_i and $\widetilde{R}_i$ crosses neither switching face, so both lie in the same open buy/hold/sell regime. If $R_i \in (1-\alpha_i, 1)$, then

$$[\mathcal{P}_\eta(R)]_i = [\mathcal{P}_\eta(\widetilde{R})]_i = 0. \tag{53}$$

For this strict-hold score, writing $m_i = \operatorname{dist}(R_i, \{1-\alpha_i, 1\})$, direct evaluation of the two sign branches also gives, for $a \neq 0$,

$$h_{\eta,i}(0; R_i) - h_{\eta,i}(a; R_i) \geq m_i|a| + \frac{\eta}{2}a^2 > 0. \tag{54}$$

Thus zero strictly dominates every nonzero scalar action at a strict-hold score; the same result applies to any two BPTT ratio estimates.

The bridge. Quadratic growth converts a small Stage-I utility gap into policy distance, ratio stability converts policy distance into ratio distance on the deployment measure, and BPTT acquisition error is added last. This chain states precisely what is needed to connect high rollout utility of $u_{\theta^\star}$ to accurate Yosida actions.

Theorem B.6 (Conditional fixed-η Stage-I-to-action bridge)**.** *Fix $\eta > 0$. Let $u_\eta^\star$ be an interior locally optimal regularized feedback satisfying*

$$u_\eta^\star = \mathcal{P}_\eta(R_\eta^\star),$$

and suppose $u_{\theta^\star}$ and $u_\eta^\star$ belong to a policy neighborhood $\mathcal{N}_\eta$ in a specified normed feedback space $\mathfrak{U}$. Fix the initial condition used by J_η, and let ν be a finite deployment measure on state–time queries. Define the Stage-I objective gap

$$\epsilon_{\theta^\star} := J_\eta(u_\eta^\star) - J_\eta(u_{\theta^\star}) \geq 0. \tag{55}$$

Assume that, for $u, v \in \mathcal{N}_\eta$,

$$J_\eta(u_\eta^\star) - J_\eta(u) \geq \kappa_{\mathrm{qg}}(\eta)\, \|u - u_\eta^\star\|_{\mathfrak{U}}^2, \tag{56}$$

$$\|R^u - R^v\|_{L^2(\nu)} \leq C_R(\eta, \nu)\, \|u - v\|_{\mathfrak{U}}. \tag{57}$$

Assume also that $V_x^w > 0$ ν-a.e. for every $w \in \mathcal{N}_\eta$, and let

$$M_\eta := \|LV_{\eta,x}^\star\|_{L^\infty(\nu)} < \infty.$$

For the acquired BPTT ratio $\widehat{R}^{\Delta}_{\mathrm{BPTT}}$, set

$$\delta_{\mathrm{acq}}(\nu) := \left\|\widehat{R}^{\Delta}_{\mathrm{BPTT}} - R_{\theta^\star}\right\|_{L^2(\nu)}, \qquad e_\eta(\nu) := C_R(\eta, \nu)\sqrt{\frac{\epsilon_{\theta^\star}}{\kappa_{\mathrm{qg}}(\eta)}} + \delta_{\mathrm{acq}}(\nu). \tag{58}$$

Then

$$\left\|\mathcal{P}_\eta(\widehat{R}^{\Delta}_{\mathrm{BPTT}}) - u_\eta^\star\right\|_{L^2(\nu)} \leq \frac{e_\eta(\nu)}{\eta}, \tag{59}$$

$$\int \left[\mathcal{G}_\eta(u_\eta^\star; DV_\eta^\star) - \mathcal{G}_\eta(\mathcal{P}_\eta(\widehat{R}^{\Delta}_{\mathrm{BPTT}}); DV_\eta^\star)\right] d\nu \leq \frac{M_\eta}{2\eta} e_\eta(\nu)^2. \tag{60}$$

Proof. See Appendix B.9. □

B.9 Proof of Theorem B.6

Quadratic growth gives

$$\|u_{\theta^\star} - u_\eta^\star\|_{\mathfrak{U}} \leq \sqrt{\frac{\epsilon_{\theta^\star}}{\kappa_{\mathrm{qg}}(\eta)}}.$$

Ratio stability then yields

$$\|R_{\theta^\star} - R_\eta^\star\|_{L^2(\nu)} \leq C_R(\eta, \nu)\sqrt{\frac{\epsilon_{\theta^\star}}{\kappa_{\mathrm{qg}}(\eta)}}.$$

Adding $\delta_{\mathrm{acq}}(\nu)$ gives

$$\left\|\widehat{R}^{\Delta}_{\mathrm{BPTT}} - R^{\star}_{\eta}\right\|_{L^2(\nu)} \le e_{\eta}(\nu).$$

Since $\mathcal{P}_{\eta}$ is $1/\eta$-Lipschitz and $u^{\star}_{\eta} = \mathcal{P}_{\eta}(R^{\star}_{\eta})$, (59) follows. The strong-concavity argument of Appendix B.6 gives

$$h_{\eta}(\mathcal{P}_{\eta}(R^{\star}_{\eta}); R^{\star}_{\eta}) - h_{\eta}(\mathcal{P}_{\eta}(\widehat{R}^{\Delta}_{\mathrm{BPTT}}); R^{\star}_{\eta}) \le \frac{1}{2\eta}\left\|\widehat{R}^{\Delta}_{\mathrm{BPTT}} - R^{\star}_{\eta}\right\|^2. \tag{61}$$

Multiplication by $LV^{\star}_{\eta,x}$ and integration prove (60).

Corollary B.7 (Robust finite-sample regime recovery)**.** *Fix a query state q. Suppose the coordinates of the pathwise BPTT sensitivity targeting $V^{\theta^{\star},\Delta}_{\eta,k}$ have conditional variance at most σ^2_{Λ}, and*

$$\partial_x V^{\theta^{\star},\Delta}_{\eta,k}(q) \ge \underline{\lambda} > 0, \qquad \max_i \left|\partial_{y_i} V^{\theta^{\star},\Delta}_{\eta,k}(q)\right| \le \overline{V}_y.$$

Let $m^{\Delta}(q) := \nabla V^{\theta^{\star},\Delta}_{\eta,k}(q)$ and let $\widetilde{m}^{\Delta}$ be its coordinatewise median-of-means estimator based on M_{MC} independent continuations. Write $\widetilde{V}^{\Delta}_x$ and $\widetilde{V}^{\Delta}_{y_i}$ for the corresponding cash and risky marginal estimates, and set $\widetilde{R}^{\Delta}_{\mathrm{BPTT},i} = \widetilde{V}^{\Delta}_{y_i}/\widetilde{V}^{\Delta}_x$. Assume M_{MC} is large enough that the coordinatewise median-of-means radius is at most $\underline{\lambda}/2$. Then, with probability at least $1-\delta$,

$$\max_i \left|\widetilde{R}^{\Delta}_{\mathrm{BPTT},i} - (R^{\Delta}_{\theta^{\star}})_i\right| \le C_{\mathrm{ratio}}\sigma_{\Lambda}\sqrt{\frac{\log(2(N+1)/\delta)}{M_{\mathrm{MC}}}}, \qquad C_{\mathrm{ratio}} = C\left(\frac{1}{\underline{\lambda}} + \frac{\overline{V}_y}{\underline{\lambda}^2}\right), \tag{62}$$

for a universal constant C. Consequently,

$$\max_i \left|\widetilde{R}^{\Delta}_{\mathrm{BPTT},i} - (R_{\theta^{\star}})_i\right| \le b_{\Delta}(q) + C_{\mathrm{ratio}}\sigma_{\Lambda}\sqrt{\frac{\log(2(N+1)/\delta)}{M_{\mathrm{MC}}}}. \tag{63}$$

By Proposition 1, every coordinate whose continuous-time fixed-policy margin

$$m^{\theta^{\star}}_i(q) := \mathrm{dist}((R_{\theta^{\star}})_i(q), \{1-\alpha_i, 1\})$$

exceeds the right-hand side of (63) *has the correct greedy regime. If*

$$b_{\theta^{\star}}(q) := \left\|R_{\theta^{\star}}(q) - R^{\star}_{\eta}(q)\right\|_{\infty}$$

is separately controlled, the same statement applies to the optimal regularized regime after adding $b_{\theta^{\star}}(q)$ to the threshold. For a positive residual margin $\bar{m} = m - b_{\Delta}(q) - b_{\theta^{\star}}(q)$, the required rollout budget is $O(\bar{m}^{-2}\log(N/\delta))$ when the variance and marginal-bound constants are held fixed; they may themselves depend on N and η.

Proof. See Appendix B.10. □

This is a fixed-query result for the median-of-means estimator; the finite-variance hypothesis alone does not give the same tail bound for the default sample mean, and neither warm-start bias nor discrete-target bias is removed by this sampling argument. Simultaneous guarantees along adaptive queries require conditional concentration and a union bound over queries (Appendix C.2), and antithetic rollouts must be grouped into independent pairs or blocks.

B.10 Median-of-means concentration

Divide M_{MC} independent pathwise sensitivities into $B_{\mathrm{MoM}} \asymp \log(2(N+1)/\delta)$ blocks, average within each block, and take the coordinatewise median. Let $m^{\Delta}_{\theta^{\star}} = \nabla V^{\theta^{\star},\Delta}_{\eta,k}(q)$. Standard median-of-means concentration under finite variance Lugosi & Mendelson (2019) and a union bound give

$$\max_j \left|\widetilde{m}^{\Delta}_j - (m^{\Delta}_{\theta^{\star}})_j\right| \le C_{\mathrm{mom}}\sigma_{\Lambda}\sqrt{\frac{\log(2(N+1)/\delta)}{M_{\mathrm{MC}}}} \tag{64}$$

with probability at least $1-\delta$. On the event that the cash-marginal error is at most $\underline{\lambda}/2$, the elementary ratio inequality

$$\left|\frac{\widetilde{a}}{\widetilde{b}} - \frac{a}{b}\right| \le \frac{2}{\underline{\lambda}}|\widetilde{a}-a| + \frac{2\overline{V}_y}{\underline{\lambda}^2}|\widetilde{b}-b| \tag{65}$$

and (64) prove (62), after absorbing universal factors into C. The triangle inequality adds $b_\Delta(q)$ to obtain (63), and adds $b_{\theta^\star}(q)$ when the target is $R^\star_\eta$. CRRA sensitivities may be heavy-tailed near the lower solvency boundary; with only a finite pth moment, $1 < p < 2$, a different robust estimator and a slower rate are required.

C FINITE LEARNED CONTROL REFINEMENT

C.1 WHAT A FINITE VIRTUAL UPDATE CAN IMPROVE

A projected rate provides a direction for an internal hypothetical cash/risky transfer. Its gain is measured before the final fee is charged, so it cannot alone certify objective improvement.

Theorem C.1 (A certified virtual-update gain and a step bound)**.** *Fix (s, z) and a comparison function W. Let $W_x(s, z) > 0$, $u = \mathcal{P}_\eta(\widehat{R}(s, z))$, and let the* virtual *update be $a = dL(z)u$ for an effective step $d \geq 0$. Suppose the entire segment $z + \theta(\Gamma_0(z, a) - z)$, $0 \leq \theta \leq 1$, lies in a feasible C^2 region for W where $\|D_z^2 W\|_{\mathrm{op}} \leq K_W$. Set $e = \|\widehat{R} - R^W\|$ at z. Then*

$$W(s, \Gamma_0(z, a)) - W(s, z) \geq dLW_x \left(\eta\|u\|^2 + \sum_i \alpha_i u_i^- - e\|u\| \right) - \frac{K_W d^2 L^2}{2} \left((\mathbf{1}^\top u)^2 + \|u\|^2 \right) \tag{66}$$

$$\geq dLW_x(\eta\|u\|^2 - e\|u\|) - \frac{K_W(N+1)d^2L^2}{2}\|u\|^2.$$

For $u \neq 0$, the sufficient conditions

$$e \leq \frac{\eta}{2}\|u\|, \qquad 0 \leq d \leq \frac{\eta W_x}{2K_W L(N+1)} \tag{67}$$

imply

$$W(s, \Gamma_0(z, a)) - W(s, z) \geq \frac{d\eta}{4} LW_x \|u\|^2 \geq 0. \tag{68}$$

For $K_W = 0$, omit the curvature restriction; $u = 0$ gives zero virtual displacement and gain. No conclusion about the actual charged endpoint follows without an execution-cost correction.

Proof. By the signwise identity for the Yosida optimizer, $\widehat{R}^\top u - c_\alpha(u) = \eta\|u\|^2$. The cost-free transfer has direction $\Gamma_0(z, dLu) - z = dL(-\mathbf{1}^\top u, u)$, and $c_\alpha(u) = \mathbf{1}^\top u + \sum_i \alpha_i u_i^-$. Its first-order value change therefore equals

$$dLW_x(R^{W\top} u - \mathbf{1}^\top u) = dLW_x \left(\eta\|u\|^2 + \sum_i \alpha_i u_i^- + (R^W - \widehat{R})^\top u \right).$$

Cauchy–Schwarz bounds the final term below by $-e\|u\|$. Taylor's integral remainder has magnitude at most $K_W d^2 L^2 [(\mathbf{1}^\top u)^2 + \|u\|^2]/2$. Using $(\mathbf{1}^\top u)^2 \leq N\|u\|^2$ yields the second bound. Under (67), the first-order term is at least $dLW_x \eta\|u\|^2/2$ and the curvature loss at most $dLW_x \eta\|u\|^2/4$. □

Scope and feasibility. The theorem concerns W at virtual states, not executed trades. Radial clipping preserves the form $a = dLu$ with a smaller effective d; a nonradial virtual increment must be checked directly through

$$W(s, \Gamma_0(z, a)) - W(s, z) \geq DW(s, z)^\top(-\mathbf{1}^\top a, a) - \frac{K_W}{2}\left((\mathbf{1}^\top a)^2 + \|a\|^2 \right).$$

Both virtual-segment feasibility and *final net-order* feasibility (31) are required.

Counterexample C.2 (A feasible oversized trade can lower continuation value). Take one asset, $\alpha = 1/10$, $\gamma = 3$, $r = 0$, $T = 1$, $\mu = \log(5/4)$, $\sigma = 1$, and initial $(x, y) = (1, 0)$. For the no-trade reference $v = 0$, write the terminal gross risky return as $G = \exp(\mu - 1/2 + B_1)$, where $B_1 \sim N(0, 1)$. Its continuation is $W(0, x, y) = \mathbb{E}[U(x + (9/10)yG)]$ on the positive feasible region, and its marginals at $(1, 0)$ are

$$W_x = 1, \qquad R^W = \frac{9}{10}\mathbb{E}G = \frac{9}{8}.$$

These derivatives follow by differentiating in the feasible directions; finite lognormal moments justify that calculation. With $\eta = 1$, the exact projected rate is $u = 1/8$. Choose virtual refinement step $d = 8$, obtain the net target $a = 1$, execute this single feasible trade (charging costs only now), and then follow the warm-start policy $v = 0$ for the remaining horizon. This is a chosen policy continuation, not a restriction on available trading times. The new state is $(0, 1)$, and exact-market terminal liquidation wealth is $(9/10)G > 0$. All CRRA moments are finite. Nevertheless,

$$J_0(\pi_{\mathrm{FR},1}) = \frac{1 - (64/81)e^3}{2} < 0 = J_0(v). \tag{69}$$

Indeed, $\mathbb{E}G^{-2} = \exp(-2\mu + 3) = (16/25)e^3$ and $(9/10)^{-2}(16/25) = 64/81$. Thus even with cost-free virtual iteration, a correct initial direction, feasible single net trade, exact market evaluation, and finite cap do not imply policy improvement. The finite-step curvature condition is substantive. The same trade and loss occur for any $\eta > 0$ by taking $u = 1/(8\eta)$ and $d = 8\eta$. Hence even $d \downarrow 0$ together with $\eta \downarrow 0$ is insufficient without control of the resulting dollar displacement and the curvature bound.

C.2 Finite inexact fixed-time refinement on the declared simulator

Fix the trained continuation policy $\bar{u}$, its declared discrete simulator and terminal payoff, and set $W_k = V_{\eta,k}^{\bar{u},\Delta}$ (conditional on the trained parameters) with target ratio $R_k^W = W_{k,y}/W_{k,x}$. All norms are Euclidean unless the infinity norm is displayed. The results below require positivity and regularity of this frozen continuation, not of an optimal value.

Theorem C.3 (Finite inexact virtual refinement: progress and stopping)**.** *Fix k, suppress k, and write $W = W_k$. Consider $M \geq 1$ virtual steps*

$$u_m = \mathcal{P}_\eta(\widehat{R}_m), \quad a_m = d_m L(\widetilde{z}_m) u_m, \quad \widetilde{z}_{m+1} = \Gamma_0(\widetilde{z}_m, a_m), \quad \widetilde{z}_0 = z_0.$$

Queries and effective radial steps may be adaptive. Suppose all virtual segments are feasible in a region where $W \in C^2$, $L > 0$, $W_x > 0$, and $\|D^2 W\|_{\mathrm{op}} \leq K_W$, and suppose the cumulative net orders remain executable after the final transaction cost. For $K_W > 0$, require

$$0 \leq d_m \leq \frac{\eta W_x(\widetilde{z}_m)}{2K_W L(\widetilde{z}_m)(N+1)}. \tag{70}$$

When $K_W = 0$, only feasibility is required. Put

$$e_m = \|\widehat{R}_m - R^W(\widetilde{z}_m)\|, \qquad r_m = \mathrm{dist}(\widehat{R}_m, \mathcal{C}_\alpha) = \eta\|u_m\|,$$
$$\omega_m = d_m L(\widetilde{z}_m) W_x(\widetilde{z}_m), \qquad S_M = \sum_{m<M} \omega_m.$$

Then, pathwise,

$$W(\widetilde{z}_{m+1}) - W(\widetilde{z}_m) \geq \frac{\omega_m}{\eta}\left(\frac{1}{2} r_m^2 - e_m^2\right). \tag{71}$$

If $S_M > 0$ and $W \leq W_{\max}$ at the endpoint, then

$$\min_{0 \leq m < M} \mathrm{dist}(R^W(\widetilde{z}_m), \mathcal{C}_\alpha)^2 \leq \frac{4\eta[W_{\max} - W(z_0)]}{S_M} + 6\frac{\sum_{m<M} \omega_m e_m^2}{S_M}. \tag{72}$$

If a successful queried virtual *state $\widetilde{z}_{m_*}$ obeys $\|u_{m_*}\|_\infty \leq \tau_{\mathrm{FR}}$, then*

$$\mathrm{dist}_\infty(R^W(\widetilde{z}_{m_*}), \mathcal{C}_\alpha) \leq \eta\tau_{\mathrm{FR}} + \|\widehat{R}_{m_*} - R^W(\widetilde{z}_{m_*})\|_\infty. \tag{73}$$

This does not certify the fee-adjusted endpoint. If R^W is K_R-Lipschitz along the fee-correction segment and $f = \sum_i \alpha_i A_i^-$, its actual residual is bounded by the right-hand side plus $K_R f$.

Proof. Theorem C.1 and (70) bound the virtual curvature loss by $\omega_m \eta \|u_m\|^2/4$. Discarding the additional nonnegative sell-direction term gives

$$W(\widetilde{z}_{m+1}) - W(\widetilde{z}_m) \geq \frac{\omega_m}{\eta}\left(\frac{3}{4} r_m^2 - e_m r_m\right) \geq \frac{\omega_m}{\eta}\left(\frac{1}{2} r_m^2 - e_m^2\right),$$

using $e_m r_m \leq r_m^2/4 + e_m^2$. Summing yields $\sum_m \omega_m r_m^2 \leq 2\eta[W_{\max} - W(z_0)] + 2\sum_m \omega_m e_m^2$. Since distance to a closed set is 1-Lipschitz, $\mathrm{dist}(R^W(\widetilde{z}_m), \mathcal{C}_\alpha)^2 \leq 2r_m^2 + 2e_m^2$. Multiplying by ω_m, summing and dividing by S_M proves the residual estimate. The box identity $\mathrm{dist}_\infty(\widehat{R}, \mathcal{C}_\alpha) = \eta\|\mathcal{P}_\eta(\widehat{R})\|_\infty$ and the triangle inequality prove the stop bound. A final application of the Lipschitz bound and (41) proves the claimed fee-adjusted residual extension. □

Effective progress, clipping, and caps. For example, uniform bounds $L \le L_+$ and $W_x \ge \lambda_- > 0$ make $d_m \le \eta\lambda_-/[2K_W L_+(N+1)]$ sufficient. If additionally $L \ge L_- > 0$, $d_m \ge d_- > 0$, and $e_m \le \varepsilon$, then (72) implies

$$\min_{m<M} \operatorname{dist}(R^W(z_m), \mathcal{C}_\alpha)^2 \le \frac{4\eta[W_{\max} - W(z_0)]}{Md_-L_-\lambda_-} + 6\varepsilon^2.$$

This is a best-iterate guarantee at fixed η, not a last-iterate or singular-optimum guarantee. Radial clipping is absorbed in d_m (including $d_m = 0$), so stalled steps give no progress; nonradial feasibility maps and fallbacks must use their actual increments in the direct Taylor bound after Theorem C.1. The stopping certificate applies only to a successfully queried state; a cap or a zero clipped trade does not imply it.

Adaptive ratio queries. The inequalities hold pathwise given the realized errors. For a probabilistic version along an adaptive path, let $\mathcal{H}_{k,m}$ contain the information before query (k,m) and require, on the event that this query is reached,

$$\mathbb{P}(e_{k,m} > \varepsilon_{k,m} \mid \mathcal{H}_{k,m}) \le \delta_{k,m}.$$

For finitely many possible queries, the tower property and a union bound give simultaneous validity with probability at least $1 - \sum_{k,m} \delta_{k,m}$. Fresh continuation samples per query are one sufficient route; reusing one Monte Carlo sample across adaptive queries needs a separate uniform analysis.

Theorem C.4 (Finite-grid accounting with a single charged net order)**.** *Let a complete refinement policy π produce virtual paths $\widetilde{Z}_{k,0} = Z_k, \dots, \widetilde{Z}_{k,M_k}$ and net orders $A_k = \sum_{m<M_k} a_{k,m}$, executing only $Z_k^+ = \Gamma_\alpha(Z_k, A_k)$, followed by $Z_{k+1} = F_k^{0,\Delta}(Z_k^+, \xi_k)$. Here $F_k^{0,\Delta}$ includes the original-model domain treatment. Assume the conditional market law given all refinement queries and decisions, integrability, and a common terminal payoff $W_K = G_\Delta$. Define*

$$\begin{aligned}
\widetilde{G}_k^W &:= W_k(\widetilde{Z}_{k,M_k}) - W_k(Z_k),\\
E_k^W &:= W_k(\widetilde{Z}_{k,M_k}) - W_k(Z_k^+),\\
G_k^W &:= W_k(Z_k^+) - W_k(Z_k) = \widetilde{G}_k^W - E_k^W,\\
\mathcal{R}_k^{W,\Delta}(z') &:= W_k(z') - \mathbb{E}\big[W_{k+1}(F_k^{0,\Delta}(z', \xi_k))\big].
\end{aligned} \tag{74}$$

Then, without assuming infinite refinement or differentiability,

$$J_0^\Delta(\pi) - W_0(z) = \mathbb{E}^\pi \sum_{k=0}^{K-1} [\widetilde{G}_k^W - E_k^W - \mathcal{R}_k^{W,\Delta}(Z_k^+)]. \tag{75}$$

For $W_k = V_{\eta,k}^{\bar{u},\Delta}$ and common initial state, put $D_\eta^\Delta(\bar{u}) := J_0^\Delta(\bar{u}) - J_\eta^\Delta(\bar{u})$. Then

$$J_0^\Delta(\pi) - J_0^\Delta(\bar{u}) = \mathbb{E}^\pi \sum_{k=0}^{K-1} [\widetilde{G}_k^W - E_k^W - \mathcal{R}_k^{W,\Delta}(Z_k^+)] - D_\eta^\Delta(\bar{u}). \tag{76}$$

If all virtual steps satisfy Theorem C.3, let

$$B_k := \sum_{m<M_k} \frac{\omega_{k,m}}{\eta} \left(\frac{1}{2} r_{k,m}^2 - e_{k,m}^2\right).$$

Then

$$J_0^\Delta(\pi) - J_0^\Delta(\bar{u}) \ge \mathbb{E}^\pi \sum_{k=0}^{K-1} [B_k - E_k^W - \mathcal{R}_k^{W,\Delta}(Z_k^+)] - D_\eta^\Delta(\bar{u}). \tag{77}$$

A nonnegative right-hand side is sufficient for improvement on the specified grid. In particular, virtual ascent alone is insufficient; fee loss and the market/reference corrections must also be covered. When $0 \le W_{k,x} \le \overline{\lambda}_{x,k}$ along each fee segment, one may replace E_k^W on the right by $\overline{\lambda}_{x,k} \sum_i \alpha_i (A_k^i)^-$.

Proof. The conditional market law gives

$$\mathbb{E}^\pi[W_{k+1}(Z_{k+1}) - W_k(Z_k)] = \mathbb{E}^\pi[G_k^W - \mathcal{R}_k^{W,\Delta}(Z_k^+)].$$

The identity $G_k^W = \widetilde{G}_k^W - E_k^W$ follows by adding and subtracting the virtual endpoint value. Summing over k telescopes and $W_K = G_\Delta$ proves (75). Use $W_0(z) = J_\eta^\Delta(\bar{u})$ and subtract $D_\eta^\Delta(\bar{u})$ for (76). The virtual increments telescope to $\widetilde{G}_k^W \ge B_k$ under Theorem C.3; subtracting the exact final fee correction proves (77). The optional fee upper bound follows from (40). □

Interpretation. E_k^W is the loss from charging the *same net target* its actual sale fees, not a sum of hypothetical per-step fees; it is nonnegative under a positive cash marginal. The market-step residual includes the mismatch between the frozen regularized continuation and the original-model transition, and neither it nor the baseline correction has a favorable sign in general. The identity covers caps, stalls, and nonradial corrections through actual endpoints, whereas its Taylor lower bound covers only qualifying radial steps; simultaneous high-probability query accuracy does not alone imply an unconditional expected-objective guarantee, since failure and fallback losses must also be controlled. An infinity-norm ratio-error radius converts to at most $\sqrt{N}$ times that radius in Euclidean norm. If $W_K \neq G_\Delta$, append $\mathbb{E}^\pi[G_\Delta(Z_K) - W_K(Z_K)]$; with a random common initial state, take $\mathbb{E}W_0(Z_0)$.

C.3 Exact-ratio convergence at fixed physical time

Two stronger conclusions are available under additional deterministic conditions. First, let $W \in C^2$, $g = \mathcal{P}_\eta(W_y/W_x)$, and $\dot{z} = L(z)(-\mathbf{1}^\top g, g)$. If this flow exists for all artificial times, stays in a compact feasible set with Lipschitz vector field, and $LW_x \geq m_0 > 0$, then its wedge residual tends to zero. Indeed, the virtual-map identity gives

$$\frac{dW}{ds} = LW_x \left[\frac{1}{\eta} \operatorname{dist}(W_y/W_x, \mathcal{C}_\alpha)^2 + \sum_i \alpha_i g_i^- \right] \geq \frac{LW_x}{\eta} \operatorname{dist}(W_y/W_x, \mathcal{C}_\alpha)^2.$$

Bounded W makes the squared residual integrable; bounded velocity and ratio derivatives make it uniformly continuous, hence convergent to zero. Accumulation points lie in the wedge, but need not be unique or optimize accumulated trades.

Second, on a feasible one-asset buying ray $z(a) = (x_0 - a, y_0 + a)$, $0 \leq a \leq a_*$, put $F(a) = W(z(a))$. If $F'(a_*) = 0$, $-M \leq F'' \leq -m < 0$, $0 < \lambda_- \leq W_x \leq \lambda_+$, and $0 < L_- \leq L \leq L_+$, then for $a_0 \in [0, a_*]$,

$$a_{k+1} = a_k + \frac{dL(z(a_k))}{\eta W_x(z(a_k))} F'(a_k), \qquad 0 < d \leq \frac{\eta \lambda_-}{L_+ M},$$

is the exact projected purchase, because $F' = W_x(R^W - 1)$. Integrating the curvature bounds gives $m(a_* - a) \leq F'(a) \leq M(a_* - a)$, so with $\rho = dL_- m/(\eta \lambda_+)$ the update stays on the ray and satisfies

$$0 \leq a_* - a_k \leq (1-\rho)^k (a_* - a_0), \qquad F(a_*) - F(a_k) \leq \tfrac{M}{2}(1-\rho)^{2k}(a_* - a_0)^2.$$

These are fixed-η results for virtual trajectories; the charged endpoint must be analyzed separately, and no uniform singular-limit rate or multidimensional net-trade optimality follows.

D Singular limit and oracle refinement

D.1 Problem-level qualitative convergence of exact optimal values

This subsection concerns only the exact optimal values V_0 and V_η of the singular and η-regularized problems. The regularized HJB can be written formally as

$$V_{\eta,t} + \mathcal{L}^0 V_\eta + \frac{LV_{\eta,x}}{2\eta} \operatorname{dist}^2\left(\frac{V_{\eta,y}}{V_{\eta,x}}, \mathcal{C}_\alpha \right) = 0. \tag{78}$$

For a smooth test function ϕ, write

$$A[\phi] := \phi_t + \mathcal{L}^0 \phi, \qquad G_i^+[\phi] := \phi_{y_i} - \phi_x, \qquad G_i^-[\phi] := (1 - \alpha_i)\phi_x - \phi_{y_i},$$

and define the proper limiting operator

$$F[\phi] := \min\left\{ -A[\phi], \min_i(-G_i^+[\phi]), \min_i(-G_i^-[\phi]) \right\}. \tag{79}$$

Assumption D.1 (Penalized-value boundedness and limiting comparison)**.** *On every compact interior cylinder $\mathcal{D}' \Subset (0,T) \times \mathcal{S}^\circ$:*

(i) V_η is a continuous viscosity solution of (78) *with the common terminal/state-constraint data;*

(ii) $\{V_\eta\}_{\eta\in(0,1]}$ *is locally equibounded. On compact bands with* $x \geq \underline{x} > 0$*, this is verified for the benchmark: the zero-rate policy gives* $V_\eta(t,x,y) \geq U(\underline{x}e^{r(T-t)})$ *uniformly in* η*, while* $V_\eta \leq V_0 \leq 1/(\gamma-1)$*. On general interior cylinders that may contain negative cash, local equiboundedness remains an assumption;*

(iii) there is $\underline{\lambda}_{\mathcal{D}'} > 0$ *such that, for all sufficiently small positive cash increments* $h > 0$ *that remain in a slightly larger compact interior cylinder,*

$$V_\eta(t, x+h, \mathbf{y}) - V_\eta(t,x,\mathbf{y}) \geq \underline{\lambda}_{\mathcal{D}'} h.$$

For the constant-coefficient CRRA–GBM benchmark this item follows from Proposition E.6 wherever its finite-value and cash-lifting conditions have been verified;

(iv) the proper operator F *in* (79)*, with the inherited terminal and state-constraint data, satisfies comparison;*

(v) the upper and lower half-relaxed limits inherit those common terminal and state-constraint data in the viscosity sense, for example through uniform boundary barriers.

Theorem D.2 (Conditional local value convergence)**.** *Under Assumption D.1,*

$$V_\eta \longrightarrow V_0 \qquad \text{locally uniformly on } (0,T)\times\mathcal{S}^\circ.$$

Proof. Let $\overline{V}$ and $\underline{V}$ be the upper and lower half-relaxed limits of V_η as $(\eta, s, z') \to (0, t, z)$. Local equiboundedness makes them finite. The cash-increment inequality in item (iii), applied at nearby touching points, gives $\phi_x \geq \underline{\lambda}_{\mathcal{D}'}$ for the tests below.

If ϕ touches $\overline{V}$ from above and some $G_i^{\pm}[\phi] \geq 0$, then $F[\phi] \leq 0$ directly. Otherwise all gradient constraints are strict: nearby upper touching points have ratios inside $\mathcal{C}_\alpha$, so the penalty vanishes and the penalized subsolution inequality gives $A[\phi] \geq 0$. Again $F[\phi] \leq 0$.

For a test from below of $\underline{V}$, nearby lower touching points give

$$A[\phi] + \frac{L\phi_x}{2\eta}\operatorname{dist}^2(\phi_y/\phi_x, \mathcal{C}_\alpha) \leq o(1).$$

The penalty is nonnegative, so $A[\phi] \leq 0$ in the limit. A positive $G_i^{\pm}[\phi]$ would keep the ratio a fixed distance outside the wedge; local lower bounds on L and ϕ_x would then make the penalty at least c/η, a contradiction. Thus all $G_i^{\pm}[\phi] \leq 0$ and $F[\phi] \geq 0$.

Item (v) supplies the common boundary data. Comparison gives $\overline{V} \leq \underline{V}$, while the reverse inequality holds by definition. Both equal the unique singular value V_0. Failure of uniform convergence on a compact set would supply a sequence contradicting one of these relaxed-limit identities. □

Remark D.3 (Scope of the problem-level value limit). Theorem D.2 concerns exact optimal values only; it implies no convergence of trading rates, free boundaries, face labels, cumulative interventions, or the performance of $\mathcal{P}_\eta(\widehat{R}^{\Delta}_{\mathrm{BPTT}})$. Appendix D.2 gives a one-sided exact-hold consequence.

D.2 Concavity and exact hold recovery on strict interiors

Concavity transfers local value convergence to gradients; a strict ratio margin then gives exact hold recovery.

Lemma D.4 (State concavity of the exact values)**.** *Fix* $t < T$ *and a convex interior state domain. Suppose admissible controls for the regularized problem can be realized on a common filtered Brownian space, the admissible class is closed under convex mixing, and the positivity and integrability requirements are preserved by that mixing. Then* $z \mapsto V_\eta(t,z)$ *is concave for every* $\eta > 0$*. The singular value* $z \mapsto V_0(t,z)$ *is concave under the analogous convexity of the singular admissible class.*

Proof. Couple admissible dollar-rate controls A^1, A^2 from initial states z^1, z^2 on the same Brownian space, and mix both with weights $\theta, 1-\theta$. The linear risky-position equation gives $Y^\theta = \theta Y^1 + (1-\theta)Y^2$. On positive liquidation wealth,

$$F_\eta(x,y,a) = rx - c_\alpha(a) - \frac{\eta\|a\|^2}{2\ell(x,y)}, \qquad \ell(x,y) = x + \sum_i (1-\alpha_i)y_i,$$

is jointly concave: c_α is convex and $\|a\|^2/\ell$ is a convex perspective. Thus $\bar{X} = \theta X^1 + (1-\theta)X^2$ satisfies $d\bar{X} \le F_\eta(\bar{X}, Y^\theta, A^\theta)ds$. On localized positive-wealth bands, local Lipschitz continuity in x and scalar ODE comparison imply $X^\theta \ge \bar{X}$. Consequently $L^\theta \ge \theta L^1 + (1-\theta)L^2 > 0$; the stated admissibility and integrability remove localization. Increasing concave utility yields

$$\mathbb{E}U(L_T^\theta) \ge \theta \mathbb{E}U(L_T^1) + (1-\theta)\mathbb{E}U(L_T^2).$$

Choose ε-optimal controls and send $\varepsilon \downarrow 0$. For singular controls, convex mixing gives the exact mixed state and the same utility argument. □

Lemma D.5 (Uniform convergence of state supergradients)**.** *Let K be a compact subset of an open state-interior cylinder with a common state-neighborhood. Suppose $V_\eta(t,\cdot)$ is concave, $V_\eta \to V_0$ locally uniformly, and V_0 is differentiable in the state with $D_z V_0(t,z)$ jointly continuous in time and state on that neighborhood. With D_z^+ denoting the concave superdifferential,*

$$\sup_{(t,z)\in K} \sup_{p \in D_z^+ V_\eta(t,z)} \|p - D_z V_0(t,z)\| \longrightarrow 0. \tag{80}$$

Proof. Concavity and local uniform boundedness on a slightly larger compact cylinder give a uniform bound on all state supergradients over K by the standard secant-slope estimate. If (80) failed, there would be $\eta_n \downarrow 0$, $(t_n, z_n) \in K$, and $p_n \in D_z^+ V_{\eta_n}(t_n, z_n)$ staying a fixed distance from $D_z V_0(t_n, z_n)$. After passing to a subsequence, $(t_n, z_n, p_n) \to (t, z, p)$. The supergradient inequality and local uniform convergence imply, for every nearby state z' at the same limiting time,

$$V_0(t, z') \le V_0(t, z) + p^\top (z' - z),$$

so $p \in D_z^+ V_0(t,z)$. Differentiability gives $D_z^+ V_0(t,z) = \{D_z V_0(t,z)\}$, contradicting the fixed separation and the continuity of $D_z V_0$ on K. □

Corollary D.6 (Eventual exact hold recovery on compact strict interiors)**.** *Let $K \Subset (0,T) \times \mathcal{S}^\circ$ have a common state-neighborhood on which $V_\eta(t,\cdot)$ is concave and $V_\eta \to V_0$ locally uniformly. Suppose V_0 is differentiable in the state with jointly continuous state gradient on that neighborhood, $V_{0,x} \ge \underline{\lambda}_K > 0$, and, with $R^0 := V_{0,y}/V_{0,x}$,*

$$m_K := \inf_{(t,z)\in K} \min_{1\le i\le N} \min\{R_i^0(t,z) - (1-\alpha_i),\, 1 - R_i^0(t,z)\} > 0. \tag{81}$$

Then, for all sufficiently small η, every state-differentiability point $(t,z) \in K$ of V_η satisfies

$$1 - \alpha_i < \frac{V_{\eta,y_i}(t,z)}{V_{\eta,x}(t,z)} < 1 \quad \textit{for all } i, \qquad \mathcal{P}_\eta\left(\frac{V_{\eta,y}}{V_{\eta,x}}(t,z)\right) = 0. \tag{82}$$

If V_η has the additional regularity needed for classical HJB verification on a domain covering the candidate trajectories and its selector is admissible, the verified regularized optimal feedback is identically zero on K for all sufficiently small η.

Proof. See Appendix D.3. □

D.3 Proof of Corollary D.6

By Lemma D.5, every supergradient $p = (p_x, p_y)$ converges uniformly to $D_z V_0$ on K. Thus $p_x \ge \underline{\lambda}_K/2$ for small η; uniform continuity of $(p_x, p_y) \mapsto p_y/p_x$ on the resulting compact set gives

$$\sup_{(t,z)\in K} \sup_{p \in D_z^+ V_\eta(t,z)} \|p_y/p_x - R^0(t,z)\|_\infty \longrightarrow 0.$$

The margin in (81) puts these ratios strictly in the wedge. At differentiability points, take $p = D_z V_\eta$ and apply Proposition 1. State-C^1 regularity extends this to every point; separately assumed classical verification and admissibility make the selector optimal.

Remark D.7 (Regularity required for the benchmark). The C^1 assumption on V_0 does not follow from interiority: cash is nondiffusive, so the HJB is degenerate. For the homogeneous CRRA benchmark, one route is to reduce by scale to portfolio-ratio variables on compact sets away from the solvency boundary, risky-coordinate faces, the terminal time, and any loss of uniform ellipticity of the reduced covariance.

Remark D.8 (Why intervention-face convergence does not follow). If $R_i^0 = 1$ on a singular purchase face, both a sequence with $R_{\eta,i} < 1$ (hold) and one with $R_{\eta,i} > 1$ (buy) can converge to the same limiting ratio. The analogous ambiguity holds at $1 - \alpha_i$. Consequently, even uniform ratio convergence does not identify eventual buy/sell labels at the faces or imply Hausdorff convergence of state-space free boundaries. Such conclusions require quantitative gradient estimates and a nondegenerate crossing or boundary-layer analysis beyond the present result.

D.4 Evaluation models and the common comparison identity

Fix the recovery coefficient $\eta > 0$ and use $\kappa \geq 0$ for the evaluation dynamics; $V_\kappa^v = J_\kappa(v)$ for a complete normalized-rate feedback v. Rate feedbacks and singular strategies belong to different control classes, and the same feedback produces different dollar trades $L_s v(s, Z_s)$ under the two evaluation laws. $\pi^\star_{\eta,\widehat{R}}$ denotes a separately assumed admissible continuous-time refinement strategy; under R4 its exact-ratio specialization is $\pi^\star_\eta = \pi^\star_{\eta,R_\eta}$.

Common assumptions for the identities below. All values are finite at the initial pair (t, z); all compared strategies, including fallbacks and solvency management, are admissible for their evaluation dynamics; and the terminal reward is $U(L_T)$. Each comparison function W is $C^{1,2}$ on a domain covering the relevant trajectories and trade segments, with $W_x > 0$ whenever a ratio is used; stochastic integrals in Itô's formula have zero expectation, and drift integrals and jump sums are integrable (localization is allowed if its exit continuation justifies the whole-horizon limit). Reference feedbacks are fixed Markov feedbacks, possibly after conditioning on their construction, and candidates may use nonanticipative auxiliary randomness. These are assumptions on the continuous-time controller, not consequences of an Euler simulation. With h_κ as defined in Appendix B.3,

$$\mathcal{G}_\kappa(u; DW) = LW_x h_\kappa(u; R^W), \qquad R^W = W_y/W_x, \tag{83}$$

and $\mathcal{L}^{\kappa,u} W = \mathcal{L}^0 W + \mathcal{G}_\kappa(u; DW)$, where $\mathcal{L}^0$ is the no-trade generator. Hence

$$\mathcal{G}_\kappa(u; DW) = \mathcal{G}_\eta(u; DW) + \frac{\eta - \kappa}{2} LW_x \|u\|^2. \tag{84}$$

D.5 One-shot improvement across the two evaluation dynamics

Theorem D.9 (Cross-model performance difference and projected improvement)**.** *Let v be a reference feedback and $W = V_\eta^v$ satisfy the common assumptions and its fixed-policy PDE. Fix $0 \leq \kappa \leq \eta$. For any admissible candidate w, evaluated under coefficient κ,*

$$J_\kappa(w) - J_\eta(v) = \mathbb{E}^{\kappa,w} \int_t^T LW_x \left[h_\eta(w; R^W) - h_\eta(v; R^W) + \frac{\eta - \kappa}{2} \|w\|^2 \right] ds. \tag{85}$$

Every integrand is evaluated at $(s, Z_s^{\kappa,w})$. Put $g = \mathcal{P}_\eta(R^W)$, and suppose the projected candidates are complete and admissible. The exact update satisfies

$$J_\kappa(g) - J_\eta(v) \geq \mathbb{E}^{\kappa,g} \int_t^T LW_x \left[\frac{\eta}{2} \|g - v\|^2 + \frac{\eta - \kappa}{2} \|g\|^2 \right] ds \geq 0. \tag{86}$$

For $w = \mathcal{P}_\eta(\widehat{R})$, with $e = \|\widehat{R} - R^W\|$,

$$J_\kappa(w) - J_\eta(v) \geq \mathbb{E}^{\kappa,w} \int_t^T LW_x \left[\frac{\eta}{2} \|g - v\|^2 - \frac{e^2}{2\eta} + \frac{\eta - \kappa}{2} \|w\|^2 \right] ds. \tag{87}$$

In particular, exact projection improves on $J_\eta(v)$ in either evaluation model. For comparison with the reference in the original model, the same right-hand side must additionally subtract $D_\eta(v) := J_0(v) - J_\eta(v)$.

Proof. The reference PDE is $W_t + \mathcal{L}^0 W + \mathcal{G}_\eta(v; DW) = 0$. Apply Itô's formula to $W(s, Z_s^{\kappa,w})$. Its expected change is $J_\kappa(w) - J_\eta(v)$. The reference PDE reduces the drift to

$$\mathcal{G}_\kappa(w; DW) - \mathcal{G}_\eta(v; DW),$$

which gives (85) by (84). Strong concavity gives $h_\eta(g; R^W) - h_\eta(v; R^W) \geq (\eta/2)\|g - v\|^2$. For the estimated maximizer w, the same calculation as the completed square in Appendix B.6 gives

$$0 \leq h_\eta(g; R^W) - h_\eta(w; R^W) \leq -\frac{\eta}{2}\|g - w\|^2 + (R^W - \widehat{R})^\top (g - w) \leq \frac{e^2}{2\eta}.$$

Subtracting this loss proves both gain bounds. Finally, $J_0(w) - J_0(v) = [J_0(w) - J_\eta(v)] - D_\eta(v)$, proving the original-baseline assertion without identifying the two occupancies. □

Meaning. The theorem guarantees improvement relative to the reference's *regularized* value; comparing $J_0(g)$ with $J_0(v)$ or with $J_\eta(g)$ are separate questions, and every expectation follows the candidate in its own evaluation model.

Corollary D.10 (When removing leakage improves the same feedback)**.** *For a fixed feedback v, assume its own regularized value V_η^v satisfies the PDE, regularity, and Itô conditions above. The identity*

$$J_0(v) - J_\eta(v) = \mathbb{E}^{0,v} \int_t^T \frac{\eta}{2} L(V_\eta^v)_x \|v\|^2 ds \tag{88}$$

holds even without a positive cash derivative. If $(V_\eta^v)_x \geq 0$ on the original-model trajectory, the difference is nonnegative. If, additionally, $0 \leq L(V_\eta^v)_x \leq M_v$ there and $\mathbb{E}^{0,v} \int_t^T \|v\|^2 ds \leq E_v$, then $0 \leq D_\eta(v) \leq \eta M_v E_v/2$. Consequently, the original-model gain bound in (87)*, with $\kappa = 0$, exceeding $\eta M_v E_v/2$ is sufficient for $J_0(w) \geq J_0(v)$.*

Proof. Apply the generator argument in the preceding proof with $w = v$ and $\kappa = 0$. The two η-regularized local criteria cancel. This argument uses derivatives directly and does not divide by $(V_\eta^v)_x$. The sign and upper bound follow by integration; subtract that upper bound in the original-baseline identity to obtain the sufficient condition. □

Counterexample D.11 (Removing leakage can hurt a nonoptimal feedback). Take $N = 1$, $\alpha = 0.1$, $r = \mu = 0$, $(x, y) = (1, 0)$, $\eta = 1$, and $T = 0.7$, initially with zero volatility. Choose smooth $\chi, \phi \in [0, 1]$ with $\chi = 0$ for $x \leq 0.1$, $\chi = 1$ for $x \geq 0.2$, $\phi = 0$ for $q \leq 0.96$, and $\phi = 1$ for $q \geq 0.99$. Use the same feedback in both models:

$$v(s, x, y) = \chi(x) \begin{cases} 1, & s \leq 0.1, \\ \phi(x + y), & 0.1 < s \leq 0.7. \end{cases}$$

While $v = 1$, $\dot{L} = -(0.1 + \kappa/2)L$ and $d(x + y)/ds = -\kappa L/2$. In the original model $x + y = 1$, so buying continues throughout, giving $L_T^0 = e^{-0.07}$ and $x_T^0 = 10e^{-0.07} - 9 > 0.32$. Under $\kappa = 1$, the first phase gives $L_{0.1}^1 = e^{-0.06}$ and $(x + y)_{0.1}^1 = (1 + 5e^{-0.06})/6 < 0.96$. Trading then stops with cash above 0.85, so $L_T^1 = e^{-0.06} > L_T^0$ and $J_0(v) < J_1(v)$.

The inequality persists for sufficiently small positive volatility: the gates ensure $x \geq 0.1$ and nonnegative buying ensures $y \geq 0$, so $L \geq 0.1$. Local Lipschitz coefficients with uniform linear growth give convergence in probability to the deterministic path as volatility vanishes, by localization, the squared-difference estimate, and Gronwall. Uniform moment bounds remove localization; bounded continuity of CRRA utility on $L \geq 0.1$ gives payoff convergence. This is therefore also a nondegenerate benchmark example. Optimal-value monotonicity does not supply the fixed-policy cash monotonicity required in Corollary D.10.

The failure need not vanish uniformly over feedback families. At fixed $T = 0.7$ and fixed positive volatility, hold cash until $T - 0.7\eta$, then run the two phases at rate $1/\eta$ for durations 0.1η and 0.6η. The time change $s = T - 0.7\eta + \eta r$ recovers the preceding drift systems with volatility $\sigma\sqrt{\eta}$. Payoff continuity gives

$$\lim_{\eta \downarrow 0} [J_0(v_\eta) - J_\eta(v_\eta)] = U(e^{-0.07}) - U(e^{-0.06}) < 0. \tag{89}$$

Neither rates nor integrated squared rates are uniformly bounded here; this is separate from convergence of optimal problem values.

D.6 From a regularized policy to a bound for the original objective

Corollary D.12 (Optimal-reference specialization and ratio-error bound)**.** *Assume the admissibility embedding in Appendix E.1, so that $b_\eta(t,z) := V_0(t,z) - V_\eta(t,z) \geq 0$. Suppose the exact regularized optimal value V_η satisfies the common classical and integrability assumptions and the HJB equation. Define*

$$R_\eta := \frac{\nabla_y V_\eta}{\partial_x V_\eta},$$

and suppose $\mathcal{P}_\eta(R_\eta)$ is an admissible optimal feedback for the regularized problem. Let $w = \mathcal{P}_\eta(\widehat{R})$ be a complete feedback, admissible in every model in which it is evaluated. Define

$$\begin{aligned} a_\eta(w;s,z) &:= \mathcal{G}_\eta(\mathcal{P}_\eta(R_\eta); DV_\eta) - \mathcal{G}_\eta(w; DV_\eta), \\ e_\eta(s,z) &:= \|\widehat{R}(s,z) - R_\eta(s,z)\|. \end{aligned} \tag{90}$$

Then $0 \leq a_\eta(w) \leq LV_{\eta,x} e_\eta^2/(2\eta)$ and

$$V_\eta - J_\eta(w) = \mathbb{E}^{\eta,w} \int_t^T a_\eta(w; s, Z_s)\, ds, \tag{91}$$

$$V_0 - J_0(w) = b_\eta + \mathbb{E}^{0,w} \int_t^T \left[a_\eta(w; s, Z_s) - \frac{\eta}{2} LV_{\eta,x} \|w\|^2 \right] ds. \tag{92}$$

In particular, if $LV_{\eta,x} \leq M_\eta$ on the relevant trajectories,

$$0 \leq V_\eta - J_\eta(w) \leq \frac{M_\eta}{2\eta} \mathbb{E}^{\eta,w} \int_t^T e_\eta(s, Z_s)^2 ds, \tag{93}$$

$$0 \leq V_0 - J_0(w) \leq b_\eta + \frac{M_\eta}{2\eta} \mathbb{E}^{0,w} \int_t^T e_\eta(s, Z_s)^2 ds. \tag{94}$$

For the exact oracle $w = \mathcal{P}_\eta(R_\eta)$, also admissible in the original model,

$$J_\eta(\mathcal{P}_\eta(R_\eta)) = V_\eta \leq J_0(\mathcal{P}_\eta(R_\eta)) \leq V_0, \qquad 0 \leq V_0 - J_0(\mathcal{P}_\eta(R_\eta)) \leq b_\eta. \tag{95}$$

Thus optimal regularized recovery loses at most the problem-level regularization gap when deployed as a rate in the original problem.

Proof. In (85), take the optimal reference $v = \mathcal{P}_\eta(R_\eta)$, so $W = V_\eta$, first with $\kappa = \eta$ and then with $\kappa = 0$. Negation and addition of b_η give (91)–(92). Optimality and the completed-square estimate in the preceding theorem give $0 \leq a_\eta(w) \leq LV_{\eta,x} e_\eta^2/(2\eta)$. Integration and discarding the nonpositive leakage term give both upper bounds; admissibility in the relevant optimal control class gives their lower bounds. For $w = \mathcal{P}_\eta(R_\eta)$, $a_\eta = 0$, and the same original-model identity yields (95). □

The two costs are problem distortion b_η and optimal-ratio error: if $e_\eta \leq \bar{e}_\eta$ on the relevant occupancies, the latter is at most $M_\eta(T-t)\bar{e}_\eta^2/(2\eta)$, so for bounded M_η value accuracy needs $\bar{e}_\eta = o(\sqrt{\eta})$, whereas rate accuracy needs $\bar{e}_\eta = o(\eta)$. Theorem D.21 bounds b_η under R1–R3 but supplies no learned-ratio bound, since $e_\eta \leq \|\widehat{R} - R^{V_\eta^v}\| + \|R^{V_\eta^v} - R_\eta\|$ and sampling and discretization address only part of the first term. If B_η is the upper bound in (94), comparison with an arbitrary original baseline satisfies $J_0(w) - J_0(v_0) \geq [V_0 - J_0(v_0)] - B_\eta$; no unconditional trained-policy ranking follows.

D.7 Continuous-time FR certificates

The following identities concern an admissible original-model singular strategy, including continuous finite-variation regulation. Its existence is assumed separately from the finite grid implementation.

Theorem D.13 (Continuous-time FR performance and singular policy gap)**.** *Let $\pi = (B, S)$ be an admissible original-model finite-variation strategy, with continuous parts B^c, S^c and jumps. At each intervention time τ_j, the entire virtual search, if any, is cost-free, and only its final net order is executed as one original-model jump from $Z_{\tau_j -}$ to Z_{τ_j}. Allow a countable collection of intervention times, provided total variation and the absolute jump-value sum satisfy the common integrability*

assumptions. Include an initial intervention, if present, with pre-intervention state z. Let W satisfy the common classical assumptions and $W(T,z) = U(L(z))$. Define

$$\begin{aligned} q_W &:= -(W_t + \mathcal{L}^0 W), \\ G_j^W &:= W(\tau_j, Z_{\tau_j}) - W(\tau_j, Z_{\tau_j -}), \\ I_W(\pi) &:= \sum_i \int_t^T (W_{y_i} - W_x)(s, Z_{s-})\, dB_s^{c,i} \\ &\quad + \sum_i \int_t^T ((1-\alpha_i)W_x - W_{y_i})(s, Z_{s-})\, dS_s^{c,i} + \sum_j G_j^W. \end{aligned} \tag{96}$$

Then the exact continuous-time identity is

$$J_0(\pi) - W(t,z) = \mathbb{E}^\pi \left[I_W(\pi) - \int_t^T q_W(s, Z_s)\, ds \right]. \tag{97}$$

Suppose $W = V_\eta$ satisfies the classical optimal HJB and the admissibility embedding of Corollary D.12. Writing $R_\eta = V_{\eta,y}/V_{\eta,x}$, its residual is

$$q_\eta = \frac{LV_{\eta,x}}{2\eta} \operatorname{dist}(R_\eta, \mathcal{C}_\alpha)^2 \geq 0. \tag{98}$$

Define the expected negative intervention contribution by

$$\begin{aligned} N_\eta(\pi) := \mathbb{E}^\pi \Bigg[&\sum_i \int_t^T (V_{\eta,x} - V_{\eta,y_i})_+\, dB_s^{c,i} + \sum_i \int_t^T (V_{\eta,y_i} - (1-\alpha_i)V_{\eta,x})_+\, dS_s^{c,i} \\ &+ \sum_j (-G_j^{V_\eta})_+ \Bigg], \end{aligned} \tag{99}$$

where continuous integrands are evaluated at (s, Z_{s-}). Also put

$$\mathcal{R}_\eta(\pi) := \mathbb{E}^\pi \int_t^T \operatorname{dist}(R_\eta(s, Z_s), \mathcal{C}_\alpha)^2\, ds.$$

If $0 < LV_{\eta,x} \leq M_\eta$ on the trajectories, then

$$V_0 - J_0(\pi) = b_\eta + \mathbb{E}^\pi \left[\int_t^T q_\eta(s, Z_s) ds - I_{V_\eta}(\pi) \right], \tag{100}$$

$$0 \leq V_0 - J_0(\pi) \leq b_\eta + N_\eta(\pi) + \frac{M_\eta}{2\eta} \mathcal{R}_\eta(\pi). \tag{101}$$

If the continuous-time trajectory satisfies $\operatorname{dist}(R_\eta(s, Z_s), \mathcal{C}_\alpha) \leq \varepsilon_\eta$ for $ds \otimes \mathbb{P}^\pi$-almost every (s, ω), then

$$0 \leq V_0 - J_0(\pi) \leq b_\eta + N_\eta(\pi) + \frac{M_\eta (T-t)}{2\eta} \varepsilon_\eta^2. \tag{102}$$

In particular, exact wedge tracking almost everywhere in physical time, nonnegative continuous intervention contributions, and nonnegative jump increments imply

$$V_\eta \leq J_0(\pi) \leq V_0, \qquad V_0 - J_0(\pi) \leq b_\eta. \tag{103}$$

Proof. Semimartingale Itô gives market drift $-q_W$, the displayed Stieltjes terms for continuous trades, and the full value change of each executed net jump. Virtual FR steps are not market interventions and therefore contribute no additional Stieltjes or jump terms. The common integrability conditions justify expectation, countable jump sums, and terminal limits, giving (97). The regularized HJB gives $q_W = q_\eta$ for $W = V_\eta$; adding b_η proves (100). Since $-\mathbb{E} I_{V_\eta} \leq N_\eta$ and $\mathbb{E} \int q_\eta ds \leq M_\eta \mathcal{R}_\eta/(2\eta)$, the gap bound follows. A uniform residual gives (102); exact tracking and nonnegative interventions give (103). Original admissibility supplies all nonnegative-gap assertions. □

The three contributions are problem distortion, adverse interventions, and wedge violation along the physical-time trajectory; endpoint wedge checks do not control the last. Theorem C.1 supplies only a virtual gain, so nonnegative *executed* jump increments must be checked after the fee correction. Purchases supported on $R_{\eta,i} = 1$ and sales on $R_{\eta,i} = 1 - \alpha_i$ contribute zero by complementarity; such boundary regulation uses normal-cone directions, not the Yosida rate, which vanishes on the faces. For an estimated-ratio strategy with $\operatorname{dist}(\widehat{R}, \mathcal{C}_\alpha) \le \widehat{\varepsilon}_\eta$ and $\|\widehat{R} - R_\eta\| \le \bar{e}_\eta$ almost everywhere in physical time,

$$\varepsilon_\eta \le \bar{e}_\eta + \widehat{\varepsilon}_\eta. \tag{104}$$

Failed queries, caps, and feasibility adjustments enter through the actual strategy, and exact ratio access does not ensure tracking or admissibility. If b_η, N_η, and $M_\eta \mathcal{R}_\eta/\eta$ all vanish, then $J_0(\pi_\eta) \to V_0$; this is a conditional continuous-time conclusion, with no market-grid limit asserted.

D.8 Continuous-time refinement compared with one-shot deployment in the original model

Corollary D.14 (Continuous-time comparison with the one-shot policy and the optimal-ratio case)**.** *Let v be a complete original-model rate feedback and $W = V_0^v$ satisfy its classical fixed-policy PDE and the common assumptions. For a continuous-time singular strategy π as above,*

$$J_0(\pi) - J_0(v) = \mathbb{E}^\pi \left[I_W(\pi) - \int_t^T \mathcal{G}_0(v; DW)(s, Z_s)\, ds \right]. \tag{105}$$

Thus an expected lower bound on I_W exceeding an expected upper bound on the displayed generator contribution certifies improvement. For the exact η-optimal feedback $\mathcal{P}_\eta(R_\eta)$, assume the hypotheses of Corollary D.12 in both evaluation models. Define

$$\begin{aligned} D_\eta &:= J_0(\mathcal{P}_\eta(R_\eta)) - V_\eta = \mathbb{E}^{0,\mathcal{P}_\eta(R_\eta)} \int_t^T \frac{\eta}{2} L V_{\eta,x} \|\mathcal{P}_\eta(R_\eta)\|^2 ds \ge 0, \\ A_\eta(\pi) &:= \mathbb{E}^\pi \left[I_{V_\eta}(\pi) - \int_t^T q_\eta(s, Z_s) ds \right]. \end{aligned} \tag{106}$$

Then

$$J_0(\pi) - J_0(\mathcal{P}_\eta(R_\eta)) = A_\eta(\pi) - D_\eta. \tag{107}$$

Under exact continuous-time wedge tracking this reduces to $J_0(\pi) - J_0(\mathcal{P}_\eta(R_\eta)) = \mathbb{E}^\pi I_{V_\eta}(\pi) - D_\eta$. Consequently, nonnegative FR intervention gains alone do not establish superiority over original-model one-shot deployment: they must cover the latter's leakage-removal benefit D_η.

Proof. For $W = V_0^v$, the fixed-policy PDE gives $-(W_t + \mathcal{L}^0 W) = \mathcal{G}_0(v; DW)$. Substitution into (97) proves the first identity. Corollary D.12 gives the displayed expression for D_η, and (97) with $W = V_\eta$ gives $A_\eta(\pi) = J_0(\pi) - V_\eta$. Subtracting yields (107). Exact tracking makes $q_\eta = 0$ almost everywhere along π, giving the specialization. □

The expectations use different policy occupancies in the same exact market; these identities neither construct a continuous regulator nor rank it universally above the one-shot policy. The finite overshoot example used a no-trade reference, not the optimal regularized value.

D.9 Derived CRRA rates on invariant comparison domains

This section proves quantitative bounds for the same original optimized values used in the paper. Its additional geometric and verification assumptions are explicit. The argument uses no restriction to future trading dates and no rate of convergence of the regularized boundary.

D.9.1 The three values and the model being compared

Fix $N < \infty$, a horizon T, and an initial time $t < T$. There are N correlated risky assets and a cash account. Let $q_i = 1 - \alpha_i$, $0 < \alpha_i < 1$, and write

$$L = x + q^\top y > 0, \qquad \theta_i = q_i y_i / L, \qquad \theta \in \mathbb{R}^N_+, \qquad X = L(1 - \mathbf{1}^\top \theta).$$

The optimized admissible class is exactly the original long-only solvency class; cash may be negative and no borrowing restriction is added to V_0 or V_η. Fix a compact comparison domain $\mathcal{K}_\theta \subset \mathbb{R}^N_+$; R3 requires only the auxiliary comparison policies and frozen-time flows to remain in it, so it constrains the competitors built in the proofs, not the original supremum. Bounds are uniform for initial ratios in $\mathcal{K}_\theta$; derivative estimates and strict-intervention lower bounds need an initial ratio in its interior, and constants may depend on $\mathcal{K}_\theta$. Let $\Sigma = aa^\top$ be a fixed positive definite return covariance with arbitrary correlations. The original dynamics are

$$dX_s = rX_s\,ds - \sum_i dB_{i,s} + \sum_i q_i\,dS_{i,s}, \tag{108}$$

$$dY_{i,s} = \mu_i Y_{i,s}\,ds + Y_{i,s}\sum_j a_{ij}\,dW_{j,s} + dB_{i,s} - dS_{i,s}. \tag{109}$$

Here B, S are nondecreasing, adapted finite-variation dollar trades; jumps and continuous singular trades are allowed. The terminal reward is

$$U(L) = \frac{L^k - 1}{k}, \qquad k = 1 - \gamma < 0, \quad \gamma > 1.$$

The regularized problem restricts trades to dollar rates Lu_i and subtracts $\eta L\|u\|_2^2/2$ from the cash drift. A positive u_i buys, a negative u_i sells. Write V_0, V_η for the optimal values, with the same utility, horizon, proportional costs and state constraints. For the *exact optimal regularized value*, put

$$R_{\eta,i} = \frac{V_{\eta,y_i}}{V_{\eta,x}}, \qquad C = \prod_{i=1}^N [q_i, 1], \qquad \mathcal{P}_\eta(R_\eta) = \frac{R_\eta - \mathrm{Proj}_C\, R_\eta}{\eta}.$$

The projection is coordinatewise, while covariance enters through R_η. $J_0(\mathcal{P}_\eta(R_\eta))$ is the original-model payoff of continuously reevaluating the exact regularized Yosida feedback, and $\pi^\star_\eta$ is the separately certified exact-ratio strategy of R4, whose single charged net order and continuation must satisfy all R4 conditions. Define

$$e_{\mathrm{reg}} = V_0 - V_\eta, \qquad e_{\mathrm{one}} = V_0 - J_0(\mathcal{P}_\eta(R_\eta)), \qquad e_\star = V_0 - J_0(\pi^\star_\eta). \tag{110}$$

There is no waiting for a trading date in any of these definitions. Fixed-time intervention paths below are parametrized by an artificial path length, during which physical time and Brownian motion do not advance.

D.9.2 Explicit structural hypotheses

All constants below may depend on N, the market and the fixed time interval, but not on η or the comparison speed M.

Assumption D.15 (R1: verification, homogeneity and original geometry)**.** *The original value has the verification regularity, including the generalized Itô formula at free boundaries, and $V_0 + 1/k$ is homogeneous of degree k. For $s < T$, its no-trade set in ratio coordinates is a closed set $\mathcal{N}_0(s) \subset \mathcal{K}_\theta$. At T we use the common terminal payoff and the verified terminal limit, without imposing this compact-set description of the terminal no-trade region. Put*

$$d(s,\theta) = \mathrm{dist}(\theta, \mathcal{N}_0(s)), \qquad \mathcal{O}_s = \mathcal{K}_\theta \setminus \mathcal{N}_0(s), \qquad Q_0 = -(V_{0,t} + \mathcal{L}^0 V_0),$$

where $\mathcal{L}^0$ is the no-trade diffusion generator in (x, y). Assume

$$0 \le Q_0(s, L, \theta) \le C_N L^k d(s,\theta), \qquad 0 < LV_{0,x} \le C_N L^k. \tag{111}$$

The original gradient constraints hold:

$$V_{0,y_i} - V_{0,x} \le 0, \qquad q_i V_{0,x} - V_{0,y_i} \le 0.$$

The derivatives needed below are continuous on interior compact sets. The bounds in this section are uniform if the initial time is moved forward within the stated interval. Test functions are defined on a neighborhood of the comparison domain, with the stated generalized Itô interpretation at its boundaries.

Assumption D.16 (R2: regularized verification and admissibility)**.** *For small $\eta > 0$, V_η is differentiable and concave in (x, y), with $V_{\eta,x} > 0$. Its classical HJB, and verification under original-model deployment of $\mathcal{P}_\eta(R_\eta)$, hold on the states used. This feedback is admissible. We assume the standard embedding comparison $V_\eta \le V_0$: using the same adapted dollar trades without leakage preserves risky holdings and increases cash. Verification formulas below include the required terminal integrability, and unconstrained local-maximization formulas are used only at interior states.*

Assumption D.17 (R3: an active, feasible inward field and two geometric certificates)**.** *There is a measurable field $v = (b, s)$, depending on (t, θ), with nonnegative coordinates, $b_i s_i = 0$, and*

$$\|b\|_1 + \|s\|_1 = 1 \quad \text{on } \mathcal{O}_t, \qquad b = s = 0 \quad \text{on } \mathcal{N}_0(t).$$

Only binding original trade directions are used:

$$\sum_i [(V_{0,y_i} - V_{0,x})b_i + (q_i V_{0,x} - V_{0,y_i})s_i] = 0. \tag{112}$$

Define its normalized wealth loss and ratio velocity by

$$c_v = \alpha^\top b, \qquad a_v = \operatorname{diag}(q)(b - s) + \theta c_v, \qquad \mathcal{T}_v\varphi = a_v \cdot D\varphi - k c_v \varphi. \tag{113}$$

The bounded-speed policies $u^M = M(b - s)$ in the regularized problem have admissible full-horizon realizations for all large M with $\eta M^2 \le 1$. The frozen-time pure-trade flow with ratio velocity a_v also has a viable, absolutely continuous realization up to its arrival at $\mathcal{N}_0(t)$. These realizations must stay in $\mathcal{K}_\theta$ and preserve the original holding and positive-liquidation constraints; no cash-sign constraint is imposed. This invariance is an assumption: compactness alone does not give it, a boundary with outward normal diffusion cannot be enforced by a bounded drift, and stopping at an artificial compact boundary without an exit-error estimate would not suffice. There are functions ψ, Φ to which the generalized Itô and chain rules apply, and positive constants, such that

$$1 \le \psi \le C_N, \qquad c_N d^2 \le \Phi \le C_N d^2, \tag{114}$$

$$\mathcal{T}_v\psi \le -c_N \psi \mathbf{1}_{\mathcal{O}_s}, \qquad |\mathcal{G}_N\psi| + |\mathcal{P}\psi| \le C_N\psi, \qquad |\psi_s| \le C_N a(s)\psi, \tag{115}$$

$$\mathcal{T}_v\Phi \le -c_N d, \qquad |\mathcal{G}_N\Phi| + |\mathcal{P}\Phi| \le C_N \mathbf{1}_{\mathcal{O}_s}, \qquad |\Phi_s| \le C_N a(s) d. \tag{116}$$

The spatial operators $\mathcal{G}_N, \mathcal{P}$ are explicitly derived in Lemma D.19. The inequalities involving second derivatives hold in the Itô sense, with no unaccounted singular second-derivative term. Finally,

$$\int_t^T a(s)ds < \infty, \qquad \int_t^T a(s)\mathbf{1}_{\{a(s)>m\}}ds \le C_N/m \quad (m \ge 1). \tag{117}$$

Bounded velocities, or a bound $a(s) \le C_N(1 + (T - s)^{-1/2})$, suffice. Section D.9.6 gives concrete constructions of these certificates. We do not assume an error exponent, a convergence rate of a free boundary, or an occupation estimate.

Assumption D.18 (R4: charged-endpoint and continuation certificate)**.** *The exact-ratio virtual search, if used, produces a net increment A_η whose* single *execution $\widehat{z} = \Gamma_\alpha(z, A_\eta)$ is feasible in the original market. Let $\pi^\star_\eta$ be an admissible original-model strategy starting with precisely this charged endpoint and a verified continuous-time continuation. Require $R_\eta(t, \widehat{z}) \in C$ and, along the subsequent continuation, $R_\eta \in C$ for Lebesgue-almost every physical time and all continuous and* executed-net-jump *intervention contributions to V_η to be nonnegative. A complementary regulator that buys only at $R_{\eta,i} = 1$ and sells only at $R_{\eta,i} = q_i$ has zero continuous contributions, but its existence and viability are hypotheses. The virtual endpoint's wedge membership alone does not imply charged-endpoint wedge membership, and a zero finite-rate feedback does not construct a singular regulator. The stronger estimate additionally requires the initial* net *order $(B, S) = (A^+_\eta, A^-_\eta)$ to satisfy the accumulated-trade certificate of Section D.9.5; exact ratio access does not supply that certificate automatically.*

D.9.3 The correlated generator and occupation estimates

Lemma D.19 (The full normalized generator)**.** *Set $\beta = \mu - r\mathbf{1}$ and $H(\theta) = \operatorname{diag}(\theta) - \theta\theta^\top$. Along the comparison policy $u^M = M(b - s)$, the drift of $L^k\varphi(s, \theta)$ divided by L^k equals*

$$\mathcal{A}_{\eta,M}\varphi = \varphi_s + \mathcal{G}_N\varphi + M\mathcal{T}_v\varphi + \frac{\eta M^2}{2}\|b - s\|_2^2 \mathcal{P}\varphi, \tag{118}$$

where

$$\mathcal{G}_N\varphi = \tfrac{1}{2}\operatorname{tr}(H\Sigma H D^2\varphi) + H(\beta - \gamma\Sigma\theta)\cdot D\varphi + k\left(r + \beta^\top\theta - \tfrac{\gamma}{2}\theta^\top\Sigma\theta\right)\varphi, \qquad \mathcal{P}\varphi = \theta\cdot D\varphi - k\varphi. \tag{119}$$

Proof. The liquidation dynamics divided by L have drift

$$r + \beta^\top\theta - Mc_v - \tfrac{1}{2}\eta M^2\|b - s\|_2^2$$

and diffusion row $\theta^\top a$. The ratio diffusion matrix is Ha; its no-trade drift is $H(\beta - \Sigma\theta)$. Trades add Ma_v, and leakage adds $\eta M^2\|b - s\|_2^2\theta/2$ to that drift. In the product rule for $L^k\varphi$, the covariance term adds $kH\Sigma\theta\cdot D\varphi$. Since $k - 1 = -\gamma$, the combined ratio drift is $H(\beta - \gamma\Sigma\theta)$. The L^k drift contributes the final term of $\mathcal{G}_N$, $-kMc_v\varphi$, and the leakage term $-k\eta M^2\|b - s\|_2^2\varphi/2$. This gives all terms in (118). In particular, no off-diagonal covariance term has been dropped. □

Lemma D.20 (Derived occupation bounds)**.** *Under R3, let $Z_s = L_s^k$ along the regularized comparison policy. Uniformly for all large M with $\eta M^2 \le 1$,*

$$\sup_{s\in[t,T]} \mathbb{E}Z_s \le C_N Z_t, \tag{120}$$

$$\mathbb{E}\int_t^T Z_s \mathbf{1}_{\mathcal{O}_s}(\theta_s)ds \le C_N Z_t/M, \tag{121}$$

$$\mathbb{E}\int_t^T Z_s d(s,\theta_s)ds \le C_N Z_t\left(\frac{d(t,\theta_t)^2}{M} + \frac{1}{M^2}\right). \tag{122}$$

Proof. Because $\|b - s\|_2^2 \le \mathbf{1}_{\mathcal{O}_s}$, (115) and (118) give

$$\mathcal{A}_{\eta,M}\psi \le C_N(1 + a(s))\psi - c_N M\psi\mathbf{1}_{\mathcal{O}_s}.$$

Apply Itô's formula with the deterministic factor $\exp(-\int_t^s C_N(1 + a(v))dv)$. This factor and ψ are bounded above and away from zero. The nonnegative terminal term then gives both (120) and (121). For each fixed M, η, the relative drift and diffusion of L are bounded on $\mathcal{K}_\theta$, so its positive and negative moments of every fixed order are finite. This justifies removing localization; uniform-in-M integrability need not be assumed to obtain these estimates. Similarly, (116) implies

$$\mathcal{A}_{\eta,M}\Phi \le -c_N Md + C_N\mathbf{1}_{\mathcal{O}_s} + C_N a(s)d.$$

On times with $a(s) \le c_N M/(2C_N)$, absorb the last term into $-c_N Md/2$. On the remaining times, boundedness of d on $\mathcal{K}_\theta$, (120), and (117) bound its integrated expected contribution by $C_N Z_t/M$. Use (121) for the other positive term. The terminal value $Z_T\Phi(T,\theta_T)$ is nonnegative, and the initial value is at most $C_N Z_t d(t,\theta_t)^2$. The resulting inequality is

$$\tfrac{1}{2}c_N M\mathbb{E}\int_t^T Z_s d(s,\theta_s)ds \le C_N Z_t d(t,\theta_t)^2 + C_N Z_t/M,$$

which proves (122). If coefficients are specified on $[t,T)$, take the verified terminal limit. No market-time mesh is involved. □

D.9.4 Value distortion and original-model one-shot deployment

Theorem D.21 (The master regularization bound)**.** *Under R1–R3, for every initial state whose ratio lies in $\mathcal{K}_\theta$ and every sufficiently large M with $\eta M^2 \le 1$,*

$$0 \le V_0 - V_\eta \le C_N L^k\left(\frac{d(t,\theta)^2}{M} + \frac{1}{M^2} + \eta M\right). \tag{123}$$

Consequently,

$$0 \le e_{\mathrm{reg}} \le C_N L^k\eta^{1/2} \qquad \text{on all of } \mathcal{K}_\theta, \tag{124}$$

$$0 \le e_{\mathrm{reg}} \le C_N L^k\eta^{2/3} \qquad \text{on } \mathcal{N}_0(t). \tag{125}$$

Proof. Apply original-value verification along the regularized comparison policy. Its original local trading criterion is zero by (112); hence

$$V_0 - J_\eta(u^M) = \mathbb{E}\int_t^T \left[Q_0(s, X_s, Y_s) + \tfrac{1}{2}\eta M^2 L_s V_{0,x}(s, X_s, Y_s)\|b - s\|_2^2\right] ds.$$

Equations (111), (121) and (122) give the right side of (123). Use $V_\eta \geq J_\eta(u^M)$ and R2 for the two inequalities. Choose $M = \eta^{-1/2}$ for (124). For (125), $d(t, \theta) = 0$, and $M = \eta^{-1/3}$ gives (125). Both choices meet $\eta M^2 \leq 1$ for small η. □

The different exponents come from an explicit balance. A fixed initial distance from the original no-trade set leaves a $1/M$ travel term. Starting in that set leaves a $1/M^2$ distance-weighted excursion term. Quadratic leakage contributes ηM in both cases. The dimension changes constants in these estimates, not these powers.

Proposition D.22 (Exact transfer identity)**.** *Under R2,*

$$J_0(\mathcal{P}_\eta(R_\eta)) - V_\eta = \mathbb{E}^{0,\mathcal{P}_\eta(R_\eta)}\int_t^T \frac{\eta}{2} L_s V_{\eta,x}(s, X_s, Y_s)\|\mathcal{P}_\eta(R_\eta)(s, X_s, Y_s)\|_2^2 ds \geq 0. \tag{126}$$

Thus $0 \leq e_{\text{one}} \leq e_{\text{reg}}$ *in every dimension.*

Proof. Set $w = \mathcal{P}_\eta(R_\eta)$ in (92); then $a_\eta(w) = 0$. R2 supplies its verification and admissibility hypotheses, and $J_0(w) \leq V_0$ gives the final comparison. □

Lemma D.23 (An interior bound on all one-shot-policy trade directions)**.** *Under R1–R3, on every interior compact space-time set whose ratios have a neighborhood in* $\operatorname{int}\mathcal{K}_\theta$*,*

$$\|\mathcal{P}_\eta(R_\eta)\|_2 \leq C_N \eta^{-1/2}$$

for small η*. A locally fixed active set is not required.*

Proof. Let $d_i^B = (-1, e_i)$ and $d_i^S = (q_i, -e_i)$ be unit-dollar trade directions in (x, y). Choose a fixed $h > 0$ so that, for every state z in the compact set and each of these finitely many directions d, the backward state $z - hd$ is feasible in a slightly larger interior compact set. The original investor at $z - hd$ can trade immediately to z, so $V_0(s, z - hd) \geq V_0(s, z)$. Concavity and (124) imply

$$\begin{aligned} D_d V_\eta(s, z) &\leq \frac{V_\eta(s, z) - V_\eta(s, z - hd)}{h} \\ &\leq \frac{V_0(s, z) - V_0(s, z - hd) + C_N\sqrt{\eta}}{h} \leq C_N\sqrt{\eta}. \end{aligned}$$

In particular this bounds the *positive* parts of every buy and sale derivative. Also $V_{\eta,x} \geq c_N > 0$ on the smaller compact set: a fixed small forward cash secant of V_0 is uniformly positive by R1, uniform convergence preserves this property for V_η, and concavity bounds $V_{\eta,x}$ below by that secant. Finally,

$$(g_{\eta,i})_+ = \frac{(D_{d_i^B} V_\eta)_+}{\eta V_{\eta,x}}, \qquad (-g_{\eta,i})_+ = \frac{(D_{d_i^S} V_\eta)_+}{\eta V_{\eta,x}}.$$

There are finitely many coordinates, which proves the bound. The argument does not require all components of $DV_\eta - DV_0$ to have a specified convergence rate. □

Theorem D.24 (Matching one-shot-deployment and regularization orders)**.** *Under R1–R3, fix an interior state* (t, z) *with ratio in* $\operatorname{int}\mathcal{K}_\theta$ *and* $Q_0(t, z) > 0$*. Then for some* $c_N, C_N, \eta_{0,N} > 0$*,*

$$c_N \eta^{1/2} \leq e_{\text{one}}(\eta; t, z) \leq e_{\text{reg}}(\eta; t, z) \leq C_N \eta^{1/2} \qquad (0 < \eta < \eta_{0,N}). \tag{127}$$

In particular both errors are $\Theta_N(\eta^{1/2})$ *at this state.*

Proof. Choose an interior space-time neighborhood of the initial state where $Q_0 \geq q_* > 0$. Its spatial boundary is a positive distance from the initial state. Stop the original-deployed one-shot-policy process on exit. Lemma D.23 bounds its drift on this neighborhood by $C_N\eta^{-1/2}$; the diffusion matrix is bounded there. For $h_\eta = \varepsilon\sqrt{\eta}$, choose a fixed $\varepsilon > 0$ sufficiently small that the stopped

drift displacement is less than one quarter of this boundary distance. The vector martingale maximal inequality bounds the probability of a diffusion displacement exceeding another quarter by $C_N h_\eta$. For small η, therefore, the process stays in the neighborhood until $t + h_\eta$ with probability at least $1/2$. For an original rate u, denote its local trading criterion against V_0 by

$$\mathcal{G}_0(u; DV_0) = L \sum_i [(V_{0,y_i} - V_{0,x})(u_i)_+ + (q_i V_{0,x} - V_{0,y_i})(-u_i)_+].$$

It is nonpositive. Original verification gives

$$e_{\text{one}} = \mathbb{E}^{0,\mathcal{P}_\eta(R_\eta)} \int_t^T [Q_0 - \mathcal{G}_0(\mathcal{P}_\eta(R_\eta); DV_0)] ds \geq \tfrac{1}{2} q_* \varepsilon \sqrt{\eta}.$$

Proposition D.22 and (124) give the remaining inequalities. The condition $Q_0 > 0$ is essential to this lower bound; no such Θ claim is made at every state. □

D.9.5 What multidimensional FR must accomplish

Lemma D.25 (An original-value-preserving initial target)**.** *Under R1 and R3, every initial state with ratio in $\mathcal{K}_\theta$ has a feasible fixed-time target z^0 with ratio in $\mathcal{N}_0(t)$, such that*

$$V_0(t, z^0) = V_0(t, z), \qquad L(z^0) \geq r_N L(z)$$

for a fixed constant $r_N > 0$ independent of η and the initial ratio.

Proof. Freeze physical time and follow the viable pure-trade flow from R3, parametrized by τ, with

$$\dot{\theta} = a_v(t, \theta), \qquad \dot{L} = -c_v L.$$

It corresponds to dollar increments $dB_i = Lb_i\, d\tau$, $dS_i = Ls_i\, d\tau$. By (116), (114) and $k < 0$,

$$\frac{d}{d\tau}\Phi(t, \theta) = \mathcal{T}_v \Phi + k c_v \Phi \leq -c_N d \leq -\widetilde{c}_N \sqrt{\Phi}$$

until arrival. The nonnegative function $\sqrt{\Phi}$ therefore decreases at a fixed positive rate while it is positive. Arrival at $\Phi = 0$ occurs in artificial time at most $C_N d(t, \theta) \leq S_N < \infty$. Since $0 \leq c_v \leq \max_i \alpha_i$, one has $L(z^0) \geq e^{-S_N \max_i \alpha_i} L(z)$. Total dollar variation is at most $L(z) S_N$. The chain rule and (112) show that V_0 is constant along this flow. If the initial ratio is in $\mathcal{N}_0(t)$, use zero trade. □

Let A be the initial *net* order resulting from a cost-free virtual search, and set $B = A^+$ and $S = A^-$. Only this order is executed. Its charged endpoint is

$$\widehat{z} = \Gamma_z(B, S) = (x - \mathbf{1}^\top B + q^\top S,\ y + B - S). \tag{128}$$

The feasible set $\mathcal{A}(z)$ consists of nonnegative B, S with nonnegative endpoint holdings and positive endpoint liquidation wealth; endpoint cash may be negative. Competing immediate trades are evaluated under their actual proportional costs. The implemented virtual path is *not* charged for its intermediate overshoots or round trips: its candidate endpoint and the residual below use exclusively $B = A^+$ and $S = A^-$. For a fixed net order, each coordinate has only one trade direction; the straight-line trade path is feasible when its charged endpoint and the initial portfolio satisfy the stated convex solvency constraints. Define

$$\mathcal{M}V_\eta(t, z) = \sup_{(B', S') \in \mathcal{A}(z)} V_\eta(t, \Gamma_z(B', S')).$$

Because $\alpha_i > 0$, feasible purchases satisfy $\alpha^\top B < L(z)$, and sales satisfy $S_i \leq y_i + B_i$. Thus feasible trade totals are bounded. We only need the supremum; where an optimum is invoked, its existence at positive wealth is included.

Proposition D.26 (A terminal-ratio certificate for accumulated FR trades)**.** *Suppose $V_\eta(t, \cdot)$ is concave and differentiable at the feasible* charged net-order *endpoint, and $R^* = R_\eta(t, \widehat{z}) \in C$. Assume $m_x^* = V_{\eta,x}(t, \widehat{z}) > 0$ and define*

$$\kappa_\eta = m_x^* \sum_{i=1}^N \left[(1 - R_i^*) B_i + (R_i^* - q_i) S_i\right] \geq 0. \tag{129}$$

Then

$$0 \le \mathcal{M}V_\eta(t,z) - V_\eta(t,\widehat{z}) \le \kappa_\eta. \tag{130}$$

In particular, the endpoint globally maximizes the immediate regularized continuation if

$$B_i > 0 \implies R_i^* = 1, \qquad S_i > 0 \implies R_i^* = q_i \quad \textit{for all } i. \tag{131}$$

These conditions are equivalent to $\kappa_\eta = 0$ under the stated wedge and positivity assumptions.

Proof. The function $F(B,S) = V_\eta(t, \Gamma_z(B,S))$ is concave because Γ_z is affine. Its derivatives at the endpoint are

$$F_{B_i} = m_x^*(R_i^* - 1) \le 0, \qquad F_{S_i} = m_x^*(q_i - R_i^*) \le 0.$$

For any feasible nonnegative B', S', its supporting hyperplane gives

$$\begin{aligned} F(B',S') - F(B,S) &\le \sum_i [F_{B_i}(B_i' - B_i) + F_{S_i}(S_i' - S_i)] \\ &\le -\sum_i [F_{B_i} B_i + F_{S_i} S_i] = \kappa_\eta. \end{aligned}$$

Take the supremum. Each summand in (129) is nonnegative, proving the final assertion. This is a global certificate, including competitors on state-constraint faces; no assertion of necessity at every constrained optimizer is needed. □

Remark D.27 (Exact ratio access and reaching the wedge are different from this certificate). Even in one dimension, a virtual first-entry stop need not satisfy complementarity after the final cash fee. In several dimensions, a later virtual update in asset j can move the ratio of an already-updated asset i, leaving a positive net purchase whose charged endpoint ratio lies in $(q_i, 1)$ and hence a positive term in (129). Exact access to R_η, no overshoot, and arrival in C therefore do not imply (131): the certificate concerns the executed cumulative *net* order, and the fee can move the endpoint out of C.

Theorem D.28 (FR error with a computable multidimensional residual)**.** *Under R1–R4, start with a ratio in $\mathcal{K}_\theta$ and suppose its* executed net order *reaches a feasible charged endpoint $\widehat{z} = \Gamma_\alpha(z, A_\eta)$ with $R_\eta(t,\widehat{z}) \in C$. Then*

$$0 \le e_\star(\eta; t, z) \le C_N L(z)^k \eta^{2/3} + \kappa_\eta. \tag{132}$$

If the charged initial net trade maximizes $\mathcal{M}V_\eta$, the κ_η term can be omitted. In particular it can be omitted under (131)*. A bound $\kappa_\eta = O_N(L^k\eta^{2/3})$ also gives $e_\star = O_N(L^k\eta^{2/3})$.*

Proof. The regularized HJB gives

$$Q_\eta := -(V_{\eta,t} + \mathcal{L}^0 V_\eta) = \frac{L V_{\eta,x}}{2\eta} \operatorname{dist}(R_\eta, C)^2.$$

Thus $Q_\eta = 0$ along the *certified physical* continuation of R4 at almost every physical time. The virtual FR path does not enter this Itô calculation. Verified Itô integration under the original dynamics, including intervention terms, and R4 imply

$$J_0(\pi_\eta^\star; t, z) \ge V_\eta(t, \widehat{z}).$$

Equality holds for a continuous complementary regulator after the initial trade. Let z^0 be the original-value-preserving target from Lemma D.25. Proposition D.26 and (125) yield

$$\begin{aligned} J_0(\pi_\eta^\star; t, z) &\ge \mathcal{M}V_\eta(t,z) - \kappa_\eta \\ &\ge V_\eta(t, z^0) - \kappa_\eta \\ &\ge V_0(t, z^0) - C_N L(z^0)^k \eta^{2/3} - \kappa_\eta \\ &\ge V_0(t,z) - C_N r_N^k L(z)^k \eta^{2/3} - \kappa_\eta. \end{aligned}$$

Absorb r_N^k into the constant. Original admissibility gives nonnegative error. If the initial intervention attains the envelope, use the exact envelope equality directly, even if a different certificate for its optimality is used. □

Corollary D.29 (Strict one-shot–refinement rate separation for fixed N)**.** *Under the hypotheses of Theorems D.24 and D.28, at a fixed interior initial state with ratio in* $\operatorname{int}\mathcal{K}_\theta$ *and* $Q_0 > 0$*,*

$$\frac{e_\star}{e_{\mathrm{one}}} \le C_N \eta^{1/6} + \frac{\kappa_\eta}{c_N\sqrt{\eta}}. \tag{133}$$

In particular, if (131) *holds, or more generally* $\kappa_\eta = O_N(\eta^{2/3})$*, then*

$$\begin{aligned} V_0 - V_\eta &= \Theta_N(\eta^{1/2}), \\ V_0 - J_0(\mathcal{P}_\eta(R_\eta)) &= \Theta_N(\eta^{1/2}), \\ 0 \le V_0 - J_0(\pi^\star_\eta) = O_N(\eta^{2/3}), &\qquad \frac{V_0 - J_0(\pi^\star_\eta)}{V_0 - J_0(\mathcal{P}_\eta(R_\eta))} = O_N(\eta^{1/6}) \to 0. \end{aligned}$$

For all sufficiently small η*,*

$$J_0(\pi^\star_\eta) > J_0(\mathcal{P}_\eta(R_\eta)), \qquad J_0(\pi^\star_\eta) - J_0(\mathcal{P}_\eta(R_\eta)) = \Theta_N(\sqrt{\eta}).$$

Even the weaker condition $\kappa_\eta = o_N(\sqrt{\eta})$ *suffices for relative improvement and eventual strict ranking, though it does not alone give the* $2/3$ *upper exponent.*

Proof. Divide (132) by the positive lower bound in (127), absorbing the fixed initial wealth factor. Under either stated $2/3$ condition the quotient tends to zero. Then $e_{\mathrm{one}} - e_\star$ is between $c_N\sqrt{\eta}/2$ and $C_N\sqrt{\eta}$ for small η. The same argument works when only the quotient tends to zero via $\kappa_\eta = o_N(\sqrt{\eta})$. □

D.9.6 Geometric certificates, including correlated corners

This section makes R3 more concrete. The existence of a feasible original-active field and its realizations remains necessary; ordinary smoothness of coefficients alone does not establish it.

Proposition D.30 (A sufficient smooth-boundary construction)**.** *Suppose the original no-trade set has an exterior distance function* d *with uniformly bounded first and second spatial derivatives off its boundary. Suppose it has a bounded smooth extension* ρ *to all of* $\mathcal{K}_\theta$*, equal to* d *outside, with bounded spatial derivatives, and* $|\rho_t| \le C_N a(t)$*. Assume* $|d_t| \le C_N a(t)$ *outside and a feasible active field satisfies*

$$Dd \cdot a_v \le -\nu_N < 0 \quad \textit{outside } \mathcal{N}_0(t).$$

Assume the corresponding piecewise functions below admit the generalized Itô formula without a singular first-derivative jump. Then the certificate part of R3 holds. A uniformly smooth convex set with a smooth signed-distance collar and suitable uniform bounds supplies the stated distance and extension properties.

Proof. Take $\psi = C_* \exp(K\rho)$, with C_* normalizing its minimum to at least one. Put $c_{\max} = \max_i \alpha_i$. Outside the no-trade set,

$$\frac{\mathcal{T}_v\psi}{\psi} = KDd \cdot a_v + |k|c_v \le -K\nu_N + |k|c_{\max}.$$

Choose $K\nu_N > |k|c_{\max} + 1$. Inside, $v = 0$. Spatial and time bounds give (115). For a possibly larger K, set $\Phi = d^2 e^{Kd}$ outside and zero inside. This function and its first derivatives match continuously at the boundary. Outside,

$$\mathcal{T}_v\Phi = e^{Kd}\big[(2d + Kd^2)Dd \cdot a_v + |k|c_v d^2\big] \le -2\nu_N d.$$

The Hessian is bounded outside, and all derivatives vanish inside except for the harmless boundary trace of the second derivative. The assumed Itô interpretation and the time bound yield (116). Bounded d yields (114). The full covariance matrix appears only in bounded coefficients of $\mathcal{G}_N$. □

Proposition D.31 (A sufficient polyhedral construction with compatible corners)**.** *Suppose on* $\mathcal{K}_\theta$ *the original no-trade set is*

$$\mathcal{N}_0(t) = \{\theta : r_j(t,\theta) := n_j^\top \theta - b_j(t) \le 0,\ j = 1, \dots, J\},$$

with fixed unit normals, absolutely continuous bounded offsets, and a uniform error bound

$$\max_j (r_j)_+ \le d(t,\theta) \le H_N \max_j (r_j)_+. \tag{134}$$

Assume the feasible active field has bounded a_v and constants $\varepsilon, \nu_N > 0$ such that, whenever $\theta \notin \mathcal{N}_0(t)$,

$$n_j^\top a_v(t,\theta) \le -\nu_N \quad \textit{for every face with } r_j(t,\theta) \ge -\varepsilon. \tag{135}$$

Set $a(t) = \sum_j |b_j'(t)|$ and assume (117)*. Then the test-function part of R3 holds, including at intersections of faces.*

Proof. All r_j and $n_j^\top a_v$ are bounded on the domain. Write the latter bound as B_N. For large K use, up to a fixed positive normalization,

$$\psi = \sum_{j=1}^J e^{K r_j}, \qquad \Phi = \sum_{j=1}^J (r_j)_+^2 e^{K (r_j)_+}.$$

Consider an outside point. Terms of $\mathcal{T}_v \psi$ from faces with $r_j \ge -\varepsilon$ are at most

$$(-K\nu_N + |k| c_{\max}) \sum_{r_j \ge -\varepsilon} e^{K r_j}.$$

At least one of these exponential weights is at least one. The remaining positive contribution is at most $J(KB_N + |k| c_{\max}) e^{-K\varepsilon}$. For sufficiently large fixed K, the total is at most a negative constant. Since ψ has a finite upper bound, this implies $\mathcal{T}_v \psi \le -c_N \psi$. Spatial derivatives are bounded and $|\psi_t| \le C_N a(t) \psi$. For every violated face let $r = r_j > 0$. Its contribution to $\mathcal{T}_v \Phi$ is at most

$$e^{Kr}\big[-\nu_N (2r + Kr^2) + |k| c_{\max} r^2 \big] \le -2\nu_N r$$

when $K\nu_N \ge |k| c_{\max}$. Therefore $\mathcal{T}_v \Phi \le -2\nu_N \sum_j (r_j)_+ \le -(2\nu_N / H_N) d$. Equation (134) gives comparability of Φ and d^2. The spatial terms in (116) are bounded and vanish in the interior of $\mathcal{N}_0$, and $|\Phi_t| \le C_N a(t) d$. Each summand is a function of a single affine face coordinate. The function $r \mapsto (r_+)^2 e^{K r_+}$ is C^1 with locally absolutely continuous derivative. The one-dimensional generalized Itô formula applied to each face coordinate therefore adds no local-time term. Summing these formulas remains valid where several faces meet. Quadratic variations use the full $H\Sigma H$; no independence assumption is made. This proves the certificate assertions with their proper Itô interpretation. □

Condition (135) is a genuine compatibility requirement: the binding slacks in (112) specify which nonnegative weights b, s may be nonzero, and positive definiteness of Σ alone does not guarantee a common feasible inward direction. The proposition does not assert that CRRA no-trade sets are polyhedral; it is a sufficient case that includes genuine corners.

D.9.7 SCOPE

The results extend the one-asset comparison argument to any *fixed* finite N satisfying the hypotheses on $\mathcal{K}_\theta$, for the original leveraged problem, with neither independent returns nor axis-aligned no-trade geometry. They do require geometric control of excursions and an admissible FR continuation. The exponents are independent of N; constants may deteriorate with covariance bounds, $\min_i \alpha_i$, inward angles, H_N, and boundary regularity, and a joint limit $N = N(\eta) \to \infty$ is not covered. For a compactly supported initial distribution with ratios in $\operatorname{int} \mathcal{K}_\theta$ that charges a compact set with $Q_0 > 0$, integrating the proofs gives averaged $e_{\mathrm{reg}}, e_{\mathrm{one}} = \Theta_N(\sqrt{\eta})$, and averaged FR error $O_N(\eta^{2/3})$ under a corresponding bound on κ_η.

Proposition D.32 (Exact wedge regulation need not dominate the original-deployed one-shot policy)**.** *Suppose $\mathcal{P}_\eta(R_\eta)$ satisfies the exact transfer identity, and the admissible exact-ratio refinement strategy $\pi_\eta^\star$ starts in the regularized wedge, has no initial value-changing trade, tracks the wedge almost everywhere in physical time, and all its later intervention contributions to V_η vanish. Then*

$$J_0(\pi_\eta^\star) = V_\eta \le J_0(\mathcal{P}_\eta(R_\eta)).$$

The inequality is strict if the transfer integral is positive.

Proof. In Theorem D.13, both the V_η residual and its intervention sum vanish for this regulator. Hence its payoff equals V_η. The exact transfer identity supplies the comparison with the one-shot policy. □

This conditional obstruction explains why the strict-intervention hypothesis in Corollary D.29 cannot be replaced by a universal initial-state claim. It does not construct such a regulator for every model, and no matching $\eta^{2/3}$ lower rate for FR is asserted. No inclusion between the regularized and original no-trade sets is assumed, and CRRA homogeneity by itself does not establish such an inclusion in a correlated multidimensional problem: the improvement follows from the initial intervention envelope and the sharper regularization bound on the original no-trade set. Applying these results to a specific market requires verifying R1–R4, which has not been done for every numerical market in the paper. The rate proof does not identify arbitrary local refinement with a multidimensional immediate optimizer.

D.9.8 Equal convergence orders do not mean equal gaps

The first two errors in (110) have the same exponent at a strict intervention state, but they are not equal. Their exact decomposition is

$$e_{\mathrm{reg}} = e_{\mathrm{one}} + D_\eta, \qquad D_\eta := J_0(\mathcal{P}_\eta(R_\eta)) - V_\eta. \tag{136}$$

The gain D_η is given by the nonnegative integral in (126). In fact the existing hypotheses imply a matching lower bound for this gain as well. Thus the benefit of removing leakage is itself of the same order as the original regularization error.

Theorem D.33 (Removing leakage yields a nonvanishing fractional improvement)**.** *Under R1–R3, fix an interior initial state (t, z) with ratio in $\operatorname{int}\mathcal{K}_\theta$ and $Q_0(t, z) > 0$. There are positive constants $c_{P,N}, c_{D,N}, C_N$, independent of small η, such that*

$$c_{P,N}\sqrt{\eta} \le e_{\mathrm{one}}, \qquad c_{D,N}\sqrt{\eta} \le D_\eta, \qquad e_{\mathrm{one}} + D_\eta = e_{\mathrm{reg}} \le C_N\sqrt{\eta}. \tag{137}$$

Consequently $D_\eta = \Theta_N(\sqrt{\eta})$, and there are $a_N, b_N > 0$ such that

$$a_N \le \frac{e_{\mathrm{one}}}{e_{\mathrm{reg}}} \le 1 - b_N < 1 \qquad \textit{for all sufficiently small } \eta. \tag{138}$$

Removing leakage therefore reduces the gap by at least a fixed positive fraction, while a fixed positive fraction remains. The constants can depend on the initial state. The result does not establish a universal fractional reduction, a limit of this ratio, or an asymptotic coefficient equal to one half.

Proof. Theorem D.24 proves the one-shot-deployment lower bound and the regularization upper bound. It remains to prove the lower bound for D_η. Choose a closed spatial ball of radius $r_* > 0$ about z, and a fixed subsequent time interval, lying in the interior and satisfying $Q_0 \ge q_* > 0$. The proof of Lemma D.23 gives a uniform lower bound on $V_{\eta,x}$ there. Since L also has a positive lower bound, put $LV_{\eta,x} \ge m_* > 0$ on this neighborhood for all small η. Let τ be the first spatial exit time of the original-deployed one-shot-policy process. Take $h_\eta = K\sqrt{\eta}$, small enough to fit within the chosen time interval. Original verification and $-\mathcal{G}_0 \ge 0$ imply

$$q_*\mathbb{E}[(\tau - t) \wedge h_\eta] \le e_{\mathrm{one}} \le C_N\sqrt{\eta}, \qquad \mathbb{P}(\tau > t + h_\eta) \le \frac{C_N}{q_*K}.$$

Choose $K \ge 4C_N/q_*$. Then spatial exit by $t + h_\eta$ has probability at least $3/4$. Write the stopped state dynamics as a no-trade drift, a controlled drift $f(\zeta_s, \mathcal{P}_\eta(R_\eta))$, and a martingale N_s, with $\zeta_s = (X_s, Y_s)$. On the ball, $\|f(\zeta, u)\| \le C_f\|u\|_2$, because the trade drift in (108)–(109) is proportional to bounded L and linear in the positive and negative parts of u. The no-trade drift and diffusion are bounded. For sufficiently small η, the accumulated no-trade drift over h_η has norm at most $r_*/4$. The martingale maximal inequality gives

$$\mathbb{P}\left(\sup_{s\in[t,t+h_\eta]} \|N_{s\wedge\tau} - N_t\| > r_*/4\right) \le C_N h_\eta \le 1/4.$$

With probability at least $1/2$, therefore, spatial exit occurs by $t+h_\eta$ while the martingale displacement stays below $r_*/4$. Continuity of the one-shot-policy state makes its exit displacement equal to r_*. On this event the controlled drift must account for at least $r_*/2$, so

$$\int_t^{\tau\wedge(t+h_\eta)} \|\mathcal{P}_\eta(R_\eta)\|_2 ds \geq \frac{r_*}{2C_f}.$$

Cauchy–Schwarz then implies, on the same event,

$$\int_t^{\tau\wedge(t+h_\eta)} \|\mathcal{P}_\eta(R_\eta)\|_2^2 ds \geq \frac{r_*^2}{4C_f^2 h_\eta}.$$

Restrict the nonnegative integral in (126) to this stopped interval. It follows that

$$D_\eta \geq \frac{\eta m_*}{2}\,\mathbb{E}\int_t^{\tau\wedge(t+h_\eta)} \|\mathcal{P}_\eta(R_\eta)\|_2^2 ds \geq \frac{m_* r_*^2}{16C_f^2 K}\sqrt{\eta}.$$

This proves (137). Divide its two lower bounds by $e_{\mathrm{reg}} \leq C_N\sqrt{\eta}$ and use (136) to obtain (138). □

The original-model occupation and cash-marginal weight in (126) are essential: subtracting an unweighted penalty along a shared path would not give D_η. Leakage removal and the remaining finite-speed loss each have square-root order at the stated strict intervention states.

D.9.9 Proof of Theorem 4

Theorem D.21 gives the regularization upper bound. Theorem D.24 supplies the matching one-shot-deployment lower bound. Theorem D.33 gives the leakage-removal lower bound and exact gap decomposition. Theorem D.28 and Corollary D.29 then give the bound and comparison only for the R4-certified *charged* net order and viable continuous-time policy. They do not establish an asymptotic rate for the implemented finite virtual refinement. All values are those of the original admissible optimization problems; only the comparison policies are assumed to remain in $\mathcal{K}_\theta$.

E Financial foundations

E.1 Admissibility embedding used in value comparisons

Whenever $V_\eta \leq V_0$ is used, assume that an admissible regularized strategy can be implemented in the original model with the same adapted *dollar* trades, admissible continuation, and terminal integrability. Risky holdings then coincide and removing the nonnegative quadratic leakage only increases cash, so $J_0 \geq J_\eta$ for this coupled construction and hence $V_0 \geq V_\eta$. This embeds strategies; it does not compare one feedback function along two different state trajectories, and positivity of an optimal cash marginal does not give $W_x > 0$ for an arbitrary frozen policy.

E.2 Admissibility and well-posedness

Let $Z = (X, Y^1, \ldots, Y^N)$ and $\mathcal{S} = \{(x, \mathbf{y}) : y_i \geq 0,\ x + \sum_i (1-\alpha_i) y_i > 0\}$. A singular strategy (B, S) is admissible if its components are adapted, right-continuous, nondecreasing, and of finite variation; the state equation (2)–(3) has a càdlàg solution; purchases, sales, and jumps preserve $\mathcal{S}$; terminal liquidation wealth is strictly positive; and the state and total variation have a finite pth moment for some $p > 1$. Simultaneous purchase and sale can be excluded because a positive spread makes such round trips weakly inferior. For the finite L^2 estimate below, additionally assume the initial and input differences on its right-hand side have finite second moments. This extra integrability is not implied by a generic $p > 1$ admissibility condition.

Proposition E.1 (Fixed-input state well-posedness)**.** *Suppose the uncontrolled drift and diffusion are locally Lipschitz on $\mathcal{S}$ and of linear growth. Each fixed admissible finite-variation input generates a unique adapted càdlàg strong solution up to the first exit from $\mathcal{S}$; admissibility extends that solution to T. For two inputs driven by the same Brownian motion and a compact $K \Subset \mathcal{S}$, define*

$$\tau_K := \inf\{s \geq t : Z_s \notin K \text{ or } \widetilde{Z}_s \notin K\} \wedge T.$$

Then

$$\mathbb{E}\sup_{t\le s\le\tau_K}|Z_s-\widetilde{Z}_s|^2\le C_K\mathbb{E}\left[|z-\widetilde{z}|^2+\sum_i\sup_{t\le s\le\tau_K}|B^i_s-\widetilde{B}^i_s|^2+\sum_i\sup_{t\le s\le\tau_K}|S^i_s-\widetilde{S}^i_s|^2\right]. \tag{139}$$

If the coefficients are globally Lipschitz, or if both trajectories are known to remain in K, one may replace τ_K by T.

Proof. The singular inputs enter additively through constant intervention directions. Stop both equations at τ_K, subtract them, apply the Burkholder–Davis–Gundy inequality to the stopped martingale part, and close the local Lipschitz estimate by Gronwall's lemma. At jump times, the jump of the state difference is the linear image of the corresponding intervention-input jumps under the constant intervention directions. Standard localization gives existence and uniqueness up to exit; the admissibility requirement prevents exit from $\mathcal{S}$ before T. □

Assumption E.2 (Localized optimization hypotheses)**.** *On a compact viable state-constraint domain $K \Subset \mathcal{S}$, assume the compactness, tightness, closedness, reward upper-semicontinuity, and measurable selection hypotheses of the regular–singular control theory of Haussmann and Suo Haussmann & Suo (1995b;a). Maximizing sequences have tight total variation, and the admissible family is nonempty.*

Remark E.3 (Positive spread and expected-turnover control)*.* On a compact stopped working band, liquidation wealth satisfies

$$dL_t = b_L(t,Z_t)\,dt + dM_t - \sum_i \alpha_i\,dB^i_t,$$

where b_L is bounded on the band. Positivity of L and optional stopping therefore give a first-moment estimate $\mathbb{E}\sum_i B^i_T \le C/\min_i \alpha_i$ for admissible maximizing sequences. The position equations and $Y^i \ge 0$ then yield a corresponding first-moment bound for sales. This supports the total-variation tightness hypothesis, but does not by itself verify the required $p > 1$ moment bounds, closure of the state-constrained admissible set, or absence of a relaxation gap.

Theorem E.4 (Localized relaxed existence and DPP)**.** *Under the preceding localized optimization assumptions, the problem has a finite value and an optimal relaxed singular control. If an optimal relaxed control is representable by a strict admissible control with the same performance, the strict problem also attains its value. A no-relaxation-gap identity alone is insufficient for this strict-attainment conclusion. Under the corresponding concatenation and measurable-selection hypotheses, the dynamic programming principle holds. Continuity requires the stated continuity and state-stability hypotheses; the value is the unique viscosity solution of the state-constraint singular HJB whenever comparison holds.*

Proof sketch. Apply the regular–singular compactness result of Haussmann & Suo (1995b) to a maximizing sequence in the specified localized formulation. Closedness makes the relaxed limit admissible, and upper semicontinuity gives it a payoff at least as large as the limiting payoffs. It is therefore a relaxed optimizer. The additional representation condition transfers this optimizer to the strict class; equality of optimal values alone gives only a strict maximizing sequence. The concatenation and measurable-selection argument of Haussmann & Suo (1995a) yields the DPP under its hypotheses. Continuity and the viscosity characterization use their corresponding regularity hypotheses, and comparison supplies uniqueness for the specified boundary data. □

E.3 Verification for the constant-coefficient CRRA–GBM benchmark

Let S^i denote the ask-price process underlying the dollar position $Y^i = \varphi^i \mathsf{S}^i$. Write $W = CB$, where B is a standard N-dimensional Brownian motion, $CC^\top = \rho$, and define

$$a := \mathrm{diag}(\sigma_1,\dots,\sigma_N)C, \qquad \vartheta := a^{-1}(\mu - r\mathbf{1}).$$

Proposition E.5 (Benchmark financial well-posedness)**.** *Assume $T < \infty$, $\gamma > 1$, $\alpha_i \in (0,1)$, $\sigma_i > 0$, and that ρ is positive definite. Then:*

(i) The measure $\mathbb{Q} \sim \mathbb{P}$ with density

$$Z_T := \frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(-\vartheta^\top B_T - \frac{1}{2}\|\vartheta\|^2 T\right) \tag{140}$$

and the shadow prices $\widetilde{\mathsf{S}}^i := (1-\alpha_i/2)\mathsf{S}^i$ form a strictly consistent price system: each shadow price lies strictly inside its bid–ask spread and $e^{-rt}\widetilde{\mathsf{S}}^i_t$ is a $\mathbb{Q}$-martingale.

(ii) Let

$$\widetilde{w}(x,\mathbf{y}) := x + \sum_i (1-\alpha_i/2) y_i. \tag{141}$$

For every admissible singular strategy and every admissible regularized-rate strategy, with terminal liquidation wealth L_T,

$$\mathbb{E}_{t,z}[Z_{t,T} L_T] \le e^{r(T-t)}\widetilde{w}(x,\mathbf{y}), \tag{142}$$

where

$$Z_{t,T} := \frac{Z_T}{Z_t} = \exp\left(-\vartheta^\top (B_T - B_t) - \frac{1}{2}\|\vartheta\|^2 (T-t)\right)$$

is the conditional density increment induced by (140)*.*

(iii) For the singular benchmark,

$$U\left(L_t e^{r(T-t)}\right) \le V_0(t,z) \le \frac{1}{\gamma-1} < \infty. \tag{143}$$

Moreover, $\mathrm{AE}(U) = 0$*, the convex conjugate*

$$U^*(q) := \sup_{\ell>0}\{U(\ell) - q\ell\} = \frac{1}{\gamma-1} - \frac{\gamma}{\gamma-1} q^{(\gamma-1)/\gamma}$$

has only polynomial growth, and Z_T has moments of every real order. Thus the displayed construction directly verifies robust no-arbitrage, finite primal value, and a finite dual candidate together with the relevant CRRA growth and integrability inputs. Under a standard Fatou-closed constrained admissible formulation satisfying the remaining hypotheses of the cited multivariate proportional-cost utility theorem, the primal value is attained and the corresponding convex-duality formulation has no gap Cvitanić & Karatzas (1996); Campi & Owen (2011).

Proof. Because a is nonsingular and ϑ is constant, Novikov's condition is immediate. Under $\mathbb{Q}$, $B^{\mathbb{Q}}_t := B_t + \vartheta t$ is Brownian and the ask-price drift becomes $\mu - a\vartheta = r\mathbf{1}$. Multiplication by the constant $1-\alpha_i/2 \in (1-\alpha_i, 1)$ preserves the discounted martingale property and places the shadow price in the relative interior of the spread, proving (i). Set $\chi_i := 1-\alpha_i/2$ and $\widetilde{W} = X + \sum_i \chi_i Y^i$. For a singular strategy, discounted shadow wealth has the form

$$d(e^{-rs}\widetilde{W}_s) = dM_s - e^{-rs}\sum_i \frac{\alpha_i}{2}(dB^i_s + dS^i_s),$$

where M is a local $\mathbb{Q}$-martingale. For an absolutely continuous rate strategy it satisfies

$$d(e^{-rs}\widetilde{W}_s) = dM_s - e^{-rs}\sum_i \frac{\alpha_i}{2}|A^i_s|\,ds - e^{-rs}\frac{\eta}{2L_s}\|A_s\|^2\,ds.$$

Since $Y^i \ge 0$ and $\chi_i > 1-\alpha_i$, one has $\widetilde{W}_s \ge L_s > 0$. Thus discounted shadow wealth is a nonnegative local supermartingale and hence a supermartingale. At terminal time $L_T \le \widetilde{W}_T$, which gives (142). Immediate liquidation followed by investment in the bank account gives the lower bound in (143); the upper bound follows from $U(\ell) \le 1/(\gamma-1)$. Finally, $\mathrm{AE}(U) = 0$, the displayed conjugate is obtained by optimizing at $\ell = q^{-1/\gamma}$, and the lognormal density (140) has all positive and negative moments. For any dual scale $y > 0$, the candidate yZ_T therefore has finite dual objective, so the dual value is finite as well. These calculations verify the model-specific market, growth, and integrability inputs. Primal attainment and the no-gap convex-duality conclusion in the stated Fatou-closed constrained class then follow by applying the cited multivariate utility theorem after identifying the scalar terminal criterion with terminal liquidation utility; they are not inferred from the localized working-band argument. This proves (iii). □

Proposition E.6 (CRRA homogeneity and a uniform cash-marginal bound)**.** *Under Proposition E.5, let V_η denote V_0 when $\eta = 0$ and the optimal quadratically regularized rate value when $\eta > 0$. On the domain considered, assume the value is finite and the admissible class is stable under positive scaling. Assume also cash lifting: an admissible adapted dollar-trade or dollar-rate process from $(x, \mathbf{y})$ can be reused from each compared state $(x + h, \mathbf{y})$, $h > 0$, with admissibility and the required integrability preserved. Finite-utility ε-optimal controls must be available at the states used in this comparison. Define*

$$G_\eta(t, x, \mathbf{y}) := 1 + (1 - \gamma)V_\eta(t, x, \mathbf{y}).$$

Then, for every $c > 0$,

$$G_\eta(t, cx, c\mathbf{y}) = c^{1-\gamma} G_\eta(t, x, \mathbf{y}). \tag{144}$$

At every differentiability point,

$$xV_{\eta,x} + \sum_i y_i V_{\eta,y_i} = G_\eta = 1 + (1 - \gamma)V_\eta. \tag{145}$$

Furthermore, put $\tau := T - t$ and

$$a_\gamma := (\gamma - 1)r + \frac{\gamma}{2(\gamma + 1)} \|\vartheta\|^2 .$$

For every $h > 0$ such that $(x + h, \mathbf{y})$ is admissible,

$$V_\eta(t, x + h, \mathbf{y}) - V_\eta(t, x, \mathbf{y}) \geq h \left(\widetilde{w}(x, \mathbf{y}) + h\right)^{-\gamma} e^{-a_\gamma \tau}. \tag{146}$$

Consequently, if $K \Subset \mathcal{S}$ is contained in a slightly larger compact $K^+ \Subset \mathcal{S}$ under sufficiently small cash perturbations and $\overline{w}_{K^+} := \sup_{z \in K^+} \widetilde{w}(z)$, then

$$\underline{\lambda}_{K^+} := \overline{w}_{K^+}^{-\gamma} \min_{0 \leq s \leq T} e^{-a_\gamma s} > 0 \tag{147}$$

satisfies $V_\eta(t, x + h, \mathbf{y}) - V_\eta(t, x, \mathbf{y}) \geq \underline{\lambda}_{K^+} h$ locally on $[0, T] \times K$. Thus $V_{\eta,x} \geq \underline{\lambda}_{K^+}$ at differentiability points, and every C^1 test function touching V_η from above or below has x-derivative at least $\underline{\lambda}_{K^+}$.

Proof. Scaling the initial state and every dollar trade or rate by c scales both the state and quadratic-over-linear leakage by c. Since $1 + (1 - \gamma)J_\eta = \mathbb{E}[L_T^{1-\gamma}]$, optimization gives (144); differentiation gives (145). Reuse a finite-utility ε-optimal control from $(x, \mathbf{y})$ after adding cash h. Risky holdings coincide. The cash difference is $D_s = he^{r(s-t)}$ for singular trades; for regularized dollar rates A_s,

$$dD_s = \left(r + \frac{\eta \|A_s\|^2}{2L_s(L_s + D_s)} \right) D_s \, ds, \qquad D_t = h.$$

Thus $D_T \geq d := he^{r\tau}$ and the payoff gain is at least $d(L_T + d)^{-\gamma}$. Put $q = \gamma/(\gamma + 1)$. The budget bound and Hölder's inequality imply

$$\mathbb{E}[(L_T + d)^{-\gamma}] \geq \frac{(\mathbb{E}[Z_{t,T}^q])^{\gamma+1}}{\mathbb{E}[Z_{t,T}(L_T + d)]^\gamma} \geq \frac{\exp[-\gamma \|\vartheta\|^2 \tau/(2(\gamma + 1))]}{[e^{r\tau}(\widetilde{w} + h)]^\gamma}.$$

Multiplying by d and letting $\varepsilon \downarrow 0$ proves (146). Replacing $\widetilde{w} + h$ by $\overline{w}_{K^+}$ gives the compact bound. Positive cash increments for test functions touching from above, and negative increments from nearby states for those touching from below, give the stated derivative bounds. □

Remark E.7 (What the cash-marginal bound does and does not verify). Proposition E.6 supplies item (iii) of Assumption D.1, uniformly on compact bands and in η, wherever its finite-value and cash-lifting conditions hold (on positive-cash bands these follow from the zero-rate policy and preserved moment conditions). It does not give a positive cash derivative for an arbitrary neural feedback, which can change its future actions in response to added cash; the fixed-policy positivity assumption in Theorem 3 remains separate.

F Experiments, diagnostics, and scope

For the coupled CRRA portfolio model, the DP/QVI reference is used only for evaluation in one and two assets. Appendix F.14 adds a separate $N = 50$ CARA/diagonal-covariance stress test whose Bellman criterion factorizes across risky coordinates; it is therefore a controlled high-dimensional regime oracle, not a 50-dimensional DP solution of the main CRRA problem. HCW rendezvous, reservoir release, and battery storage use their own dynamics and evaluation protocols.

F.1 WARM-START AND ACQUISITION-ERROR ABLATION

The one-dimensional ablation varies Stage-I training iterations and the number M_{MC} of independent frozen-continuation paths. Each cell reports the mean with standard deviation below under the common held-out reference protocol. The cash- and risky-marginal tables diagnose the two numerator/denominator components of the ratio, while the ratio and label tables are the quantities that enter the margin-based recovery mechanism most directly. These MAEs are empirical diagnostics for the error terms in the conditional bridge of Theorem B.6; they are not direct estimates of the global policy gap $V_\eta - J_\eta(\mathcal{P}_\eta(\widehat{R}^{\Delta}_{\mathrm{BPTT}}))$.

Warm start and continuation budget. Table 5 separates training effort and sampling budget in the one-asset diagnostic. With $M_{\mathrm{MC}} = 2048$, the reported mean ratio MAE is 0.0214 before Stage-I training and 0.0015 after 500 iterations. At 500 iterations, increasing M_{MC} from 16 to 2048 reduces the mean MAE from 0.0150 to 0.0015. The full tables report standard deviations and boundary violations. These comparisons motivate checking both continuation quality and acquisition accuracy instead of using rollout utility or sample count alone as a proxy for regime recovery.

Table 3: V_x MAE over algorithmic parameters (mean; standard deviation below).

Stage I	M_{MC}							
iterations	16	32	64	128	256	512	1024	2048
0	0.0575 (0.0125)	0.0428 (0.0095)	0.0334 (0.0066)	0.0276 (0.0061)	0.0235 (0.0085)	0.0211 (0.0100)	0.0198 (0.0119)	0.0192 (0.0130)
10	0.0820 (0.0206)	0.0638 (0.0185)	0.0540 (0.0182)	0.0495 (0.0200)	0.0464 (0.0213)	0.0449 (0.0224)	0.0443 (0.0229)	0.0441 (0.0231)
50	0.0467 (0.0036)	0.0327 (0.0030)	0.0234 (0.0029)	0.0166 (0.0024)	0.0119 (0.0018)	0.0088 (0.0015)	0.0066 (0.0015)	0.0051 (0.0015)
100	0.0451 (0.0065)	0.0314 (0.0057)	0.0225 (0.0044)	0.0157 (0.0030)	0.0109 (0.0022)	0.0077 (0.0017)	0.0057 (0.0016)	0.0042 (0.0014)
200	0.0437 (0.0036)	0.0314 (0.0031)	0.0224 (0.0023)	0.0156 (0.0015)	0.0109 (0.0010)	0.0074 (0.0006)	0.0052 (0.0005)	0.0035 (0.0004)
500	0.0430 (0.0041)	0.0319 (0.0035)	0.0226 (0.0027)	0.0149 (0.0016)	0.0109 (0.0011)	0.0076 (0.0009)	0.0058 (0.0007)	0.0039 (0.0006)

Table 4: V_y MAE over algorithmic parameters (mean; standard deviation below).

Stage I	M_{MC}							
iterations	16	32	64	128	256	512	1024	2048
0	0.0388 (0.0045)	0.0308 (0.0039)	0.0257 (0.0038)	0.0227 (0.0039)	0.0208 (0.0044)	0.0195 (0.0047)	0.0187 (0.0050)	0.0181 (0.0053)
10	0.0700 (0.0185)	0.0517 (0.0157)	0.0421 (0.0147)	0.0368 (0.0147)	0.0338 (0.0150)	0.0322 (0.0151)	0.0316 (0.0152)	0.0314 (0.0152)
50	0.0355 (0.0028)	0.0246 (0.0023)	0.0175 (0.0020)	0.0125 (0.0016)	0.0090 (0.0012)	0.0068 (0.0010)	0.0054 (0.0009)	0.0046 (0.0009)
100	0.0348 (0.0045)	0.0242 (0.0038)	0.0171 (0.0029)	0.0118 (0.0018)	0.0081 (0.0012)	0.0056 (0.0008)	0.0040 (0.0007)	0.0030 (0.0006)
200	0.0339 (0.0028)	0.0242 (0.0024)	0.0172 (0.0018)	0.0118 (0.0011)	0.0081 (0.0007)	0.0055 (0.0004)	0.0038 (0.0003)	0.0027 (0.0002)
500	0.0334 (0.0031)	0.0244 (0.0026)	0.0175 (0.0020)	0.0112 (0.0012)	0.0084 (0.0008)	0.0058 (0.0006)	0.0046 (0.0005)	0.0031 (0.0004)

Table 5: Ratio V_y/V_x MAE over algorithmic parameters (mean; standard deviation below).

Stage I	M_{MC}							
iterations	16	32	64	128	256	512	1024	2048
0	0.0408 (0.0059)	0.0313 (0.0057)	0.0268 (0.0069)	0.0246 (0.0081)	0.0231 (0.0092)	0.0222 (0.0102)	0.0216 (0.0109)	0.0214 (0.0114)
10	0.0322 (0.0065)	0.0258 (0.0063)	0.0229 (0.0066)	0.0210 (0.0069)	0.0201 (0.0071)	0.0196 (0.0073)	0.0194 (0.0074)	0.0193 (0.0075)
50	0.0161 (0.0035)	0.0115 (0.0029)	0.0085 (0.0024)	0.0066 (0.0020)	0.0052 (0.0018)	0.0044 (0.0018)	0.0039 (0.0018)	0.0037 (0.0018)
100	0.0155 (0.0038)	0.0108 (0.0027)	0.0079 (0.0020)	0.0059 (0.0015)	0.0043 (0.0011)	0.0033 (0.0009)	0.0026 (0.0009)	0.0021 (0.0009)
200	0.0150 (0.0020)	0.0103 (0.0013)	0.0074 (0.0009)	0.0052 (0.0006)	0.0037 (0.0004)	0.0027 (0.0004)	0.0020 (0.0004)	0.0016 (0.0004)
500	0.0150 (0.0019)	0.0106 (0.0014)	0.0078 (0.0010)	0.0055 (0.0007)	0.0038 (0.0005)	0.0027 (0.0003)	0.0020 (0.0002)	0.0015 (0.0002)

Table 6: Boundary-violation ratio over algorithmic parameters in percent (mean; standard deviation below).

Stage I	M_{MC}							
iterations	16	32	64	128	256	512	1024	2048
0	52.82 (8.30)	49.63 (13.78)	48.24 (17.25)	46.98 (19.69)	44.19 (21.57)	42.66 (26.49)	42.86 (27.15)	41.73 (26.88)
10	51.56 (6.37)	50.56 (6.03)	50.03 (6.24)	49.10 (7.45)	47.84 (9.66)	46.98 (11.27)	46.84 (11.53)	46.78 (11.97)
50	26.45 (8.12)	21.33 (5.82)	15.88 (5.57)	13.42 (3.88)	11.30 (3.71)	9.90 (4.19)	9.57 (4.38)	9.90 (4.89)
100	34.49 (7.24)	26.18 (6.85)	20.73 (7.94)	15.48 (6.57)	11.63 (4.24)	8.77 (2.36)	7.38 (2.32)	6.98 (2.92)
200	38.60 (3.16)	30.37 (4.85)	26.11 (2.84)	19.47 (3.17)	13.36 (2.17)	10.83 (2.82)	8.64 (3.57)	7.84 (3.67)
500	41.20 (4.06)	33.02 (4.42)	27.97 (3.78)	20.80 (4.47)	14.15 (1.86)	10.37 (1.23)	7.77 (1.32)	6.11 (1.12)

F.2 Full single-asset boundary-violation grid

Table 7: Boundary-violation ratio in percent (mean with standard deviation in parentheses over independently trained seeds).

	$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.001$	
η	Direct policy	Regime detector	Direct policy	Regime detector	Direct policy	Regime detector
0.001	56.54 (1.44)	5.58 (3.02)	12.29 (0.02)	5.78 (2.74)	5.18 (1.53)	4.25 (1.55)
0.01	57.01 (0.69)	3.06 (0.76)	13.29 (1.73)	6.58 (1.83)	6.78 (3.89)	3.85 (0.50)
0.1	56.88 (1.23)	2.59 (0.49)	15.15 (6.27)	3.79 (0.38)	8.64 (2.25)	4.32 (0.01)

F.3 One-dimensional regularization-bias sweep

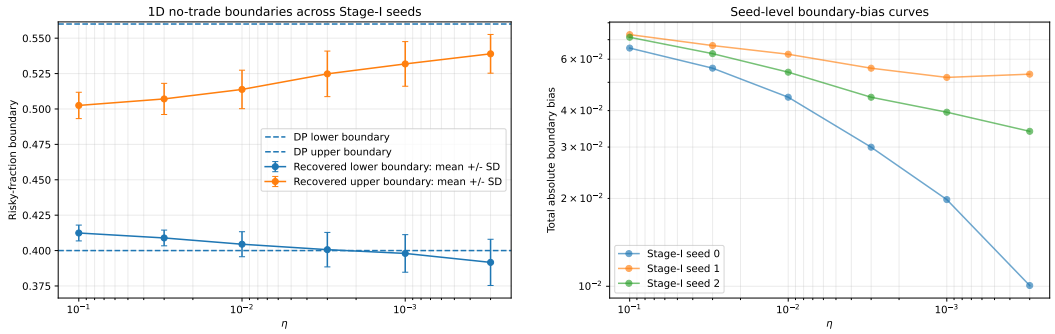
The sweep fixes $\alpha = 0.01$ and uses $\eta \in \{10^{-1}, 3\times10^{-2}, 10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}\}$. For every η, Stage I is retrained independently for three seeds under the matched regularized simulator. Each Stage-I run uses $1{,}500$ training steps, training batch size $2{,}048$, validation batch size $4{,}096$, and validation every 50 iterations. Stage II uses a $1{,}501$-point query grid, $K = 40$ time steps, $4{,}096$ continuation rollouts per replicate, three continuation replicates, and exchange smoothing 3×10^{-4}. The numerical grid-level reference values used in this experiment, reported to four decimals, are $b^-_{\mathrm{DP}} = 0.4000$ and $b^+_{\mathrm{DP}} = 0.5600$; they are not asserted to be exact continuum free-boundary locations.

$$\mathcal{E}_{\mathrm{bdry}}(\eta, s) := \left|\widehat{b}^-_{\eta,s} - b^-_{\mathrm{DP}}\right| + \left|\widehat{b}^+_{\eta,s} - b^+_{\mathrm{DP}}\right|. \tag{148}$$

Table 8 reports the seed-level aggregation. The final column first computes (148) within each seed and then reports its mean and sample standard deviation; it therefore does not benefit from cancellation between seeds. The mean discrepancy decreases at every successive value of η, while its standard deviation increases in the small-η regime. Across all 18 seed–η combinations, both thresholds are bracketed, the estimated cash marginal remains positive, and no action clipping, policy saturation, or numerical projection occurs.

Table 8: One-dimensional regularization sweep across three independently trained Stage-I seeds. Boundary columns report mean $\pm$ sample standard deviation. The final column reports the mean $\pm$ sample standard deviation of the per-seed total absolute discrepancy $\mathcal{E}_{\mathrm{bdry}}(\eta, s)$.

η	Recovered lower boundary	Recovered upper boundary	Total boundary discrepancy
10^{-1}	0.412410 ± 0.005585	0.502502 ± 0.009298	0.069909 ± 0.003877
3×10^{-2}	0.408913 ± 0.005550	0.507098 ± 0.010999	0.061815 ± 0.005517
10^{-2}	0.404485 ± 0.008833	0.513801 ± 0.013593	0.053678 ± 0.008981
3×10^{-3}	0.400651 ± 0.012175	0.524807 ± 0.016086	0.043437 ± 0.012992
10^{-3}	0.397982 ± 0.013269	0.531818 ± 0.015766	0.037100 ± 0.016213
3×10^{-4}	0.391664 ± 0.016319	0.538952 ± 0.013676	0.032452 ± 0.021672



(a) Recovered lower and upper boundaries, with mean ± sample standard deviation across Stage-I seeds.

(b) Seed-level total absolute discrepancy curves.

Figure 10: Detailed one-dimensional regularization diagnostics. The lower boundary reaches the DP reference near $\eta = 3 \times 10^{-3}$ and then slightly overshoots for smaller η, whereas the upper boundary continues to move toward its reference. Thus the robust statement is monotone reduction of the *mean per-seed total discrepancy*, not coordinatewise monotonicity for every seed.

F.4 Three-asset projected controls

The projected controls underlying the three-asset regime partition in Figure 6 are displayed here at zero pairwise correlation. The induced $3^3 = 27$ regime-partition figure itself appears only in the main text.

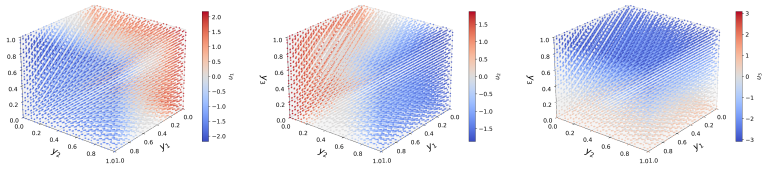


Figure 11: Three-asset projected controls at zero pairwise correlation. The corresponding induced 27-regime partition appears in Figure 6.

F.5 Full three-asset correlation sweep

The following grid expands the one-axis slices to the complete (ρ_{12}, ρ_{13}) design.

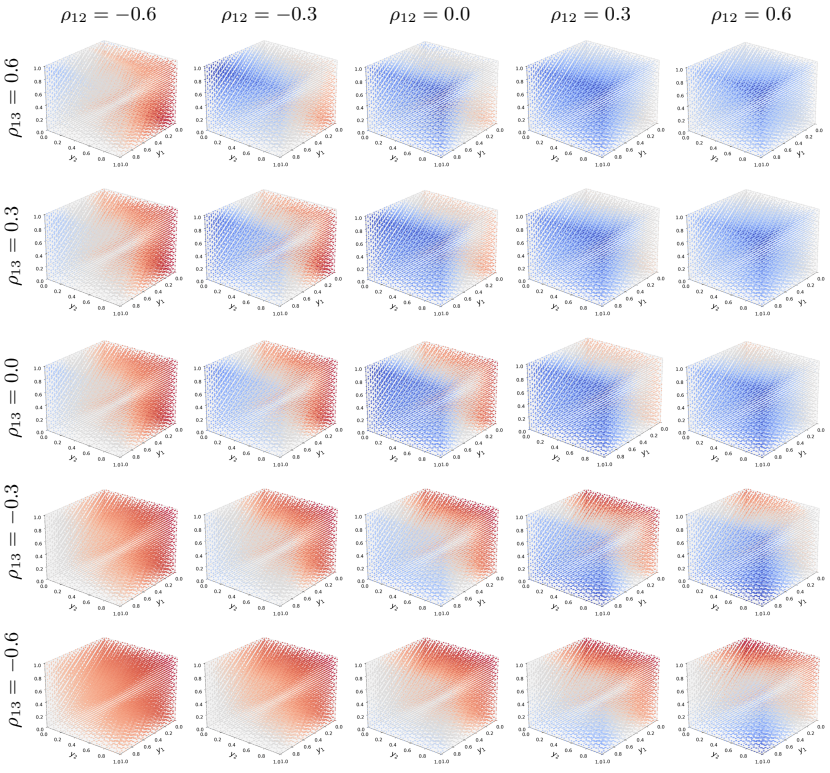
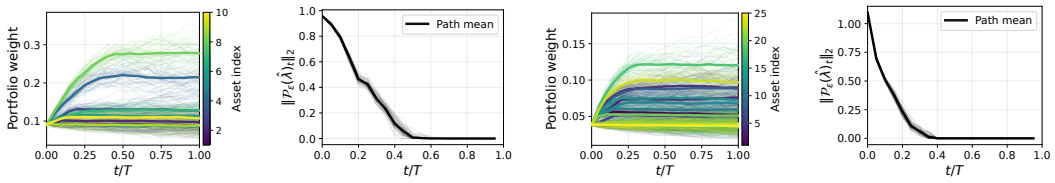


Figure 12: $N = 3$ correlation sweep. Each panel shows $[\mathcal{P}_\eta(\widehat{R})]_1$ under a different pair (ρ_{12}, ρ_{13}). Cross-asset dependence enters through the frozen continuation value gradient rather than the coordinatewise recovery map.

F.6 Full temporal diagnostics

These panels show the full pathwise portfolio weights and trading norms of $\mathcal{P}_\eta^{[1]}(\widehat{R}_{BPTT})$.



(a) $N = 10$: risky weights (b) $N = 10$: projected norm (c) $N = 25$: risky weights (d) $N = 25$: projected norm

Figure 13: High-dimensional temporal behavior. Thin curves are simulated paths and thick curves are pathwise means. Projected trading intensity declines toward maturity for $N = 10$ and $N = 25$.

F.7 Symmetric perturbation-loss diagnostic

Regularized target and empirical scope. Stage III is the exact local maximizer of the η-regularized Hamiltonian for its supplied ratio (Proposition 1). For the true regularized continuation ratio this is the corresponding regularized action; using frozen-policy estimated ratios need not achieve a globally optimal J_η. The diagnostic below accordingly perturbs the *one-shot Stage-III control* and measures differences in J_η, not J_0. It probes average local curvature at the reported perturbation scales, without certifying that the one-shot control is stationary, globally optimal, or equivalent to the Stage-IV refined policy. For each dimension, let $\mathcal{D}_N$ be the fixed symmetric collection of normalized

perturbation directions used in the implementation, so $\mathcal{D}_N = -\mathcal{D}_N$. At the evaluated policy u, we report

$$F_N(\delta) := \mathrm{mean}_{d\in\mathcal{D}_N}\big[J_\eta(u+\delta d) - J_\eta(u)\big],$$

using common random numbers for paired evaluations. Symmetry implies $F_N(-\delta) = F_N(\delta)$ exactly, so the first-order directional terms cancel whether or not u is stationary. When J_η is twice differentiable,

$$F_N(\delta) = \frac{\delta^2}{2|\mathcal{D}_N|}\sum_{d\in\mathcal{D}_N} d^\top \nabla^2 J_\eta(u) d + o(\delta^2).$$

The statistic therefore measures average symmetric perturbation loss or local curvature at the tested scale. It neither tests first-order stationarity nor rules out an individual improving direction.

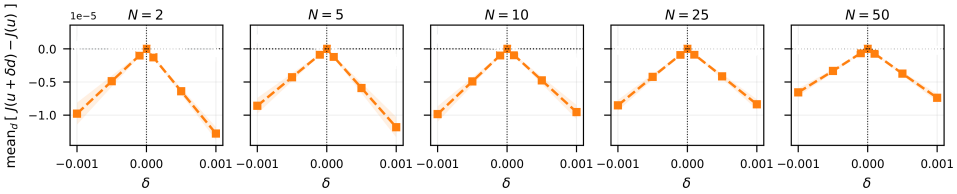


Figure 14: Common-random-number symmetric perturbation-loss diagnostics $\big(\mathbb{E}\big[J_\eta(\mathcal{P}_\eta(\widehat{R}_{BPTT}) + \delta d) - J_\eta(\mathcal{P}_\eta(\widehat{R}_{BPTT}))\big]\big)$ from $N = 2$ to $N = 50$. Because the direction set is symmetric, first-order terms cancel identically; the curves measure average local loss/curvature rather than stationarity or global optimality. The tested objective is the η-regularized J_η, not the original-market J_0 or the fixed-time-refined controller's objective.

F.8 Non-financial spacecraft-rendezvous transfer benchmark

Minimum-fuel HCW problem. For a single chaser in nondimensional Hill coordinates, write $Z_t = (r_t, v_t) \in \mathbb{R}^6$, $r_t = (x_t, y_t, z_t)$, $v_t = (v_{x,t}, v_{y,t}, v_{z,t})$. With reference-orbit mean motion ω, the Hill–Clohessy–Wiltshire drift and velocity-input matrix are

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 3\omega^2 & 0 & 0 & 0 & 2\omega & 0 \\ 0 & 0 & 0 & -2\omega & 0 & 0 \\ 0 & 0 & -\omega^2 & 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0_{3\times 3} \\ I_3 \end{bmatrix}. \tag{149}$$

The continuous minimum-fuel model is

$$dZ_t = AZ_t\,dt + B\,d\Xi_t + \sigma B\,dW_t, \qquad J_0(t,z;\Xi) = \mathbb{E}_{t,z}\left[\Phi(Z_T) + \alpha\sum_{j=1}^{3}\mathrm{TV}_{[t,T]}(\Xi^j)\right], \tag{150}$$

Here Ξ is a finite-variation velocity impulse and W is three-dimensional Brownian motion. The minimized value is

$$V_0(t,z) = \inf_{\Xi} J_0(t,z;\Xi). \tag{151}$$

The docking penalty is

$$\Phi(z) = \frac{1}{2} z^\top Q_f z, \qquad Q_f = \mathrm{diag}(12,12,12,3,3,3). \tag{152}$$

Regularized family and implemented simulator. DPO warm-starts the finite-rate family

$$dZ_t = (AZ_t + Bu_t)\,dt + \sigma B\,dW_t,$$
$$J_\eta(t,z;u) = \mathbb{E}_{t,z}\left[\Phi(Z_T) + \int_t^T \left(\alpha\|u_s\|_1 + \frac{\eta}{2}\|u_s\|_2^2\right) ds\right]. \tag{153}$$

For $\Delta t = T/K$, $A_d = e^{A\Delta t}$, and $\Delta v_k = u_k \Delta t$, the simulator uses exact HCW drift:

$$Z_{k+1} = A_d\big(Z_k + B\Delta v_k\big) + \sigma\sqrt{\Delta t}\, B\xi_k, \qquad \xi_k \sim \mathcal{N}(0, I_3). \tag{154}$$

The warm DPO policy is trained with the regularized discrete cost

$$J_\eta^\Delta = \mathbb{E}\left[\Phi(Z_K) + \sum_{k=0}^{K-1}\Big(\alpha\|\Delta v_k\|_1 + \frac{\eta}{2}\|u_k\|_2^2\Delta t\Big)\right]. \tag{155}$$

Figure 8 instead evaluates the original cost

$$J_0^\Delta = \mathbb{E}\left[\Phi(Z_K) + \alpha\sum_{k=0}^{K-1}\|\Delta v_k\|_1\right], \tag{156}$$

where lower is better.

Minimization-form Pontryagin–Yosida action. For continuation costate $\lambda = \nabla_z V_\eta(t, z)$, the switching score is

$$s := B^\top\lambda. \tag{157}$$

The control-dependent Hamiltonian is

$$h_\eta^{\mathrm{HCW}}(u; s) = s^\top u + \alpha\|u\|_1 + \frac{\eta}{2}\|u\|_2^2. \tag{158}$$

Its coordinatewise optimality condition is

$$0 \in s_i + \alpha\,\partial|u_i| + \eta u_i. \tag{159}$$

Hence the unique minimizer is

$$[\mathcal{P}_\eta^{\mathrm{HCW}}(s)]_i = \frac{[-s_i-\alpha]_+ - [s_i-\alpha]_+}{\eta} = \begin{cases} -(s_i+\alpha)/\eta, & s_i < -\alpha, \\ 0, & |s_i| \le \alpha, \\ -(s_i-\alpha)/\eta, & s_i > \alpha. \end{cases} \tag{160}$$

Equivalently,

$$\mathcal{P}_\eta^{\mathrm{HCW}}(s) = 0 \quad\Longleftrightarrow\quad \|s\|_\infty \le \alpha, \qquad \min_u h_\eta^{\mathrm{HCW}}(u; s) = -\frac{1}{2\eta}\sum_{i=1}^{3}(|s_i|-\alpha)_+^2. \tag{161}$$

The no-thrust cube is exact; the reversed sign reflects cost minimization with symmetric fuel costs.

Continuation recovery and projection depth. After freezing DPO, feedback BPTT at $(t_k, Z_k^{(m)})$ retains state-to-action Jacobians to estimate $\widehat{\lambda}_k^{(m)}$. The fixed-time update is

$$\begin{aligned} s_k^{(m)} &= B^\top\widehat{\lambda}_k^{(m)}, \\ u_k^{(m)} &= \mathcal{P}_\eta^{\mathrm{HCW}}(s_k^{(m)}), \\ Z_k^{(m+1)} &= Z_k^{(m)} + B\big(u_k^{(m)}\Delta t\big). \end{aligned} \tag{162}$$

The virtual loop applies no drift, disturbance, or fuel charge. Finite depth $\mathcal{P}_\eta^{[d]}$ uses at most d maps, stopping at no-thrust entry. Adaptive $\mathcal{P}_\eta^*$ stops at entry or the stated cap/stall condition; it denotes a finite implementation, not an operator limit. The net impulse

$$\Delta v_k^{\mathrm{net}} = \sum_m u_k^{(m)}\Delta t \tag{163}$$

is executed and charged $\alpha\|\Delta v_k^{\mathrm{net}}\|_1$ once, before physical time advances.

Configuration. Table 9 specifies the experiment. Computation uses nondimensional Hill coordinates; only plots use physical units.

Table 9: Configuration for the stochastic HCW transfer experiment.

Quantity	Value
Chasers / state / action dimension	$1/6/3$
Horizon / decision steps	$T = 1.5$, $K = 24$, $\Delta t = 0.0625$
Nondimensional mean motion	$\omega = 1$
Display length / reference mean motion	$100\,\mathrm{m}$ / $1.13 \times 10^{-3}\,\mathrm{s}^{-1}$
Linear fuel coefficient	$\alpha = 0.12$
Quadratic warm-start regularizer	$\eta = 0.30$
Velocity-disturbance scale	$\sigma = 0.20$
Terminal weights	$q_{\mathrm{pos}} = 12$, $q_{\mathrm{vel}} = 3$
Nominal initial state	$(1, -0.55, 0.25, 0, 0.12, -0.03)$
DPO network	3 hidden layers, width 128
DPO optimization	5000 updates, batch 256, learning rate 8×10^{-4}
BPTT continuation budget	2048 antithetic paths per query
Held-out outer rollouts	256
Adaptive-map safeguards	cap 500, switching margin 10^{-5}, stall tolerance 10^{-6}

F.9 Non-financial two-reservoir singular-release transfer benchmark

Original one-sided singular-control problem. This one-sided benchmark has reservoir levels

$$H_t = (H_t^1, H_t^2)^\top \in \mathbb{R}^2$$

and coordinatewise nondecreasing cumulative release $\Xi_t = (\Xi_t^1, \Xi_t^2)^\top$, with dynamics

$$dH_t = (AH_t + b)\,dt + \Sigma\,dW_t - d\Xi_t, \tag{164}$$

where W is two-dimensional Brownian motion. The experiment minimizes

$$J_0(t,h;\Xi) = \mathbb{E}_{t,h}\left[\int_t^T \ell(H_s)\,ds + \alpha\sum_{i=1}^{2}\mathrm{TV}_{[t,T]}(\Xi^i) + \Phi(H_T)\right], \qquad V_0(t,h) = \inf_{\Xi} J_0(t,h;\Xi), \tag{165}$$

with quadratic level-regulation costs

$$\ell(h) = \frac{1}{2}(h-h^\star)^\top Q(h-h^\star), \qquad \Phi(h) = \frac{1}{2}(h-h^\star)^\top Q_f(h-h^\star). \tag{166}$$

Releases cannot push a coordinate below $h_{\min}$; additive disturbances are not projected.

Regularized warm-start family and one-sided Yosida action. Stage-I DPO uses nonnegative rates $u_t \in \mathbb{R}^2_+$:

$$\begin{aligned} dH_t &= (AH_t + b - u_t)\,dt + \Sigma\,dW_t, \\ J_\eta(t,h;u) &= \mathbb{E}_{t,h}\left[\int_t^T \left(\ell(H_s) + \alpha\mathbf{1}^\top u_s + \frac{\eta}{2}\|u_s\|_2^2\right)ds + \Phi(H_T)\right]. \end{aligned} \tag{167}$$

For $\lambda = \nabla_h V_\eta(t,h)$, the control-dependent Hamiltonian is

$$h_\eta^{\mathrm{res}}(u;\lambda) = (-\lambda + \alpha\mathbf{1})^\top u + \frac{\eta}{2}\|u\|_2^2, \qquad u \geq 0. \tag{168}$$

Therefore each coordinate has the unique minimizer

$$[\mathcal{P}_\eta^{\mathrm{res}}(\lambda)]_i = \frac{[\lambda_i - \alpha]_+}{\eta}. \tag{169}$$

Equivalently, if

$$\mathcal{C}_\alpha^{\mathrm{res}} = (-\infty, \alpha]^2,$$

then

$$\mathcal{P}_\eta^{\mathrm{res}}(\lambda) = \frac{1}{\eta}\left(\lambda - \mathrm{Proj}_{\mathcal{C}_\alpha^{\mathrm{res}}}(\lambda)\right), \qquad \mathcal{P}_\eta^{\mathrm{res}}(\lambda) = 0 \iff \lambda \in \mathcal{C}_\alpha^{\mathrm{res}}. \tag{170}$$

Thus $\lambda_i \leq \alpha$ gives exact coordinatewise no-release, directly from the state marginal without cash normalization.

Discrete simulator and original-model evaluation. Set $\Delta t = T/K$, $A_d = e^{A\Delta t}$, and $h_{\mathrm{eq}} = -A^{-1}b$. A release $\Delta\Xi_k \geq 0$ precedes the physical transition:

$$H_{k+1} = h_{\mathrm{eq}} + A_d\big(H_k - \Delta\Xi_k - h_{\mathrm{eq}}\big) + \sqrt{\Delta t}\,\Sigma\xi_k, \qquad \xi_k \sim \mathcal{N}(0, I_2). \tag{171}$$

All held-out controllers use the original discrete objective

$$J_0^{\Delta} = \mathbb{E}\left[\sum_{k=0}^{K-1} \left\{\ell(H_k - \Delta\Xi_k)\Delta t + \alpha \mathbf{1}^\top \Delta\Xi_k\right\} + \Phi(H_K)\right]. \tag{172}$$

The warm policy executes feasible $\Delta\Xi_k = u_k\Delta t$ and adds $\frac{\eta}{2}\|u_k\|_2^2\Delta t$ during training only. Its smooth positive-rate network does not impose exact no-release.

Low-dimensional DP reference. An independent grid-DP reference solves the same discrete $\eta = 0$ problem. For feasible post-release levels $q \leq h$ on a rectangular grid $\mathcal{G}$,

$$\begin{aligned} V_K^{\Delta}(h) &= \Phi(h), \\ V_k^{\Delta}(h) &= \min_{\substack{q\in\mathcal{G}\\ q\leq h}} \left[\alpha\mathbf{1}^\top(h-q) + \ell(q)\Delta t + \mathbb{E}\left[V_{k+1}^{\Delta}\left(h_{\mathrm{eq}} + A_d(q - h_{\mathrm{eq}}) + \sqrt{\Delta t}\Sigma\xi\right)\right]\right]. \end{aligned} \tag{173}$$

Equivalently,

$$V_k^{\Delta}(h) = \alpha\mathbf{1}^\top h + \min_{\substack{q\in\mathcal{G}\\ q\leq h}} \left[-\alpha\mathbf{1}^\top q + \ell(q)\Delta t + \mathbb{E}[V_{k+1}^{\Delta}(F(q,\xi))]\right]. \tag{174}$$

The constrained minimization is a two-dimensional prefix minimum on 101×101 states in $[0.25, 2.25]^2$, using fifth-order tensor-product Gauss–Hermite quadrature and bilinear continuation interpolation. Deployment uses the lower grid cell to keep release targets feasible; a grid hold executes zero release, without snapping to the grid. DP is a numerical reference, not a certified continuous-time optimum.

Frozen-continuation recovery and refinement depth. With the trained policy frozen, feedback BPTT at $(t_k, H_k^{(m)})$ retains state-to-action Jacobians to estimate the regularized continuation gradient $\widehat{\lambda}_k^{(m)}$. Refinement updates

$$\begin{aligned} u_k^{(m)} &= \mathcal{P}_\eta^{\mathrm{res}}\left(\widehat{\lambda}_k^{(m)}\right), \\ H_k^{(m+1)} &= H_k^{(m)} - \delta_{\mathrm{FR}} u_k^{(m)}. \end{aligned} \tag{175}$$

Only intervention feasibility clips each displacement; the virtual loop applies no drift, disturbance, running cost, or release cost. Depths $\mathcal{P}_\eta^{[1]}$, $\mathcal{P}_\eta^{[2]}$, and $\mathcal{P}_\eta^{[4]}$ use at most one, two, and four maps. Adaptive $\mathcal{P}_\eta^{*}$ stops when all coordinates satisfy

$$\widehat{\lambda}_i - \alpha \leq \tau_{\mathrm{sw}},$$

or at a stated stall/cap. The net release executes once in the $\eta = 0$ simulator. Nonnegative increments imply

$$\alpha\left\|\sum_m \Delta\Xi_k^{(m)}\right\|_1 = \alpha\sum_m \left\|\Delta\Xi_k^{(m)}\right\|_1,$$

so aggregation does not discount release costs.

Configuration and evaluation protocol. Table 10 gives the configuration. The six methods in Table 2 share 64 held-out initial states and disturbance paths under $\eta = 0$. Paired differences $J_0 - J_0^{\mathrm{DP}}$ use these common random numbers: adaptive -0.0002 ± 0.0005 is statistically indistinguishable from the numerical DP reference, not evidence of outperforming its optimum.

Table 10: Configuration for the two-reservoir singular-release transfer benchmark.

Quantity	Value
State / action dimension	$2/2$
Horizon / decision steps	$T = 2$, $K = 32$, $\Delta t = 0.0625$
Drift matrix A	$\left[\begin{smallmatrix} -0.35 & 0.08 \\ 0.18 & -0.30 \end{smallmatrix}\right]$
Affine inflow b	$(0.22, 0.17)^\top$
Diffusion matrix Σ	$\left[\begin{smallmatrix} 0.055 & 0 \\ 0.018 & 0.050 \end{smallmatrix}\right]$
Target level $h^\star$	$(0.85, 0.90)^\top$
Nominal initial level	$(1.65, 1.45)^\top$
Initial standard deviation	$(0.08, 0.08)^\top$
Running weight Q	$\left[\begin{smallmatrix} 1.20 & 0.15 \\ 0.15 & 1.00 \end{smallmatrix}\right]$
Terminal weight Q_f	$\left[\begin{smallmatrix} 3.00 & 0.25 \\ 0.25 & 2.50 \end{smallmatrix}\right]$
Linear release coefficient	$\alpha = 0.28$
Quadratic warm-start regularizer	$\eta = 0.45$
Minimum intervention level	$h_{\min} = 0.25$
DPO network	3 hidden layers, width 128, sigmoid output
DPO optimization	2500 updates, batch 384, learning rate 10^{-3}
Warm-policy rate cap	$u_{\max} = 2.5$
BPTT continuation budget	24 antithetic paths per query
Finite refinement depths	$1, 2, 4$
Adaptive refinement safeguards	cap 24, switching margin 2×10^{-3}, stall tolerance 10^{-7}
DP reference	101^2 states, fifth-order Gauss–Hermite quadrature
Held-out outer rollouts	64 common paths

F.10 Non-financial battery-storage transfer benchmark

Discrete storage problem. The policy/DP state is energy and OU price (E_k, P_k), $E_k \in [0, \bar{E}]$; additive cash X_k need not enter the CARA grid state. With $\Delta t = T/K$, $\phi = e^{-\kappa \Delta t}$, and $\sigma_\Delta = \sigma_P \sqrt{(1-\phi^2)/(2\kappa)}$, neural control and DP share

$$P_{k+1} = \bar{p} + \phi(P_k - \bar{p}) + \sigma_\Delta \xi_k, \tag{176}$$

$$E_{k+1} = E_k + u_k^{\text{exec}} \Delta t, \tag{177}$$

$$X_{k+1} = X_k + \left[-a(P_k)(u_k^{\text{exec}})^+ + b(P_k)(u_k^{\text{exec}})^- - \frac{\eta}{2}(u_k^{\text{exec}})^2 \right] \Delta t, \tag{178}$$

where $a(p) = \max\{p + s/2, p_{\min}\}$, $b(p) = \max\{p - s/2, p_{\min}\}$, and u_k^{exec} is the proposed rate projected onto

$$\left[\max\{-\bar{u}, -E_k/\Delta t\}, \min\{\bar{u}, (\bar{E} - E_k)/\Delta t\}\right].$$

Both use a price interval spanning 4.5 stationary standard deviations around $\bar{p}$. Terminal wealth and CARA certainty equivalent are

$$W_T = X_T + b(P_T)E_T, \qquad \text{CE}_\gamma(W_T) = -\frac{1}{\gamma} \log \mathbb{E}[e^{-\gamma W_T}].$$

Frozen-policy feedback BPTT estimates $R = V_E/V_X$, giving the local criterion and unconstrained map

$$h_\eta^{\text{bat}}(u; R, p) = Ru - a(p)u^+ + b(p)u^- - \frac{\eta}{2}u^2, \qquad \mathcal{P}_\eta^{\text{bat}}(R, p) = \frac{[R - a(p)]_+ - [b(p) - R]_+}{\eta}, \tag{179}$$

The interval $b(p) \le R \le a(p)$ maps exactly to zero before rate/capacity projection. Because capacity faces alter the feasible cone, interior metrics exclude $E = 0, \bar{E}$. Here R is the energy-to-cash marginal ratio. This local-criterion map is not an established exact finite-step Bellman maximizer.

Configuration. We use $T = 1$, $K = 24$, $\bar{p} = 1$, $\kappa = 3$, $\sigma_P = 0.35$, spread $s = 0.10$, $p_{\min} = 0.05$, $\bar{E} = 1$, $\bar{u} = 4$, $\eta = 0.05$, $\gamma = 2$, and initial state $(X_0, E_0, P_0) = (2, 0.5, 1)$. The DP uses a 101×101 energy–price grid and nine Gauss–Hermite nodes. The direct policy uses three hidden layers of width 128, 3,000 updates, batch size 1,024, and learning rate 10^{-3}. Online regime detector uses 512 antithetic frozen-policy continuations per query. The ratio student uses three hidden layers of width 128, 30,000 teacher-labelled states, and 3,000 updates. Each of seeds $800, 801, 802$ is evaluated with 1,024 common-random-number outer paths; region maps are evaluated at $t/T = 0.5$.

Table 11: Non-financial battery-storage transfer benchmark, mean across seeds 800–802. Parentheses in CE and latency are sample standard deviations across runs. Geometry is evaluated against the held-out grid-DP policy at $t/T = 0.5$; the interior metric excludes $E = 0$ and $E = \bar{E}$. DP variation comes only from the outer evaluation draws.

Method	CE (run SD)	Regime acc.	Interior hold IoU	ms/state (run SD)
DP	2.502085 (0.002606)	1.0000	1.0000	0.0039 (0.0001)
Direct policy	2.500904 (0.003384)	0.8777	0.0093	0.0065 (0.0002)
Online regime detector	2.500602 (0.002943)	0.9802	0.8743	0.5384 (0.0246)
Distilled regime detector	2.502134 (0.002600)	0.9857	0.8838	0.0114 (0.0030)

Seedwise geometry and teacher retention. Across seeds, direct/online/distilled DP-reference regime accuracies are 0.8777 ± 0.0011, 0.9802 ± 0.0006, and 0.9857 ± 0.0004; interior hold IoUs are 0.0093 ± 0.0074, 0.8743 ± 0.0058, and 0.8838 ± 0.0034. Against the online teacher, the student's interior regime accuracy is 0.9919 ± 0.0016 and hold IoU is 0.9290 ± 0.0134, indicating active-set retention. Paired terminal wealth places the online detector below DP in all three seeds and not uniformly above direct control. The student exceeds its teacher in these comparisons with slightly improved geometry, suggesting empirical denoising without a general improvement guarantee. Its CE of 2.502134 versus DP's 2.502085 does not establish superiority given grid error, Monte Carlo error, and seed variation.

Table 12: Paired common-random-number terminal-wealth differences in the battery benchmark. Each entry is mean difference with Monte Carlo standard error in parentheses. These are paired terminal-wealth diagnostics, not paired certainty-equivalent tests.

Seed	regime-corrected policy − DP	regime-corrected policy − direct policy	regime detector − distilled
800	−0.001766 (0.000334)	−0.000749 (0.000419)	−0.001717 (0.000345)
801	−0.001076 (0.000279)	0.000241 (0.000334)	−0.001442 (0.000281)
802	−0.001344 (0.000282)	−0.001141 (0.000330)	−0.001269 (0.000285)

F.11 Online evaluation cost

Table 13: Online cost of one core policy call across dimensions. A batched pass contains 2,560 query states and includes one continuation simulation, one feedback-BPTT marginal-value estimate, and one dead-zone map. These numbers do not include repeated fixed-time regime refinement closure.

N	Direct policy $u_{\theta^\star}$		Regime detector $\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})$	
	batch (s)	decision (ms)	batch (s)	decision (ms)
2	0.064	0.025	2.44	0.953
5	0.070	0.027	2.34	0.914
10	0.068	0.027	2.39	0.934
25	0.073	0.029	2.31	0.902
50	0.073	0.029	2.53	0.988

Table 14: Stage-II wall-clock time per query state under different continuation budgets. Research implementations are not otherwise optimized.

Hardware	Configuration	K	M_{MC}	Time/query (s)
CPU (Intel Xeon)	Base	16	256	0.03
	Scalability A	50	1024	0.25
	Scalability B	100	4096	1.80
GPU (NVIDIA A100)	Base	16	256	0.018
	Scalability A	50	1024	0.04
	Scalability B	100	4096	0.17

F.12 Default two-asset configuration

The following is the common low-dimensional reference setting unless otherwise stated.

Table 15: Configuration for the two-asset transaction-cost benchmark and the low-dimensional DP/QVI reference.

Category	Parameter	Value
Problem	Horizon T	1.0
	Time steps K	20
	Risk-free rate r	0.02
	Drifts (μ_1, μ_2)	$(0.08, 0.10)$
	Volatilities (σ_1, σ_2)	$(0.20, 0.25)$
	Risk aversion γ	3.0
	Transaction cost α	0.03
	Random seed	800
Regime detector	Regularization η	0.05
	No-trade classification tolerance	0.01
	Warm-up batch size	256
	Continuation rollouts M_{MC}	1024
DP/QVI	Position-grid size	60
	Cash-grid scaling factor	1000
	Newton tolerance	10^{-7}
	Maximum Newton iterations	40

F.13 Configuration for the high-dimensional objective experiment

The objective experiment in Figure 7 uses $N \in \{5, 10, 20, 50\}$ and the long-only/no-borrowing closed simplex $x \geq 0$, $y_i \geq 0$. Initial gross wealth is one, with cash weight 0.20 and risky weight 0.80. Assets are split into two sectors: the initial risky sector totals are $(0.08, 0.72)$, while drifts are generated so that the frictionless Merton target has sector totals $(0.70, 0.10)$ under the dimension-specific covariance matrix. This construction keeps the initial state away from the target while preserving feasibility across dimensions.

Table 16: Configuration used for the reported high-dimensional objective comparison. Method-specific implementation diagnostics are logged separately.

Category	Parameter	Value
Problem	Horizon T	1.0
	Time steps K	20
	Risk-free rate r	0.045
	Risk aversion γ	3.0
	Transaction cost α	0.005
	Recovery regularization η	0.01
	Initial cash weight	0.20
	Initial risky sector totals	$(0.08, 0.72)$
Market design	Volatility	0.20 per asset
	Correlation	fixed-seed positive-definite factor construction
	Correlation strength parameter	0.70
	Frictionless target sector totals	$(0.70, 0.10)$
Continuation/evaluation	Held-out outer paths	256
	BPTT inner continuations M_{MC}	2048
	Common random numbers	yes
	Antithetic sampling	yes
Refinement/distillation	Finite depths	$1, 2, 3$
	Adaptive cap $M_{\max}$	500
	Position stall tolerance	10^{-4}
	Teacher outer paths per depth	64
	Student train/validation split	disjoint by teacher path ID

F.14 High-dimensional regime recovery against a separable CARA DP oracle

Purpose and scope. The main high-dimensional CRRA experiment in Figure 7 evaluates J_0 and executed activity, because a coupled N-asset DP/QVI boundary is not available for $N \geq 5$. To test *regime recovery itself* at high dimension without treating objective performance as a proxy for geometry, we use a separate controlled portfolio benchmark in which a reliable held-out DP regime oracle remains available at $N = 50$. As in the battery transfer benchmark (Appendix F.10), this supplementary experiment uses CARA utility; it is not used to replace or redefine the main CRRA objective experiment.

Separable oracle construction. There is one global risk-free cash account x, unrestricted borrowing/lending, long-only risky holdings $y_i \geq 0$, independent risky returns, and sell-only proportional transaction cost α; purchases are made at face value. We use

$$U(W) = -e^{-aW}, \qquad a = 2, \qquad T = 1, \qquad K = 20, \qquad r = 0.02, \qquad \alpha = 0.05.$$

The two risky-asset types

$$(\mu, \sigma) \in \{(0.06, 0.21), (0.07, 0.24)\}$$

are repeated cyclically across the $N = 50$ coordinates. Under CARA translation invariance, additive trading cash flows, and independent risky returns, the exponential Bellman criterion factorizes as

$$M_k(x, \mathbf{y}) = e^{-A_k x} \prod_{i=1}^{N} m_{k,\tau(i)}(y_i), \qquad A_k = a\, e^{r(T-t_k)}, \tag{180}$$

where $\tau(i)$ denotes the asset type. Consequently the optimal post-trade target and buy/hold/sell label separate coordinatewise. We therefore solve one high-resolution one-dimensional Bellman problem for each distinct asset type rather than constructing a 50-dimensional grid. Each one-dimensional oracle uses 2001 risky-position nodes on $[0, 2]$ ($\Delta y = 10^{-3}$), seven Gauss–Hermite nodes, and may jump from the current grid node to any other post-trade grid node at a decision time. Thus the reference is a high-resolution discrete DP oracle, subject to its time/grid/quadrature discretization, rather than an asserted continuous-time exact boundary.

The frozen test set contains 1000 global states. Decision times are balanced over the $K = 20$ dates, cash is sampled on $[0, 0.5]$, and each risky coordinate is sampled on $[0.05, 1]$, leaving a substantial buffer to the upper oracle-grid boundary. This gives $1000 \times 50 = 50{,}000$ held-out coordinate labels. The oracle labels are used only for evaluation. Every learned baseline is instead one *global* 50-output network: it sees all risky coordinates jointly and is not supplied with asset-block labels or separate per-asset models.

Neural and ratio baselines. The direct policies use three hidden layers of width 256. DPO samples a uniformly random decision time at each optimizer step and maximizes the mean statewise CARA certainty equivalent using four antithetic market paths per training state; boundary-state augmentation places one randomly selected risky coordinate on $y_i = 0$ with probability $1/7$. The regularized continuation uses $\eta = 10^{-3}$ and the BPTT ratio uses $M_{\mathrm{MC}} = 2048$ antithetic continuations per query. Coordinatewise magnitude masks use dollar-order thresholds $\tau \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. The BPTT hold gate retains the sign and magnitude of $u_{\theta,0}$ outside the ratio-implied hold set, whereas $\mathcal{P}_\eta(\widehat{R})$ reconstructs the action directly from the estimated ratio.

For oracle label $\ell_{s,i} \in \{-1, 0, 1\}$ and predicted label $\widehat{\ell}_{s,i}$, we report

$$\mathrm{RegAcc} = \frac{1}{nN} \sum_{s=1}^{n} \sum_{i=1}^{N} \mathbf{1}_{\{\widehat{\ell}_{s,i} = \ell_{s,i}\}}, \tag{181}$$

$$\mathrm{ActionRecall} = \mathbb{P}(\widehat{\ell} \neq 0 \mid \ell \neq 0), \qquad \mathrm{HoldRecall} = \mathbb{P}(\widehat{\ell} = 0 \mid \ell = 0). \tag{182}$$

Regime accuracy is therefore the full three-class buy/hold/sell score, while the two recalls separately diagnose missed interventions and missed holds.

What the oracle comparison shows. Results are reported in Table 1 of the main text. The two unmasked neural policies u_θ are active on every evaluated coordinate: their action recall is 1, but their hold recall is 0. Coordinatewise magnitude masking creates the expected precision–recall tradeoff. At $\tau = 10^{-3}$ it retains most true interventions but recovers few holds; at $\tau = 10^{-2}$ it recovers most holds but misses most interventions; and at $\tau = 10^{-1}$ it degenerates to the all-hold rule. Hence a single magnitude cutoff does not recover both sides of the switching classification.

In contrast, across five training seeds the feedback-BPTT realization $\mathcal{P}_\eta(\widehat{R}_{\mathrm{BPTT}})$ attains buy/hold/sell regime accuracy 0.8663 ± 0.0007 (range $[0.8653, 0.8668]$), together with action recall 0.8979 ± 0.0006 and hold recall 0.8301 ± 0.0012. The BPTT hold gate has the same action/hold recalls but lower three-class accuracy, 0.8507 ± 0.0054, because, after deciding that a coordinate is active, it delegates the buy-versus-sell sign to $u_{\theta,0}$ rather than reconstructing that sign from the ratio. The distilled ratio $\widehat{R}_\theta$ retains substantial recovery performance, with regime accuracy 0.7651 ± 0.0092, action recall 0.8322 ± 0.0127, and hold recall 0.6883 ± 0.0102. By contrast, the PINN ratio $\widehat{R}_{\mathrm{PINN}}$ does not recover the oracle partition in this configuration: its action recall is 0.0015 ± 0.0007 while its hold recall is 1.0000 ± 0.0000, making it close to an all-hold classifier. Thus the high-dimensional recovery result is specific to the feedback-BPTT ratio path and its distilled approximation rather than a generic property of every ratio estimator.

Interpretation and limitation. Table 1 closes a narrower empirical gap than the main objective experiment: it gives a direct quantitative regime-recovery test at $N = 50$ against held-out DP labels, instead of inferring geometry from J_0 or trading activity. It deliberately obtains those labels through CARA separability and diagonal covariance. Accordingly, it does *not* claim a 50-dimensional DP solution of the correlated CRRA problem, nor does it prove Hausdorff convergence or full free-boundary recovery in that nonseparable setting. The coupled CRRA $N \geq 5$ results should therefore be read as high-dimensional performance evidence, while this controlled benchmark provides direct high-dimensional regime-classification evidence.

F.15 Objective, executed active-time fraction, online cost, and refinement depth

The objective comparison in Figure 7 uses the original proportional-cost evaluation objective, while fixed-time intermediate updates are virtual and cost-free and only the final feasible net order is

charged. Hence refinement depth changes both the candidate displacement and the amount of online ratio computation. The experiment does not identify either of these effects in isolation.

Controller comparisons and evaluation protocol. All objective comparisons use the unregularized proportional-cost simulator ($\eta_{\text{eval}} = 0$) and common-random-number outer paths within each dimension. The directly trained policies $u_{\theta,0}$ and $u_{\theta,\eta}$ form the neural-policy baselines. The magnitude masks act *coordinatewise* on $u_{\theta,0}$:

$$u_i^{\text{thr}} = u_{\theta,0,i}\,\mathbf{1}_{\{|u_{\theta,0,i}|\geq\tau\}}, \qquad \tau \in \{10^{-3}, 10^{-2}, 10^{-1}\}.$$

The BPTT hold gate instead sets $u_{\theta,0,i}$ to zero when its estimated ratio lies in $[1-\alpha_i, 1]$. These baselines suppress selected coordinates of the raw policy but do not reconstruct the active-region action magnitude. PINN, frozen-policy BPTT, and the supervised student R_θ each supply ratios to $\mathcal{P}_\eta$; student teacher trajectories are disjoint from held-out evaluation paths (Appendix A.2). Objective bars include depths $[1], [2], [3]$ and adaptive $[\star]$. To avoid crowding, the five activity curves show only $u_{\theta,\eta}$, the BPTT hold gate, and adaptive PINN, BPTT, and student-ratio controllers.

Active-time measurement and interpretation. For the feasible net trading rate u_s^{exec} at normalized physical time s, define the fraction of the horizon with an executed trade by

$$A(\pi) := \mathbb{E}_\pi\left[\int_0^1 \mathbf{1}_{\{\|u_s^{\text{exec}}\|_\infty > 0\}}\, ds\right]. \tag{183}$$

With P outer paths and K uniformly spaced physical decisions, the corresponding empirical fraction is

$$\widehat{A}(\pi) = \frac{1}{PK}\sum_{p=1}^{P}\sum_{k=0}^{K-1} \mathbf{1}_{\{\|u_{k,p}^{\text{exec}}\|_\infty > 0\}}. \tag{184}$$

We use the same outer market-noise bank within each dimension for controller comparisons. The indicator is evaluated after feasibility handling and net-order construction; intermediate virtual refinement actions never enter it. No-Action is excluded from the 18-method objective comparison. The figure displays activity curves for five preselected controllers, while all 18 activity means are retained in the reported data. Activity uncertainty is not shown; quantifying it requires pathwise activity measurements and paired differences, not the outer-path standard errors of J_0.

An inactive executed order is not necessarily an interior ratio-wedge hold: actions may also be blocked by portfolio constraints, termination safeguards, or a failed-query fallback. This whole-portfolio indicator is active if *any* coordinate trades; it measures neither turnover nor the number of traded assets. The regularized direct policy is active throughout the measured horizon at all four dimensions. Adaptive BPTT refinement has measured activity fractions $0.5201, 0.5285, 0.6701, 0.7354$, respectively, with higher displayed sample-mean J_0. Adaptive PINN trades on still smaller fractions (0.1777–0.3416), yet has lower mean objective than the regularized direct policy; the distilled adaptive controller is fully active at $N = 20$. Several coordinatewise magnitude masks also retain objective means close to their neural-policy baseline. These observations do not imply that minimal activity maximizes utility, that an optimal intermediate activity fraction has been identified, or that depth alone explains the objective differences. Objective error bars estimate outer-path standard errors, not variability across training seeds or paired uncertainty for policy gaps.

Figure 15(a) therefore plots held-out J_0 against the measured complete decision time per state. For a ratio-based controller, this timing covers the state input through ratio evaluation, all required fixed-time re-queries and stopping checks, feasibility handling, and construction of the final net order. It is distinct from a single core BPTT call or a pure network-forward benchmark. The student R_θ replaces online BPTT by one feed-forward ratio evaluation per queried virtual state, so its latency should be interpreted as amortization of the teacher rather than as an additional ratio-recovery theorem.

Panel (b) reports the disjoint termination labels defined in Appendix A.5. The presence of depth-limit or stall events is important when interpreting $[k]$ and $[\star]$: a finite refinement endpoint is not synonymous with entry into the analytical hold wedge. Likewise, a small observed residual is distinct from the unknown true ratio error and does not certify global boundary accuracy or oracle accumulated-trade optimality.

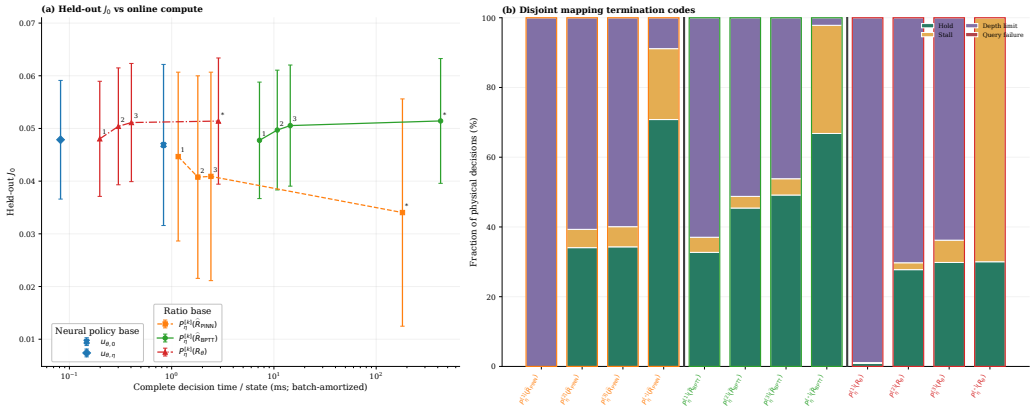


Figure 15: Diagnostics accompanying the high-dimensional objective experiment. **(a)** Held-out original-model objective J_0 versus measured complete online decision time per state (batch-amortized); labels $1, 2, 3, \star$ denote fixed-time mapping depth. Direct neural policies require only a feed-forward policy decision, whereas PINN, BPTT, and student-ratio methods include their actual ratio-map deployment cost. **(b)** Fraction of physical decisions ending in numerical hold, stall, finite depth limit, or ratio-query failure. The categories are mutually exclusive deployment diagnostics and are not certificates of singular-optimal boundary entry.

The magnitude-threshold baselines in the main figure are coordinatewise, and the BPTT hold gate is also coordinatewise. Their thresholds are fixed before the final held-out comparison rather than selected from the displayed test objective. These ablations suppress coordinates of the *unregularized* base $u_{\theta,0}$ which is directly trained over the original problem and consequently isolate the influences of η.

F.16 Limitations and scope of the theory

Numerical target and admissibility. BPTT differentiates the declared simulator and payoff. Raw Gaussian Euler steps preserve neither long-only holdings nor positive liquidation almost surely; samples do not establish near-zero CRRA integrability. Boundary rules, exit payoffs, and failed-query fallbacks belong to that specification. Post-trade feasibility does not ensure subsequent solvency (Appendix A.3, Appendix A.4).

Finite learned refinement. Theorems C.3–C.4 require a positive frozen-policy cash marginal, bounded curvature, feasible virtual segments, and effective virtual-step bounds, plus feasible charged net execution; optimal-value positivity does not supply these conditions. It bounds the best queried residual, not the last iterate or optimal boundary. Noisy steps need not increase W, and tolerance, stall, cap, and failure events differ. Performance improvement also requires control of the final execution-fee correction, market residual, and signed baseline correction. Virtual wedge entry may be lost after the cash deduction. Aggregate objective comparisons do not establish these local conditions or bound adaptive-query errors and failure-event losses.

Continuous-time recovery and oracle limits. Continuous one-shot recovery assumes a classical fixed-policy value and full-horizon admissibility; nondiffusive cash makes regularity substantive. Discrete-to-continuous identification requires controlling b_Δ, and warm-start-to-ratio accuracy requires the stated coverage and stability. Optimal-value convergence implies strict-interior hold recovery only with concavity, smoothness, positivity, and margin conditions, not free-boundary or cumulative-control convergence. Oracle rate separation requires R1–R4, viable comparison flows, charged-endpoint wedge entry, and a net-trade certificate, unverified for each experimental market. Local ascent and wedge entry do not certify optimal cumulative trades; finite refinement does not construct a continuous singular regulator. No dimension-uniform constants or implemented-method singular-limit rate are asserted.

Financial and transfer scope. The financial foundations retain the stated duality, closedness, and admissibility hypotheses, and Euler comparisons are not exact-market gap certificates. Transfer results concern specific linear-cost structures. Battery tests use a low-dimensional discrete regularized problem with capacity-dependent feasible cones; three-seed value/latency estimates do not establish superiority to grid DP. Financial boundary-bias results likewise cover one asset and three seeds.