Structure-Guided Reinforcement Learning

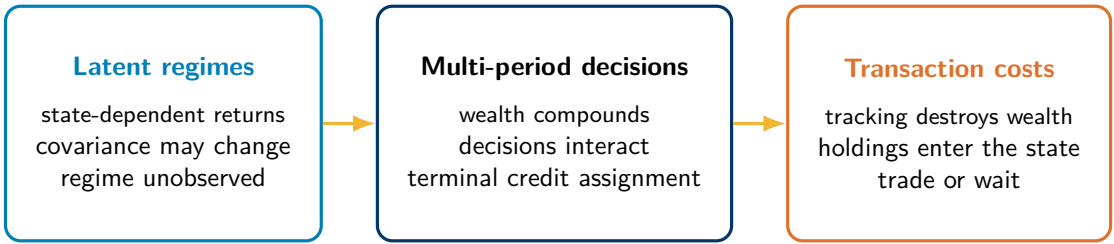
for Partially Observed Multi-Period Portfolio Optimization with Transaction Costs

Takahiro Kobayashi

33rd International Congress of Actuaries — ICA2026
Tokyo — 10 November 2026

Filtering → control structure → selective learning → costly implementation

Why Is Dynamic Portfolio Optimization Difficult?



The controller must infer the market state, choose risk, and decide when rebalancing is worth its cost.

Actuarial relevance: repeated rebalancing and long-horizon targets matter for insurers and pension funds; this paper is a portfolio-control building block, not a complete ALM model.

Dynamic Programming Is Natural — but Grids Grow Quickly

Bellman benchmark

$$V_t(s) = \min_a \mathbb{E}[r_t(s, a, s') + V_{t+1}(s') \mid s, a].$$

- Converts a multi-period problem into backward recursions.
- Highly effective when the state and action spaces remain manageable.

State expansion

$$s_t = (t, X_t, p_t) \quad \longrightarrow \quad s_t = (t, X_t, p_t, \boldsymbol{w}_t).$$

- With costs, current holdings are part of the state.
- For d risky assets, both holdings and actions are multidimensional.

The curse of dimensionality is a numerical representation problem, not a failure of Bellman's principle.

From Dynamic Programming to Structure-Guided RL

Precommitted objective $\Rightarrow$ fixed-ω embedding

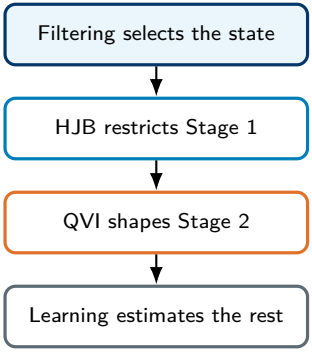
$$\min_u \operatorname{Var}(X_T^u) \quad \text{s.t.} \quad \mathbb{E}[X_T^u] = z.$$

$$\Longrightarrow \quad \sup_{\omega \in \mathbb{R}} \inf_u \left\{ \mathbb{E}[(X_T^u - \omega)^2] - (\omega - z)^2 \right\}.$$

Fixed-ω embedding: for fixed ω, the embedded problem admits the dynamic-programming / HJB formulation below.

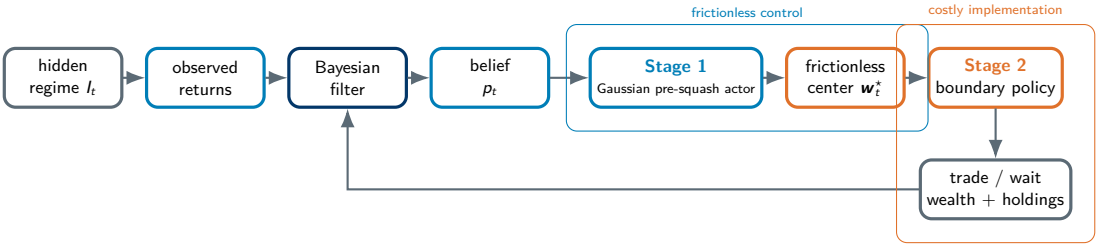
Continuous-time control object

$$0 = v_t + \min_\pi \left\{ \mathcal{G}^\pi v + \Lambda \int \pi(\boldsymbol{u}) \log \pi(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \right\}, \quad v(T, \cdot) = h.$$



RL is simulation-based function approximation here—not a black-box replacement for stochastic control.

Two-Stage Regime-Aware Portfolio-Control Architecture



Stage 1: belief-state, entropy-regularized, precommitted mean–variance control.

Stage 2: asymmetric no-trade boundaries optimized with proportional costs.

Belief-State Reduction under Common Covariance

Shared innovation dynamics

$$\mathrm{d}X_t = \boldsymbol{u}_t^\top \sigma\big(\widehat{\boldsymbol{\rho}}(p_t)\,\mathrm{d}t + \mathrm{d}\widehat{\boldsymbol{B}}_t\big),$$

$$\mathrm{d}p_t = \underbrace{[\lambda_2 - (\lambda_1+\lambda_2)p_t]}_{a(p_t)}\,\mathrm{d}t + \underbrace{p_t(1-p_t)(\boldsymbol{\rho}^{(1)} - \boldsymbol{\rho}^{(2)})^\top}_{b(p_t)^\top}\,\mathrm{d}\widehat{\boldsymbol{B}}_t.$$

- $p_t = \mathbb{P}(I_t = 1 \mid \mathcal{F}_t^S)$.
- The true regime is never supplied to the controller.
- (t, X_t, p_t) is the observable Stage 1 state.
- Cross variation creates the HJB hedging term.

$$\mathrm{d}\langle X, p\rangle_t = \boldsymbol{u}_t^\top \sigma b(p_t)\,\mathrm{d}t.$$

Common covariance gives the theorem-level continuous-time Wonham reduction.

Stage 1: HJB-Restricted Actor and Critic

Belief-state HJB: the action-dependent Hamiltonian is quadratic

$$0 = v_t + a(p)v_p + \tfrac{1}{2}\|b(p)\|^2 v_{pp} + \min_{\pi} \int_{\mathbb{R}^d} \left[\tfrac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Sigma}\boldsymbol{u}\, v_{xx} + \boldsymbol{u}^\top \sigma\{\widehat{\rho}(p)v_x + b(p)v_{xp}\} + \Lambda \log \pi(\boldsymbol{u})\right] \pi(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}.$$

Complete the square ⇒ Gaussian actor

$$\boldsymbol{\pi}^* = \mathcal{N}\left(-\frac{(\sigma^\top)^{-1}[\widehat{\rho}(p)v_x + b(p)v_{xp}]}{v_{xx}}, \frac{\Lambda}{v_{xx}}\boldsymbol{\Sigma}^{-1}\right).$$

Terminal condition ⇒ quadratic critic

$$v(t,x,p) = \frac{(x-\omega)^2}{f(t,p)} + g(t,p) - (\omega - z)^2.$$

Reported implementation: the pre-squash draw $z_k \sim \boldsymbol{\pi}^*$ is executed as $u_k = a_{\max}|X_k| \tanh(z_k/(a_{\max}|X_k|))$ — a bounded push-forward of the Gaussian law, not itself Gaussian.

Theory fixes Gaussianity, wealth-linearity, and wealth-quadraticity; learning fits low-dimensional time/belief coefficients.

Regime-Dependent Covariance: Euler Prediction + Exact Correction

Exact conditional correction

$$\Delta \log \boldsymbol{S}_k \mid I_k = i \sim \mathcal{N}\Big((\boldsymbol{\mu}^{(i)} - \tfrac{1}{2}\,\mathrm{diag}\,\boldsymbol{\Sigma}^{(i)})\Delta t, \boldsymbol{\Sigma}^{(i)}\Delta t\Big),$$

$$p_k = \frac{p_{k|k-1}\phi_1(y_k)}{p_{k|k-1}\phi_1(y_k) + (1 - p_{k|k-1})\phi_2(y_k)}.$$

Euler Markov prediction → exact Gaussian Bayes correction

Moment-matched control layer

$$\widehat{\boldsymbol{b}}(p) = p(\boldsymbol{\mu}^{(1)} - r\mathbf{1}) + (1-p)(\boldsymbol{\mu}^{(2)} - r\mathbf{1}),$$

$$\overline{\boldsymbol{\Sigma}}(p) = p\boldsymbol{\Sigma}^{(1)} + (1-p)\boldsymbol{\Sigma}^{(2)}.$$

$$\mathrm{d}X_t \approx \boldsymbol{u}_t^\top \widehat{\boldsymbol{b}}(p_t)\mathrm{d}t + \boldsymbol{u}_t^\top \overline{\sigma}(p_t)\mathrm{d}\boldsymbol{B}_t.$$

Quadratic Hamiltonian retained as an approximation

Exact Bayesian correction for the adopted discrete-time observation model ≠ exact continuous-time control reduction.

Regime-Dependent Covariance: Exploratory Single-Seed Evidence

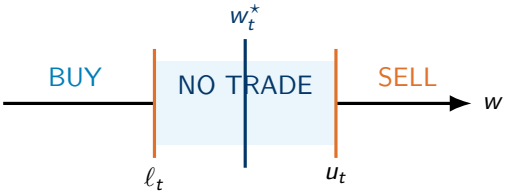
Method	Mean X_T	SD X_T	$\mathbb{P}(X_T \geq z)$
Equal weight	1.078	0.271	0.316
Minimum variance	1.078	0.259	0.312
Approximate HJB RL	**1.132**	**0.172**	**0.342**

Scope

Single training seed — exploratory evidence only.

The result supports operational implementability of the approximate filtering/control architecture, not optimality or multi-seed robustness.

Transaction Costs Create an Intervention Problem



Current holdings enter the state:

$$(t, x, p) \quad \longrightarrow \quad (t, x, p, \boldsymbol{w}).$$

QVI benchmark

$$\min\{-V_t - \mathcal{L}V,\ \mathcal{M}V - V\} = 0.$$

- Continuation: $-V_t - \mathcal{L}V = 0$ and $V < \mathcal{M}V$.
- Intervention: $V = \mathcal{M}V$ after paying proportional cost.
- The contact set defines the trade boundary.

Transaction costs change the geometry of control, not merely the reward function.

Stage 2: Learn the Boundary around the Frictionless Center

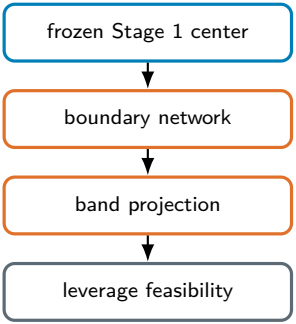
Asymmetric center-relative geometry

$$\underline{\boldsymbol{w}}_k = \boldsymbol{w}_k^\star - \mathrm{softplus}(h_\phi^\ell(o_k)) - \varepsilon,$$
$$\overline{\boldsymbol{w}}_k = \boldsymbol{w}_k^\star + \mathrm{softplus}(h_\phi^u(o_k)) + \varepsilon,$$
$$\boldsymbol{w}_k^{\mathrm{band}} = \min\left(\max(\boldsymbol{w}_k^-, \underline{\boldsymbol{w}}_k), \overline{\boldsymbol{w}}_k\right).$$

Evaluation-time proportional gross-leverage rescaling

$$\boldsymbol{w}_k^{+,\mathrm{eval}} = \frac{\boldsymbol{w}_k^{\mathrm{band}}}{\max\{1,\ \|\boldsymbol{w}_k^{\mathrm{band}}\|_1/1.5\}}.$$

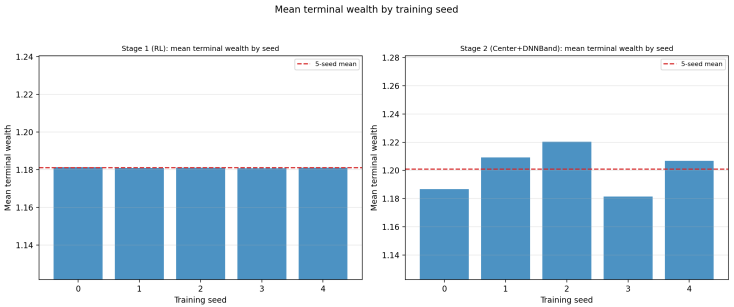
Evaluation-only; applied identically to every Stage 2 comparator; a proportional rescaling toward the origin, not the generic Euclidean projection onto an L^1 ball.



- Direct pathwise terminal objective.
- No Stage 2 critic.
- Not an exact QVI solution.

A trade may result from a band crossing or a leverage-cap violation.

Stage 1: Stable Five-Seed Performance



Intervals are computed across five training-seed-level statistics.

HJB-guided RL

1.1811

mean terminal wealth

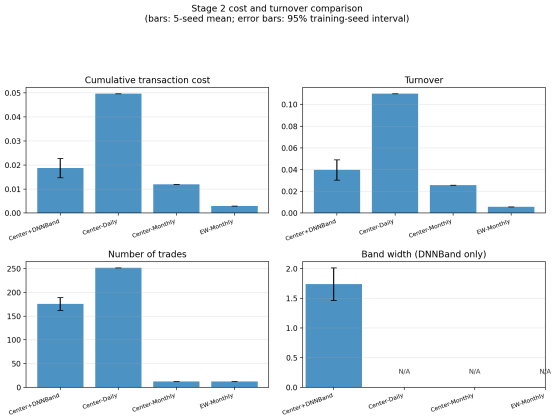
$[1.1808, 1.1814]$

95% training-seed interval

- $\mathbb{P}(X_T \geq z) = 0.566$
- Target MSD $= 0.0714$
- Between-seed SD ≈ 0.000232

Fixed evaluation design: this is initialization stability, not market-distribution robustness.

Stage 2: Cost and Turnover under the Evaluated Policy



"N/A" means that the comparator has no learned no-trade band.

Center+DNNBand	
Net X_T	**1.2009**
Cumulative tc	**0.0187**
Mean trades	**175.8**
Daily center	252.0

Paired vs daily center

$+0.0910$

$[0.0709, 0.1112]$

Attribute trading behavior to the evaluated band-plus-leverage-cap policy, not to the learned band alone.

What Is — and Is Not — Established

Supported by the reported experiments

- Stage 1 is stable across five training seeds.
- Center+DNNBand trades less than daily center rebalancing.
- It has higher displayed net wealth than reported costly comparators.
- Regime-dependent covariance is operationally implementable.

Not established

- Stage 2 dominates Stage 1 under a common objective.
- The regime-dependent approximation is optimal.
- SJM is superior to HMM.
- Direct Boundary solves the transaction-cost QVI.
- The model is a complete ALM system.

$$\text{Stage 2 net} - \text{Stage 1} = 0.0198, \qquad 95\% \text{ interval } [-0.0003, 0.0400].$$

The interval includes zero, and the two stages optimize different terminal objectives.

Actuarial Relevance and Future QVI-Guided Learning

Potential portfolio-control uses

- Regime-aware return-seeking versus defensive sleeves.
- Transaction-aware duration or credit overlays.
- Funding or accumulation targets.
- Transparent rebalancing rules for governance.

Prospective applications only; fixed-income, LDI, liabilities, and solvency capital were not evaluated.

Next methodological step

trained Direct Boundary policy

⇓ structured initialization ⇓

QVI-guided Stage 2 curriculum

- Introduce continuation/intervention residuals gradually.
- Use target networks and obstacle-compatible losses.
- Retain interpretable intervention geometry.

A complete ALM extension would add liabilities, surplus/funding states, capital, liquidity, and eligibility constraints.

Takeaways

1. **Belief-state reduction** converts latent-regime uncertainty into an observable control state.
2. **HJB-guided Stage 1** restricts the pre-squash actor and critic to Gaussian and wealth-quadratic forms.
3. **Transaction costs change control geometry**: holdings enter the state and a no-trade region replaces a point target.
4. **Selective function approximation** learns primarily the residual Stage 2 boundary geometry.

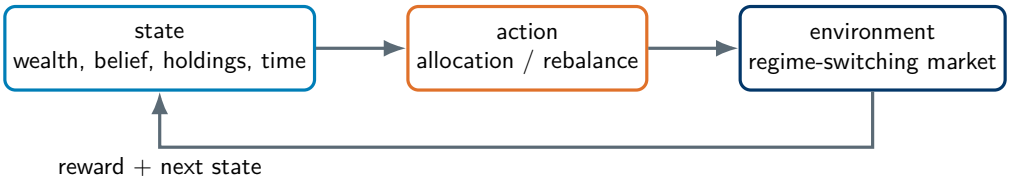
Use learning where theory is incomplete—and preserve theory where it is informative.

Thank you

Questions?

Structure-guided learning for partially observed portfolio control

Backup A1 — What Is Reinforcement Learning?



- A policy maps the information state into an action distribution.
- Simulation supplies trajectories for policy and value approximation.
- In this paper, filtering and stochastic control strongly restrict the state and policy classes before learning begins.

Backup A2 — DP, RL, and the Curse of Dimensionality

$$V_t(s) = \min_a \mathbb{E}[r_t(s, a, s') + V_{t+1}(s') \mid s, a].$$

DP

Conceptual benchmark; exact recursion once a numerical representation is chosen.

Grid pressure

Continuous wealth, belief, holdings, and multidimensional actions multiply grid requirements.

RL representation

Parameterized value and policy functions fitted from simulation.

RL changes the numerical representation; it does not make high-dimensional control free.

Backup A3 — Wonham Filtering Detail

$$
\begin{aligned}
\mathrm{d}\widehat{\boldsymbol{B}}_t &= \sigma^{-1}\mathrm{d}\log \boldsymbol{S}_t - \sigma^{-1}\big(\widehat{\boldsymbol{\mu}}(p_t) - \tfrac{1}{2}\,\mathrm{diag}\,\boldsymbol{\Sigma}\big)\,\mathrm{d}t,\\
\mathrm{d}p_t &= [\lambda_2 - (\lambda_1+\lambda_2)p_t]\mathrm{d}t + p_t(1-p_t)(\boldsymbol{\rho}^{(1)} - \boldsymbol{\rho}^{(2)})^\top \mathrm{d}\widehat{\boldsymbol{B}}_t,\\
\mathrm{d}X_t &= \boldsymbol{u}_t^\top \sigma\big(\widehat{\boldsymbol{\rho}}(p_t)\mathrm{d}t + \mathrm{d}\widehat{\boldsymbol{B}}_t\big).
\end{aligned}
$$

- Common observation covariance makes $\widehat{\boldsymbol{B}}$ Brownian in the price filtration.
- The same innovation drives wealth and belief, producing $\boldsymbol{u}^\top \sigma b(p) v_{xp}$ in the generator.
- This is the theorem-level continuous-time information-state reduction.

Backup A4 — Stage 1 HJB Derivation

Define

$$A = v_{xx}\boldsymbol{\Sigma}, \qquad q = \sigma[\widehat{\rho}(p)v_x + b(p)v_{xp}].$$

Then the control-dependent objective is

$$\int \left[\tfrac{1}{2}\boldsymbol{u}^\top A\boldsymbol{u} + q^\top \boldsymbol{u} + \Lambda \log \pi(\boldsymbol{u})\right] \pi(\boldsymbol{u})\mathrm{d}\boldsymbol{u}.$$

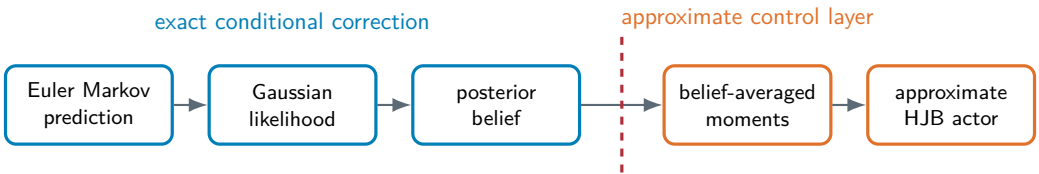
Completing the square gives

$$\Lambda\,\mathrm{KL}(\pi\|\varphi^*) - \tfrac{1}{2}q^\top A^{-1}q - \frac{\Lambda}{2}\log\det(2\pi\Lambda A^{-1}), \quad \varphi^* = \mathcal{N}(-A^{-1}q, \Lambda A^{-1}).$$

With $v = (x-\omega)^2/f + g - (\omega - z)^2$,

$$\boldsymbol{\pi}^* = \mathcal{N}\left(-(x-\omega)(\sigma^\top)^{-1}\left[\widehat{\rho}(p) - b(p)\frac{f_p}{f}\right], \frac{\Lambda f}{2}\boldsymbol{\Sigma}^{-1}\right).$$

Backup A5 — Regime-Dependent Covariance: Exact vs Approximate



Exact conditional on the model

Bayes normalization of specified regime-wise Gaussian emission densities after the predicted probability.

Not established in this paper

Exact continuous-time filter, separation principle, identification of the wealth-belief cross coefficient, or verification theorem.

Backup A6 — Transaction-Cost QVI

$$\min\{-V_t - \mathcal{L}V,\ \mathcal{M}V - V\} = 0.$$

Continuation

$-V_t - \mathcal{L}V = 0,\ V < \mathcal{M}V$: leave holdings unchanged.

Intervention

$V = \mathcal{M}V$: pay proportional cost and move to a post-trade position.

Complementarity

Both residuals are nonnegative and at least one is zero.

Implementation status

Joint learning of the QVI critic, obstacle residual, trigger, and intervention map did not converge reliably under the tested optimization scheme. The reported Direct Boundary policy learns the inaction boundary directly and is not claimed to solve this QVI.

Backup A7 — Full Numerical Tables

Stage 1: common covariance, five-seed means

Method	Mean X_T	Std. X_T	$\mathbb{P}(X_T \geq z)$	Target MSD
Equal weight	1.0543	0.2688	0.304	0.0934
Minimum variance	1.0530	0.2567	0.291	0.0874
HJB-guided RL	**1.1811**	0.2667	**0.566**	**0.0714**

Stage 2: transaction-cost-aware, five-seed means

Method	Net X_T	Cum. tc	Turnover	Trades	Band width	$\mathbb{P}(X_T \geq z)$
Center+DNNBand	**1.2009**	0.0187	0.0397	175.8	1.740	0.490
CenterOnly daily	1.1099	0.0496	0.1100	252.0	N/A	0.426
CenterOnly monthly	1.1229	0.0120	0.0258	12.0	N/A	0.437
Equal weight monthly	1.0372	0.0029	0.0059	12.0	N/A	0.281

Intervals aggregate five training-seed-level statistics; the same 1000 evaluation paths are reused across seeds. The regime-dependent-covariance table is single-seed exploratory: Approximate HJB RL $(1.132, 0.172, 0.342)$ for mean, SD, and target probability.

Backup A8 — Research Positioning

Literature	Latent regime	Belief state	HJB structure	Reg.-dep. cov.	Costs	Learned boundary
Wu & Li	yes	yes	yes	no	no	no
Classical no-trade literature	no	no	yes	no	yes	analytic/numeric
Deep PDE methods	varies	varies	PDE	varies	varies	varies
Mulvey et al.	no	no	yes	no	yes	yes
This paper	**yes**	**yes**	**yes**	**approx.**	**yes**	**yes**

Positioning: filtering + HJB restrictions + regime-dependent approximation + selective boundary learning under costs.

Backup A9 — Stage 1 Actor Signal: Martingale + Full-Horizon Tail

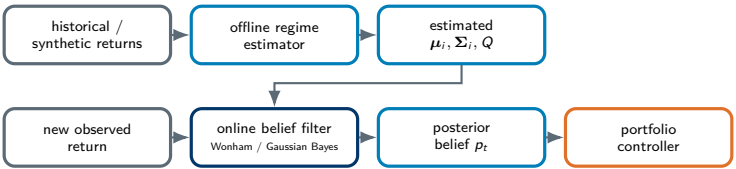
$$G_k = (X_T - \omega)^2 - (\omega - z)^2 - \Lambda \sum_{i=k}^{K-1} H_i \Delta t, \qquad \Delta M_k = V_{k+1} - V_k - \Lambda H_k \Delta t,$$

$$r_k^{\text{tail}} = G_k - V_k, \qquad s_k = 0.7\ \text{Norm}(\Delta M_k) + 0.3\ \text{Norm}(r_k^{\text{tail}}).$$

- The critic target G_k is full-horizon / Monte Carlo, not a one-step TD target.
- ΔM_k is the local martingale-based signal; r_k^{tail} is the full-horizon residual.
- The mixed quantity s_k is an actor policy-gradient cost weight, not a supervised regression target.
- The 0.3 tail weight is an implementation refinement for finite-sample credit assignment, not a control theorem.

The actor log-density uses the exact tanh change-of-variables Jacobian; the entropy term uses a single-sample transformed-entropy estimate — Gaussian entropy plus the sampled log-Jacobian — with the Jacobian correction detached in the gradient.

Backup A10 — Offline Regime Estimation Is Not the Online State



Five-seed common-covariance benchmark → threshold / drawdown heuristic.

Exploratory regime-dependent-covariance experiment → statistical jump model (SJM).

HMM: alternative/comparator, not the reported estimator for the five-seed common-covariance benchmark.

The controller does not receive the ex-post hard regime label or true hidden regime; online allocation uses the posterior belief state p_t.

Backup A11 — Direct Boundary Policy and Evaluation-Time Feasibility

$$\boldsymbol{w}_k^\star = \boldsymbol{u}^\star(t_k, x_k, p_k)/x_k, \qquad \underline{\boldsymbol{w}}_k = \boldsymbol{w}_k^\star - g_k^\ell, \qquad \overline{\boldsymbol{w}}_k = \boldsymbol{w}_k^\star + g_k^u,$$
$$\boldsymbol{w}_k^{\mathrm{band}} = \min\Big\{\max\{\boldsymbol{w}_k^-, \underline{\boldsymbol{w}}_k\}, \overline{\boldsymbol{w}}_k\Big\} \quad \text{(componentwise)}.$$

Evaluation-time feasibility map

$$\boldsymbol{w}_k^{+,\mathrm{eval}} = \frac{\boldsymbol{w}_k^{\mathrm{band}}}{\max\{1,\ \|\boldsymbol{w}_k^{\mathrm{band}}\|_1/1.5\}}.$$

- The band $[\underline{\boldsymbol{w}}_k, \overline{\boldsymbol{w}}_k]$ is learned during training.
- The hard 1.5 gross-leverage rule applies only in evaluation.
- The map is a proportional rescaling, not the generic Euclidean L^1-ball projection.
- The cap may trigger a trade even when the pre-trade weight lies inside the learned band.

Backup A12 — Stage 2 Wealth Accounting and Training Objective

$$TC_k = \kappa\, x_k^- \,\|\boldsymbol{w}_k^+ - \boldsymbol{w}_k^-\|_1, \qquad x_k^+ = x_k^- - TC_k, \qquad U(x_T) = \log\big(\max(x_T, 10^{-10})\big).$$

- The training objective contains the physical proportional transaction cost through the wealth recursion, plus turnover, gap-width, and gross-leverage regularization.
- $\lambda_{\text{turnover}} = 5 \times 10^{-4}$, $\lambda_{\text{gap}} = 10^{-5}$, $\lambda_{\text{gross lev}} = 10^{-5}$.
- The hard 1.5 leverage cap is evaluation-only.

Stage 2 therefore does not optimize the same terminal objective as Stage 1.

Gross terminal wealth uses the same realized trade path but does not deduct transaction costs; it is not a separately re-simulated no-cost counterfactual.

Backup A13 — Selected References

- W. M. Wonham (1964). Some applications of stochastic differential equations to optimal nonlinear filtering. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 2(3):347–369.
- B. Wu and L. Li (2024). Reinforcement learning for continuous-time mean–variance portfolio selection in a regime-switching market. *Journal of Economic Dynamics and Control*, 158:104787.
- Y. Jia and X. Y. Zhou (2022). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23:1–55.
- J. Xiong and X. Y. Zhou (2007). Mean–variance portfolio selection under partial information. *SIAM Journal on Control and Optimization*, 46(1):156–175.
- M. H. A. Davis and A. R. Norman (1990). Portfolio selection with transaction costs. *Mathematics of Operations Research*, 15(4):676–713.
- S. E. Shreve and H. M. Soner (1994). Optimal investment and consumption with transaction costs. *The Annals of Applied Probability*, 4(3):609–692.
- J. M. Mulvey, Y. Sun, M. Wang, and J. Ye (2020). Optimizing a portfolio of mean-reverting assets with transaction costs via a feedforward neural network. *Quantitative Finance*. doi:10.1080/14697688.2020.1729994.
- A. O. Aydınhan, P. N. Kolm, J. M. Mulvey, and Y. Shu (2024). Identifying patterns in financial markets: extending the statistical jump model for regime identification. *Annals of Operations Research*. doi:10.1007/s10479-024-06035-z.