LEARNING HOW TO BE RISK-SENSITIVE: STATE-DEPENDENT RISK FUNCTIONALS FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting risk preferences to the state expands the decisions a risk-sensitive agent can implement. We prove a constructive separation within a specified conditional value at risk (CVaR) family: a stationary state-dependent risk map strictly implements an outer-optimal policy beyond the implementation capacity of every time-dependent, state-independent map. We develop a learning procedure that selects risk maps through empirical planning and separate outer-utility evaluation before fitting a neural critic. This separates the value of a risk preference from the fidelity of its learned execution. Controlled experiments recover optimal utility and verify policy fidelity separately. In a complementary sequential model, adaptive risk exceeds the specified shared-risk baseline under both design and shifted transition laws, matching direct optimization with equal total data at the largest evaluated budget. Exact certificates verify that learned state-only maps implement the target policy under both laws. In financial risk management, a simulated currency-liquidation benchmark with hold-or-sell-all policies shows higher mean revenue than selected shared and time-dependent risk controls. Historical Bitcoin quote replay adds evidence of temporal transfer: a frozen risk map improves lower-tail proceeds over those controls in two later periods on Coinbase's exchange (GDAX in the dataset). Together, these results connect the expressive power of state-dependent risk to a practical procedure for learning explicit valuation rules and verifying their execution.

1 INTRODUCTION AND POSITIONING

Should an agent treat risk in the same way at every state? A preference for downside protection can be appropriate before entering a dangerous branch of a decision process, yet suppress a useful opportunity once the agent reaches a different state. We show that these requirements can conflict: no shared risk functional in a specified spectral family implements the desired decisions, even when its parameters change with time. Allowing the functional to depend on the physical state resolves the conflict.

Our question concerns agents that choose actions through a nested risk valuation. We distinguish the *outer objective*, which measures the value of the resulting behavior, from the *inner risk functional*, which ranks available actions. Learning a risk map means selecting which inner functional to use in each state so that the resulting policy serves the outer objective. The risk map provides an explicit interface between outcome distributions and action rankings: its quantile weights specify the local valuation, and its action margins specify whether the policy executes that valuation. We study the capacity of this structured controller and learn maps with verifiable execution. Direct outer-objective optimization supplies a performance reference; the additional question is which useful decisions admit implementation through local spectral valuations. This makes risk sharing a restriction on decision capacity and execution fidelity a separately testable property.

The distinction exposes two different obstacles. First, parameter sharing can prevent the desired behavior even with exact values. Second, an approximate critic can choose useful actions for the wrong reason: its policy may score well externally while failing to implement its selected risk

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.

functional. We address the first obstacle through a sequential separation theorem and the second through a procedure that selects a risk map from observed outcomes before fitting its neural critic.

Our contributions are:

- **A structural reason to adapt risk.** A stationary state-dependent map implements an outer-optimal policy that every time-only map in the specified spectral class fails to implement. Strict margins make the inner implementation separation robust to bounded perturbations.
- **A separation of selection and execution.** Outcome-based selection avoids ranking candidate risk maps through their fitted neural critics. We evaluate achieved utility and fidelity to the selected functional separately.
- **Evidence from controlled and external environments.** Fresh confirmation recovers optimal utility in all 80 units and faithful implementation in 79. A separate ten-seed study corroborates the advantage over global risk; Currency Exchange demonstrates improved mean revenue over global risk; historical Bitcoin replay demonstrates temporal transfer of a selected map's downside benefit over shared and time-dependent risk controls.

1.1 Positioning in the literature

Coherent, spectral and dynamic risk. Coherent risk measures formalize monotonicity, translation equivariance, positive homogeneity and subadditivity (Artzner et al., 1999). Spectral measures express risk preferences through weighted outcome quantiles (Acerbi, 2002). Markov risk measures support recursive valuation and risk-averse dynamic programming (Ruszczyński, 2010). These foundations specify the admissible functionals and sequential valuation used here; our theorem characterizes the effect of sharing their parameters across physical states.

Risk-sensitive control and distributional RL. Policy gradients address coherent objectives (Tamar et al., 2015); regret and approximation guarantees cover risk-sensitive and distributional RL (Liang & Luo, 2024; Chen et al., 2024), episodic CVaR (Wang et al., 2023), low-rank representation learning (Zhao et al., 2024), and iterated CVaR (Deng et al., 2025). Distributional and quantile critics estimate outcome laws (Bellemare et al., 2017; Dabney et al., 2018b); implicit quantile networks also support distorted-quantile action selection (Dabney et al., 2018a). The objective and Bellman operator must agree (Lim & Malik, 2022); static CVaR decompositions require particular care in policy optimization (Hau et al., 2023). We learn a nested risk map and test the decisions it implements.

Moghimi & Ku (2025) optimize a static spectral measure of total return and derive intermediate risk preferences to interpret the resulting policy. Our question instead concerns the implementation capacity of a specified nested spectral family: we select its local maps against a separate outer objective, then assess their neural execution.

State-dependent preferences and local objectives. State dependence itself is established: Wang & Xu (2023) define state-dependent distortion risk on acts and study preference-robust decisions under ambiguity in distortions and beliefs. Their weighted-CVaR representation provides a foundation for our use of spectral preferences. Likewise, Schlosser (2020) calibrates local subproblems toward target return distributions; Schlosser (2022) uses state-dependent risk aversion for finite-horizon mean–variance optimization. Our distinction is the sequential implementation question: which outer-optimal decisions are excluded by sharing a specified inner spectral functional? We supply a constructive obstruction and a strict state-dependent implementation, then develop and evaluate a procedure for selecting useful maps and checking their execution (Appendix E).

Risk-sensitive offline learning. Rigter et al. (2023) use model ensembles and risk aversion to address epistemic and aleatoric uncertainty in offline RL. Yoo & Woo (2025) protect against risk across plausible alternative models. These address uncertainty in learned dynamics; we examine sharing of the valuation rule. The public Currency Exchange environment of Rigter et al. (2023) supplies our continuous-state application, with matched candidate counts and paired simulator evaluation.

2 STATE-DEPENDENT RISK AS A DECISION MECHANISM

The agent has two distinct decisions to make: which action to take and which risk functional should rank those actions. The outer objective defines what counts as useful behavior. The inner functional determines how the agent trades favorable and adverse outcomes at each decision. Keeping these roles separate allows us to ask whether a restriction on risk preferences prevents a policy that the outer objective would favor.

2.1 VALUATION, SHARING AND IMPLEMENTATION

Consider a finite-horizon MDP with physical state s, stage $t \in \{0, \dots, H-1\}$, reward R and next state S'. Throughout, rewards are maximized. Define the lower-reward CVaR utility and its mixture by

$$C_\alpha(Y) = \frac{1}{\alpha} \int_0^\alpha F_Y^{-1}(u)\, du, \qquad \rho_\lambda(Y) = \sum_{j=1}^{5} \lambda_j C_{\alpha_j}(Y), \quad \lambda \in \Delta^4, \tag{1}$$

where $\alpha = (1/8, 1/4, 1/2, 3/4, 1)$ and Δ^4 is the probability simplex. These are concave reward utilities; their negatives have the corresponding loss-risk convention. Mixing valid CVaR bases supplies the structural constraints used here (Artzner et al., 1999; Acerbi, 2002).

For a deterministic Markov target policy π and a map $\lambda_{t,s}$, set

$$V_H^{\pi,\lambda}(s) = 0, \tag{2}$$

$$Q_t^{\pi,\lambda}(s,a) = \rho_{\lambda_{t,s}}\big(R + \gamma V_{t+1}^{\pi,\lambda}(S') \mid s, a\big), \tag{3}$$

$$V_t^{\pi,\lambda}(s) = Q_t^{\pi,\lambda}(s, \pi_t(s)). \tag{4}$$

This recursion defines a nested valuation by composing conditional one-step functionals (Ruszczyński, 2010).

Definition 1 (Implementation)**.** *A map implements π if $Q_t^{\pi,\lambda}(s, \pi_t(s)) \geq Q_t^{\pi,\lambda}(s,a)$ for every stage, state and alternative action. Strict implementation requires strict inequalities for distinct actions.*

Target-policy continuation values are defined even when implementation fails. Backward induction makes an implemented policy Bellman-consistent for that map. Our impossibility statements allow favorable resolution of ties.

A global map has $\lambda_{t,s} = \lambda$; a time-only map has $\lambda_{t,s} = \lambda_t$; a stationary state-only map has $\lambda_{t,s} = \lambda_s$. Global maps belong to both other classes, but those two classes are not generally nested. State-time maps allow both indices. The two-band adaptive subclass embeds $w \in [1/2, 1]$ as $\lambda(w) = (0, 0, 2w-1, 0, 2(1-w))$.

The fixed outer objective is

$$J(\pi) = \mathbb{E}_\pi \sum_{t=0}^{H-1} \gamma^t u(R_t), \qquad u(r) = r - 5\mathbf{1}\{r < 0\}. \tag{5}$$

An outer-optimal policy maximizes J. The exact outer planner is a decision reference; an empirical utility planner instead estimates rewards and transitions from allowed data. These references evaluate the outer objective directly; risk maps govern the inner valuation.

2.2 A SEQUENTIAL OBSTRUCTION TO SHARED RISK

Consider two decisions that a shared risk preference must support. At state X, Growth offers an equal chance of rewards zero and eight, whereas Flat gives three. Favoring Growth requires sufficiently valuing its high outcome. At another state, R, Risky incurs a small loss and may lead to X, whereas Safe avoids that exposure. A shared functional that preserves Growth also values the chance of reaching X too highly. Changing the shared functional with time does not remove this conflict between states at the same stage.

Define the minimum implementation margin, using the target-policy continuations,

$$m(\pi, \lambda; M) := \min_{t,s,\, a \neq \pi_t(s)} \Big[Q_t^{\pi,\lambda}(s, \pi_t(s)) - Q_t^{\pi,\lambda}(s,a) \Big]. \tag{6}$$

Thus $m \geq 0$ means implementation and $m > 0$ means strict implementation.

Theorem 1 (Sequential separation under risk sharing)**.** *Fix the specified five-base CVaR family. There exist a family $\mathcal{M}$ of stationary three-state, two-stage MDPs and constants $\varepsilon, \delta > 0$ such that every $M \in \mathcal{M}$ has a unique outer-optimal deterministic Markov policy π^*_M under the fixed outer objective, with:*

1. Time-only obstruction: *every time-only map λ satisfies $m(\pi^*_M, \lambda; M) < -\varepsilon$.*
2. Stationary state-only implementation: *a single state-only map $\bar{\lambda}$, common to the family, satisfies $m(\pi^*_M, \bar{\lambda}; M) \geq \delta$.*

Every state–stage decision cell is reachable under π^*_M.

Appendix A gives the full proof: outer optimality and reachability, the uniform time-only obstruction, and a common stationary state-only witness with explicit positive margins.

Mechanism. The target chooses Safe at R, Growth at X, and Hold at an absorbing state Z. Under shared risk, preserving Growth forces a high continuation value at X. The same root functional must then value Risky's chance of reaching that state, contradicting the required Safe comparison. State dependence separates these demands: expectation preserves Growth at X, while lower-tail CVaR protects Safe at R. Physical states recur across stages, making state and time distinct sources of adaptation.

The full construction, proof and quantitative perturbation bound appear in Appendix A. The bound preserves the inner implementation separation under bounded perturbations; preserving outer optimality additionally requires controlling the outer-utility gaps. The theorem covers the whole specified time-only simplex, including global maps. Its strict state-only witness establishes an expressiveness separation for this family.

Sequential coupling creates the separation. For a fixed immediate reward distribution, changing a risk coefficient changes an action ranking. In a sequential problem it also changes the continuation values on which earlier rankings depend. The theorem exploits this coupling: a valuation that supports the desired action at one state constrains how a different state values reaching it. The conflict persists across the entire specified time-only simplex. Increasing the shared candidate bank or improving its optimizer cannot resolve this particular obstruction.

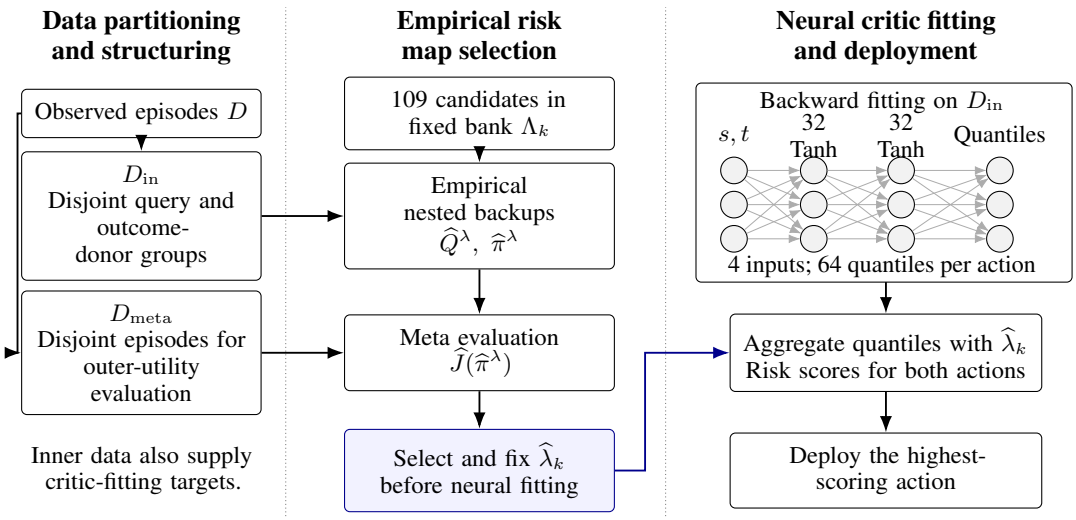
The state-only witness removes the sharing restriction without using a different map at every stage. The result concerns policy implementation: several spectra can produce identical action rankings, while the existence of a strictly implementing map distinguishes the two classes here. The learning question is whether sampled outcomes suffice to select and execute such a map.

Beyond the original parameter interval. A separate exact study evaluates 200 frozen draws on a broader parameter grid within the same transition graph. It certifies 11 separations and finds 90 time-only implementation witnesses; 99 cases remain unresolved by the supplied certificate and witness searches. These results show that both obstruction and shared-risk sufficiency occur in the tested family. The unresolved cases leave the prevalence of separation open. Appendix H.1 retains every draw and specifies the sampling population.

3 Learning useful risk preferences

We learn a risk map from observed trajectories, without true dynamics or a supplied target policy. A fixed bank contains rules assigning spectral weights by state, time, or neither; the declared outer objective judges their decisions.

Why use a risk-map interface? The map specifies how uncertain outcomes are valued at each state, independently of the critic that executes it. This makes discrepancies diagnosable: a faithfully executed map can induce poor decisions, while a useful critic can depart from its selected valuation. We therefore select maps by outcome-based evaluation and check execution separately. The interface supplies an explicit valuation rule against which action rankings can be tested when local valuation is part of the controller specification. Direct optimization remains a utility reference.



Training: quantile regression + cross-fitted score-margin loss; 100 updates/stage, learning rate 0.003.

Figure 1: Selection and deployment architecture for the controlled confirmation. Inner outcomes define each candidate's empirical policy; disjoint meta data select the risk map by outer utility. After the map is fixed, stagewise critics predict 64 quantiles per action, aggregated under that map to choose an action. Nodes depict layers schematically. Quantile regression follows Dabney et al. (2018b), augmented by the score-margin loss. Exact model values are reserved for post-selection utility and fidelity checks.

3.1 Utility and fidelity identify distinct learning outcomes

For example, a critic can choose the externally preferred Safe action while its map ranks Risky higher; correcting the critic then restores Risky. In ten developmental cases, exact execution restored fidelity but lost outer optimality in all ten. Empirical reselection recovered both properties, motivating the procedure tested on new instances, data and initializations in Section 4.1.

3.2 Select the functional before fitting its critic

We separate the question of which map is useful from the question of how well a neural critic executes it. Let Λ_k be the fixed bank for sharing class k. Inner trajectories D_{in} define empirical outcome laws; disjoint meta trajectories D_{meta} supply an independent estimate of the policy's outer objective. The selection rule is

$$\widehat{\lambda}_k = \arg\max_{\lambda \in \Lambda_k} \widehat{J}_{D_{\text{meta}}}\left(\widehat{\pi}^{\lambda}_{D_{\text{in}}}\right). \tag{7}$$

The equation has a simple interpretation: use one dataset to determine what each risk map would do, then a different dataset to judge the consequences.

Backward planning applies each candidate functional to observed rewards plus estimated scalar continuation values and chooses the highest-scoring action. The separate meta model evaluates the resulting policy's outer utility. Ties use a frozen rule; neither neural scores nor exact simulator values determine selection.

Only after selection is locked do we fit a neural quantile critic for the chosen map, using the inner data. Its $N = 64$ heads $z_{t,i}(s,a)$ per action approximate quantiles of reward plus scalar nested continuation. Deployment uses

$$S_t(s,a) = \sum_{i=0}^{N-1} \beta_i(\widehat{\lambda}_{k,t,s}) z_{t,i}(s,a), \qquad a_t \in \arg\max_a S_t(s,a), \tag{8}$$

where $\beta_i(\lambda)$ is the CVaR mixture's mass on quantile bin $[i/N, (i+1)/N]$ (Appendix B). The fixed map supplies the state- or stage-specific weights; we also retain the empirical policy.

3.3 What is learned, and what is checked?

Utility measures whether the policy serves the external goal; *fidelity* checks agreement with the selected map's exact nested optimizer. Both are evaluated exactly in the controlled MDPs after policies are locked.

Execution guarantee (Proposition 2). Fix a risk map λ. Let $Q^{*,\lambda}$ be its exact optimal nested scores and $\Delta_{t,s} > 0$ the gap between its best and second-best actions. For deployed critic scores S, at each decision cell,

$$2 \max_a |S_t(s,a) - Q_t^{*,\lambda}(s,a)| < \Delta_{t,s} \quad \Longrightarrow \quad \arg\max_a S_t(s,a) = \arg\max_a Q_t^{*,\lambda}(s,a).$$

If this holds at every cell, execution is faithful everywhere. This sufficient condition concerns the selected map; outer utility is evaluated separately. Appendix B gives the proof; experiments check fidelity directly.

Each sharing class receives 109 candidates. Adaptive maps use the two-band subclass; shared controls use five-base mixtures. The finite banks are not nested. Neural architectures and fitting budgets are matched across classes; direct utility planners use inner-only or pooled data as stated controls. Appendix B specifies the recursions, critic architecture and losses; Appendix G records the splits and verification.

4 Experimental evidence

We test recovery of the structural advantage and performance where sharing suffices, then examine another sequential model under shifted dynamics. Currency Exchange tests offline selection with continuous states; historical Bitcoin quotes test temporal transfer. Each study has its own objective and protocol. Appendix E.1 maps the comparisons to candidate coverage and validated components; Appendix F adds a slippery CliffWalking control retaining shared risk.

4.1 Can learning recover the structural advantage?

This experiment tests whether outcome-based selection recovers optimal utility and faithful risk implementation where sharing is obstructed, while retaining optimal utility where global risk suffices. Four fresh parameter draws use the necessary family of Theorem 1, and four controls replace the Safe reward 0.9 by 2. Each instance has ten fresh data and initialization seeds. The parameter draws, banks, selection rule, training budget and criteria were frozen before collecting 50,000 inner and 50,000 meta episodes per unit. All 80 units completed and passed provenance, data, selection and checkpoint verification; no unit was replaced or excluded.

Table 1: Fresh confirmation over 40 necessary cases and 40 global-sufficient controls. Optimality is exact outer-utility optimality; fidelity is agreement with the selected map's exact nested optimizer at all six decisions. Empirical and neural policies agree for all 240 method–unit pairs, yielding the same counts, but their confirmation criteria were evaluated separately.

Setting	Risk sharing	Optimal ↑	Faithful ↑	Joint ↑	Regret ↓
Necessary	State-only	**40/40**	39/40	**39/40**	**0**
	Global	0/40	36/40	0/40	0.656667
	Time-only	0/40	**40/40**	0/40	0.333333
Control	State-only	**40/40**	**40/40**	**40/40**	**0**
	Global	**40/40**	38/40	38/40	**0**
	Time-only	22/40	38/40	20/40	0.148250

Adaptive risk wins where state dependence is necessary. State-only selection attains the utility optimum in every necessary case, whereas neither shared control does so in any case (Table 1). Mean paired advantages are 0.656667 over global risk and 0.333333 over time-only risk. The predeclared 95% seed-block bootstrap intervals are $[0.656667, 0.656667]$ and $[0.333333, 0.333333]$: each of the ten seed-block differences is identical. These intervals describe seed variation conditional on the four

realized instances. The theoretical obstruction excludes the full specified shared class in these cases; the empirical comparison evaluates the frozen finite banks.

Adaptation retains optimal utility where sharing suffices. Both state-only and global risk attain optimal utility in all 40 controls. State-only selection is also faithful in every control. Its joint success counts are 39/40 in necessary cases and 40/40 in controls, exceeding the predeclared 36/40 thresholds; both paired lower bounds are positive. Empirical and neural deployment therefore each pass their confirmation criterion. Per-instance joint counts are $(10, 10, 10, 9)$ for necessary cases and $(10, 10, 10, 10)$ for controls. Both direct empirical utility planners attain optimal utility in all 80 cases, so adaptive risk matches these performance references.

The remaining fidelity exception is localized. In unit 37, both deployments choose the outer-optimal Safe action at initial state R, but the selected risk map's exact optimizer prefers Risky. The empirical Safe-minus-Risky score is 0.014800, versus -0.068259 under exact valuation. Thus the mismatch is already present before neural fitting and is distinct from negligible checkpoint roundoff. The frozen choice is retained under the predeclared tie rule. Across all state-only selections, 79/80 policies are jointly faithful and optimal.

What the comparisons establish. The result confirms the correction on fresh data, initializations and parameter draws within the constructed family. Finite-bank coverage explains the controls' time-only deficit: its banks lack an exact optimal map in 20 control units, whereas the global banks contain one. Architecture and training budgets are matched across risk methods, but banks are not nested. Full reproduction details, ties and numerical diagnostics are given in Appendix G.

4.2 State-dependent risk across sequential models and transition shifts

A separate fully observed model has three physical states, two actions and three stages. We evaluate design and shifted transition laws using policies locked before evaluation. Seven methods, three training budgets, ten seeds and two laws give 420 evaluations. Each inner budget has an equal, disjoint meta budget. Lower budgets are dependent, nested prefixes within a seed. This complementary study uses the original selector, which fits a critic for every candidate and scores its neural policy.

Table 2: Mean exact outer utility over ten training seeds in the three-stage model. Columns show inner episode budgets; an equal meta budget is available for selection or pooled direct training. The full three-budget table and paired uncertainty analysis appear in Appendix I. Higher is better.

	Design law		Shifted law	
Method	2,000	12,000	2,000	12,000
State-only risk	9.4455	9.4843	8.3561	8.4163
State-time risk	9.4729	9.4843	8.4046	8.4163
Rich global risk	8.7349	8.7149	7.2856	7.2452
Planner, inner data	9.3931	9.4610	8.2933	8.3824
Neural, inner data	9.4014	9.4610	8.3014	8.3824
Planner, pooled data	9.4612	9.4843	8.3876	8.4163
Neural, pooled data	9.4726	9.4843	8.3994	8.4163

Adaptive risk exceeds the rich global baseline in all displayed conditions. At the largest budget, state-only utility is 9.4843 versus 8.7149 on the design law and 8.4163 versus 7.2452 on the shifted law. Exact certificates separately verify that three frozen learned state-only maps implement the target policy under both laws, while the specified global class cannot. This complements the learning comparison with a verified implementation result.

The pooled direct controls match the adaptive values at the largest budget. The predeclared lower-budget data-efficiency comparison yields uncertainty intervals spanning zero: averaged over the two lower design-law budgets, state-only minus pooled planner is -0.007872 (97.5% paired bootstrap interval $[-0.034192, 0.017471]$), and the corresponding difference against pooled neural utility is -0.013557 ($[-0.051012, 0.017471]$). This study corroborates the global-risk advantage, with relative data efficiency against pooled direct control remaining unresolved. Its earlier selection rule and previously examined shifted law make it complementary evidence to the fresh confirmation.

4.3 LIQUIDATION APPLICATIONS: SELECTION AND TEMPORAL TRANSFER

We test whether outcome-based selection yields useful adaptive risk maps beyond the finite constructions. Currency Exchange (Rigter et al., 2023) is a simulated liquidation task with continuous prices and sale fractions. Our evaluated policies restrict actions to hold or sell all; this is not unrestricted continuous-action learning. The public offline dataset and transition-accounting audit are described in Appendix K. We retain its horizon, rewards and absence of forced terminal liquidation.

Objective and protocol. Here the outer selection target is lower-tail $\mathrm{CVaR}_{0.1}(G)$ of complete episode revenue. This empirical extension pairs a static episodic outer objective with nested inner valuation; the theorem addresses the additive outer utility defined in Section 2. We estimate price dynamics from 8,192 offline episodes using a censored Gaussian autoregression and use a disjoint set of 4,096 episodes for the meta model. Public accounting and the price-model family are known, but drift and volatility are estimated. No reference trajectories enter fitting or selection.

Each risk class receives 109 candidates and 512 meta-model rollouts per candidate. Global and time-only maps use five CVaR bases; state-only maps use 17 constant and 92 sigmoid price-dependent mixtures of expectation and $\mathrm{CVaR}_{0.5}$. Thus candidate counts are matched, but the finite banks are not nested or identical in expressivity. After locking selection, we evaluate on 4,096 fresh, paired simulator episodes. A risk-neutral planner and immediate liquidation provide interpretable controls. The comparison evaluates one development fit per method, with paired uncertainty from fresh evaluation episodes.

Table 3: Currency Exchange revenue on fresh evaluation episodes. Higher is better in both columns. CVaR is the average lower-tail revenue at mass 0.1; values are revenue units. Bold marks the largest point estimate; paired uncertainty intervals are reported in the text.

Method	Mean revenue	$\mathrm{CVaR}_{0.1}$
State-dependent risk	103.560	**92.410**
Rich global risk	103.091	91.855
Time-dependent risk	100.453	91.999
Risk-neutral planner	**137.330**	35.766
Immediate sale	100.084	91.169

Improved mean revenue with competitive downside protection. State-dependent risk improves mean revenue over rich global risk by 0.468 (paired 95% bootstrap interval $[0.342, 0.597]$) and over time-only risk by 3.106 ($[2.820, 3.406]$). Its corresponding CVaR improvements are 0.555 ($[-0.307, 1.325]$) and 0.411 ($[-1.005, 1.652]$): the highest observed tail revenue is accompanied by uncertainty spanning zero in both tail differences. Relative to risk neutrality, adaptive risk increases CVaR by 56.643 ($[53.627, 59.438]$), sacrificing 33.771 in mean revenue. The result quantifies the mean-revenue cost of downside protection. Intervals use 1,000 paired episode bootstrap replicates and quantify evaluation uncertainty conditional on this fit.

An earlier restricted 25-candidate search selected the same constant map for all three classes and produced identical performance. The richer search above selected a nonconstant state map; both developmental outcomes are retained. The external application complements the replicated synthetic evidence with feasible offline selection and a promising return–risk trade-off.

The comparison conditions on one fit, known accounting and the specified price-model family. It establishes a positive paired mean difference against shared risk, with tail differences unresolved. Historical replay next tests whether a selected map's benefit transfers to a later observed period.

Historical-market temporal transfer. Currency Exchange tests selection in an estimated simulator; historical Bitcoin quotes test transfer to a later observed market period. We use one-minute best bids and asks from Dryad[1] on Bitstamp, Bitfinex and Coinbase's exchange (labelled GDAX in the dataset). An agent holds or sells all inventory within 60 minutes, with next-minute bid execution and mandatory terminal liquidation. A 64-state-per-stage empirical model with 32 outcome donors is

[1]Dataset DOI: `https://doi.org/10.5061/dryad.q2bvq83rn`.

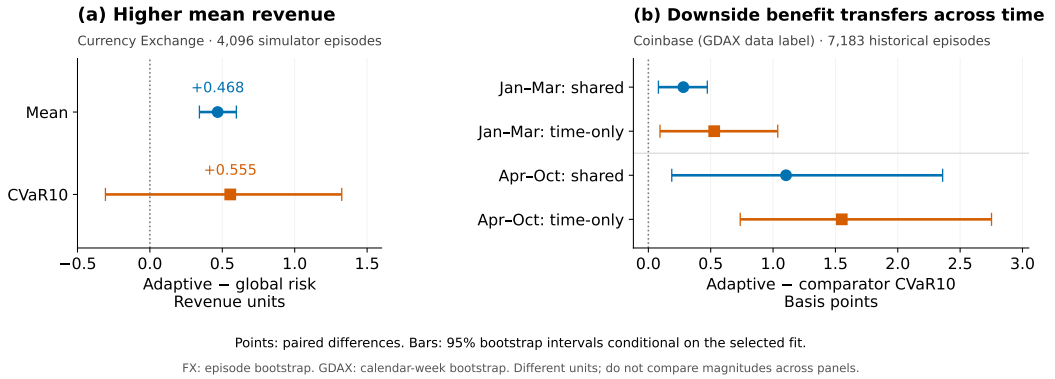


Figure 2: Complementary application evidence for adaptive risk. (a) Currency Exchange improves mean revenue over global risk, with a positive tail point estimate whose interval includes zero. (b) A map selected on 2018 data improves historical Coinbase/GDAX tail proceeds over shared and time-only risk in both fixed 2019 periods. Bars are paired 95% bootstrap intervals conditional on each selected fit: episode resampling for FX, calendar-week resampling for GDAX. Positive differences favor adaptation; panel units differ. Other venues tie, and immediate sale retains stronger GDAX downside protection (Appendix D).

fitted on 512 early-2018 episodes. Separate 2018 episodes select maps by episodic $\mathrm{CVaR}_{0.1}$ before evaluation on all eligible 2019 episodes. This application evaluates empirical planning directly. The state and time banks each contain all 109 shared maps plus 96 varying maps; nesting is explicit, while search budgets differ.

The locked GDAX map improves $\mathrm{CVaR}_{0.1}$ over shared risk by 0.281 bp ($[0.081, 0.473]$) in January–March and 1.104 bp ($[0.188, 2.360]$) in April–October. Across 7,183 episodes, its tail advantages are 0.875 bp ($[0.232, 1.730]$) over shared and 1.231 bp ($[0.621, 2.037]$) over time-only risk. The shared-risk mean difference is 0.048 bp ($[-0.048, 0.149]$). Bitstamp (6,986 episodes) and Bitfinex (7,158) select shared maps and tie. Thus one selected adaptive map transfers a downside benefit to two subsequent periods, complementing the simulator's mean-revenue result. Immediate sale has 0.284 bp stronger GDAX tail proceeds; five training-block refits per venue yield no adaptive tail gain. Temporal transfer of a locked fit and fitting stability are distinct. Appendix D reports these controls, execution stresses and the full protocol.

5 Discussion and Conclusion

Sharing a risk functional across states can exclude useful decisions even when its parameters vary with time. Within the specified CVaR family, a stationary state-dependent map implements the outer-optimal policy while every time-only map fails. Learning recovers optimal utility in 80/80 confirmation units and faithful execution in 79.

The risk map specifies how uncertain outcomes should be valued, separately from the network ranking actions. A good return alone cannot establish that the network follows that rule. Separate selection and execution checks help locate errors in either component. Direct controls also reach the synthetic optimum; the added value is an explicit valuation rule with testable execution.

The applications test empirical planning. Currency Exchange shows higher mean revenue than shared risk, with uncertain tail differences for one fit. A frozen GDAX map improves downside proceeds over shared and time-only risk in two later periods. Other venues tie, immediate sale retains better downside, and refits show fitting sensitivity. Temporal transfer and repeatable discovery of a useful controller are distinct questions.

Broader risk families, stable estimation and faithful execution in larger problems remain to be studied. The central result is that state dependence expands implementable decisions within the specified family, while utility and fidelity provide separate tests of the learned controller.

AI USE STATEMENT

AI tools assisted manuscript drafting and editing, checks of mathematical proofs, and literature-review verification through comparison of bibliographic records with primary sources. They also assisted code debugging.

REPRODUCIBILITY STATEMENT

The theoretical result appears in Section 2.2, with the full construction and proofs in Appendix A. The learning procedure appears in Section 3, with equations and critic details in Appendix B. Appendix G records data generation, seed streams, selection, verification and statistical criteria. Appendix H.1 reports exact certificate coverage. The external-application appendices specify data sources, preprocessing, selection and evaluation protocols. These details distinguish reproduction of model fitting and policy selection from reanalysis of saved evaluation returns. Independent saved-evidence replay verifies 240 selected-map records in the controlled confirmation and selection and return statistics for 327 Currency Exchange candidates. This checks recorded evidence rather than rerunning training or historical-market fitting.

REFERENCES

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, July 2002. ISSN 0378-4266. doi: 10.1016/s0378-4266(02)00281-9. URL http://dx.doi.org/10.1016/S0378-4266(02)00281-9.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, July 1999. ISSN 1467-9965. doi: 10.1111/1467-9965.00068. URL http://dx.doi.org/10.1111/1467-9965.00068.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458, 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Yu Chen, Xiangcheng Zhang, Siwei Wang, and Longbo Huang. Provable risk-sensitive distributional reinforcement learning with general function approximation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7748–7791, 2024. URL https://proceedings.mlr.press/v235/chen24bf.html.

Ziteng Cheng, Anthony Coache, and Sebastian Jaimungal. Eliciting risk aversion with inverse reinforcement learning via interactive questioning, 2025. URL https://arxiv.org/abs/2308.08427v2. Version 2, revised 24 December 2025.

Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1096–1105, 2018a. URL https://proceedings.mlr.press/v80/dabney18a.html.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b. doi: 10.1609/aaai.v32i1.11791.

Zilong Deng, Simon Khan, and Shaofeng Zou. Near-optimal sample complexity for iterated CVaR reinforcement learning with a generative model. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3907–3915, 2025. URL https://proceedings.mlr.press/v258/deng25b.html.

Jia Lin Hau, Erick Delage, Mohammad Ghavamzadeh, and Marek Petrik. On dynamic programming decompositions of static risk measures in markov decision processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51734–51757, 2023. doi: 10.52202/075280-2254. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a264726ebd222124514a32bf0143b83d-Abstract-Conference.html.

Hao Liang and Zhiquan Luo. Regret bounds for risk-sensitive reinforcement learning with lipschitz dynamic risk measures. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1774–1782, 2024. URL https://proceedings.mlr.press/v238/liang24a.html.

Shiau Hong Lim and Ilyas Malik. Distributional reinforcement learning for risk-sensitive policies. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c88a2bd0e793550d0e885aa6e31ca277-Abstract-Conference.html.

Mehrdad Moghimi and Hyejin Ku. Beyond CVaR: Leveraging static spectral risk measures for enhanced decision-making in distributional reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44571–44593, 2025. URL https://proceedings.mlr.press/v267/moghimi25a.html.

Marc Rigter, Bruno Lacerda, and Nick Hawes. One risk to rule them all: A risk-sensitive perspective on model-based offline reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 77520–77545. Curran Associates, Inc., 2023. doi: 10.52202/075280-3389. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f49287371916715b9209fa41a275851e-Paper-Conference.pdf.

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, 125(2):235–261, July 2010. ISSN 1436-4646. doi: 10.1007/s10107-010-0393-3. URL http://dx.doi.org/10.1007/s10107-010-0393-3.

Rainer Schlosser. Risk-sensitive control of markov decision processes: A moment-based approach with target distributions. *Computers & Operations Research*, 123:104997, 2020. doi: 10.1016/j.cor.2020.104997.

Rainer Schlosser. Heuristic mean-variance optimization in markov decision processes using state-dependent risk aversion. *IMA Journal of Management Mathematics*, 33(2):181–199, 2022. doi: 10.1093/imaman/dpab009. URL https://academic.oup.com/imaman/article/33/2/181/6276404.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/024d7f84fff11dd7e8d9c510137a2381-Abstract.html.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35864–35907, 2023. URL https://proceedings.mlr.press/v202/wang23m.html.

Wei Wang and Huifu Xu. Preference robust state-dependent distortion risk measure on act space and its application in optimal decision making. *Computational Management Science*, 20(1):45, 2023. doi: 10.1007/s10287-023-00475-x.

Gwangpyo Yoo and Honguk Woo. Model risk-sensitive offline reinforcement learning. In *International Conference on Learning Representations*, pp. 82343–82362, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/cce0df8a454e6a5d230048befe2ba0fe-Abstract-Conference.html.

Yulai Zhao, Wenhao Zhan, Xiaoyan Hu, Ho-fung Leung, Farzan Farnia, Wen Sun, and Jason Lee. Provably efficient CVaR RL in low-rank MDPs. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 37468–37497, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/a2e3b4132ab2e0b7a21e6e75da7f91a9-Paper-Conference.pdf.

A Formal proof of sequential separation

A.1 MDP family and admissible risk maps

For $q \in [1/2, 3/5]$, let M_q have state space $\mathcal{S} = \{R, X, Z\}$, horizon $H = 2$, discount $\gamma = 97/100$ and initial distribution $\mu_0(s) = 1/3$. Each state has the two actions named below. The table specifies the joint reward–successor law, identical at both stages; a deterministic entry has probability one. At X/Growth, the two reward atoms each have probability $1/2$ and successor X.

State/action	Reward	Next state
R/Safe	$9/10$	R with $1/100$, otherwise Z
R/Risky	$-1/10$	X with q, otherwise Z
X/Growth	$0, 8$ equiprobably	X
X/Flat	3	X
Z/Hold	0	Z
Z/Loss	-1	Z

Set $\mathcal{M} = \{M_q : q \in [1/2, 3/5]\}$ and fix $u(r) = r - 5\mathbf{1}\{r < 0\}$. The outer objective is $J_{M_q}(\pi) = \mathbb{E}_{\mu_0,\pi} \sum_{t=0}^{1} \gamma^t u(R_t)$. The target π^* chooses Safe, Growth and Hold at both stages. The admissible spectra are $\lambda \in \Delta^4$ on the bases $\alpha = (1/8, 1/4, 1/2, 3/4, 1)$. A time-only map is any pair $(\lambda_0, \lambda_1) \in (\Delta^4)^2$, shared across physical states. All inner values below use the target-policy recursion from Section 2, with terminal value zero, whether or not the map implements π^*. Figure 3 summarizes the conflict between the two states before the formal argument.

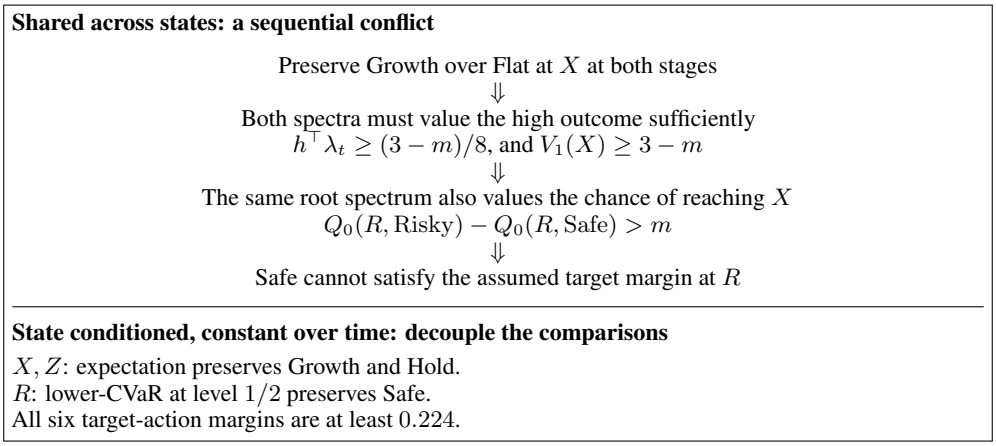


Figure 3: Mechanism of Theorem 1, for $q \in [0.5, 0.6]$ and $m = 0.04$. The upper panel assumes all target margins are at least $-m$ and derives a contradiction. It concerns the entire specified five-base time-only simplex, not a failed search over finitely many candidates. The lower panel is an exact supplied witness, not a learned-agent result. Arrows express proof implications rather than sampled trajectories.

A.2 Outer optimality and reachability

Lemma 1 (Unique target and reachable decision cells)**.** *For every $M_q \in \mathcal{M}$, π^* uniquely maximizes J_{M_q} among deterministic Markov policy tables. Every state–stage cell has positive occupancy under π^*.*

Proof. Let $U_t(s)$ denote the optimal outer continuation value from stage t, with $U_2 = 0$. At stage one, expected utilities of the target and alternative actions are, respectively,

s	R	X	Z
target	$9/10$	4	0
alternative	$-51/10$	3	-6.

Hence $U_1 = (9/10, 4, 0)$ in (R, X, Z) order, with unique target actions. At stage zero, the two actions at X share successor X, and those at Z share successor Z. Their target-minus-alternative utility gaps are thus one and six. At R, the outer action values are

$$U_0(R, \text{Safe}) = \frac{9}{10} + \gamma \frac{1}{100}\frac{9}{10} = \frac{90873}{100000}, \qquad U_0(R, \text{Risky}) = -\frac{51}{10} + 4\gamma q \le -\frac{693}{250}.$$

Safe is therefore the unique maximizer. Backward induction establishes the target as the unique optimal action at every cell.

The initial distribution has full support. Under the target, the stage-one distribution is $(1/300, 1/3, 199/300)$, also with full support. Any different deterministic policy either differs at stage zero, where its strict Bellman gap has positive initial mass, or agrees there and differs at a positively occupied stage-one cell. Even optimal subsequent actions cannot recover a strict Bellman gap. Its initial expected utility is therefore smaller. □

A.3 Uniform obstruction for time-only maps

Lemma 2 (Two-atom CVaR identity)**.** *If B_p is Bernoulli with success probability p and $v \ge 0$, then*

$$C_\alpha(c + vB_p) = c + v\frac{(p + \alpha - 1)_+}{\alpha}.$$

Proof. The lower quantile function equals c up to mass $1 - p$ and $c + v$ above that mass. Integrating it over $[0, \alpha]$ gives the identity. □

Lemma 3 (Time-only obstruction)**.** *For every $q \in [1/2, 3/5]$ and every time-only map λ, $m(\pi^*, \lambda; M_q) < -1/25$.*

Proof. Fix an arbitrary q and arbitrary $(\lambda_0, \lambda_1) \in (\Delta^4)^2$. Write $V_t = V_t^{\pi^*,\lambda}$, $Q_t = Q_t^{\pi^*,\lambda}$, $\varepsilon = 1/25$ and $h = (0, 0, 0, 1/3, 1/2)$. Suppose, for contradiction, that every target-action margin is at least $-\varepsilon$. By Lemma 2, the spectral score of the Growth reward is $8h^\top \lambda_t$. Growth and Flat share deterministic successor X, so their continuations cancel by translation equivariance:

$$Q_t(X, \text{Growth}) - Q_t(X, \text{Flat}) = 8h^\top \lambda_t - 3 \ge -\varepsilon, \qquad t = 0, 1.$$

Consequently $h^\top \lambda_t \ge (3 - \varepsilon)/8$ and $V_1(X) = 8h^\top \lambda_1 \ge 3 - \varepsilon > 0$. Target continuation gives $V_1(R) = 9/10$ and $V_1(Z) = 0$.

Define $r_j(q) = (q + \alpha_j - 1)_+/\alpha_j$. Since $q \ge 1/2$, $r_j(q) \ge r_j(1/2) = h_j$ componentwise. Nonnegative spectral weights and Lemma 2 therefore imply

$$Q_0(R, \text{Risky}) = -\frac{1}{10} + \gamma V_1(X) r(q)^\top \lambda_0 \ge -\frac{1}{10} + \gamma\frac{(3 - \varepsilon)^2}{8},$$
$$Q_0(R, \text{Safe}) \le \frac{9}{10} + \gamma\frac{1}{100}\frac{9}{10} = \frac{90873}{100000}.$$

The second inequality follows from $C_\alpha(Y) \le \mathbb{E}Y$, preserved by convex mixtures. Subtraction yields

$$Q_0(R, \text{Risky}) - Q_0(R, \text{Safe}) \ge \frac{26807}{500000} = \varepsilon + \frac{6807}{500000} > \varepsilon.$$

Thus the target Safe action has margin strictly below $-\varepsilon$, a contradiction. The argument uses the arbitrary map's own target continuations, without assuming the target is optimal under that map. □

A.4 A common stationary state-only witness

Lemma 4 (Uniform strict implementation)**.** *Let e_j denote the jth simplex vertex. The stationary map $\bar\lambda_R = e_3$, $\bar\lambda_X = \bar\lambda_Z = e_5$ satisfies $m(\pi^*, \bar\lambda; M_q) \ge 28/125$ for every $q \in [1/2, 3/5]$.*

Proof. The map uses $C_{1/2}$ at R and expectation at X, Z. All terminal inner action margins equal one, and the target terminal continuations are $(9/10, 4, 0)$. At stage zero, common deterministic

successors again cancel, leaving margin one at X, Z. At R, the Safe outcome equals $9/10$ with probability $99/100$, so its lower-half CVaR is $9/10$. Lemma 2 gives

$$Q_0^{\pi^*,\bar{\lambda}}(R, \text{Risky}) = -\frac{1}{10} + 4\gamma(2q-1)_+.$$

The Safe-minus-Risky margin is therefore

$$1 - 4\gamma(2q-1)_+ \geq 1 - 4\frac{97}{100}\frac{1}{5} = \frac{28}{125} > 0.$$

This lower bound applies to all six decision cells. Neither the map nor the bound depends on the choice of q within the family. □

A.5 Conclusion of the separation proof

Proof of Theorem 1. Take the family $\mathcal{M}$ defined above, $\varepsilon = 1/25$ and $\delta = 28/125$. Lemma 1 supplies the unique outer-optimal policy and reachability for every member. Lemma 3 establishes the obstruction for every admissible time-only map. Lemma 4 supplies one stationary state-only map valid throughout the family with uniform positive margin. These are precisely the quantified claims of Theorem 1. □

A.6 Robustness of inner implementation

Proposition 1 (Bounded perturbations of inner implementation)**.** *Fix the discount and bases. Use a common finite joint outcome label space for each action, including zero-mass labels with nominal rewards in* $[-1, 8]$*. Let reward changes be at most* δ *and probability total variation at most* ϵ *per action. Set*

$$E_1 = \delta + 72\epsilon, \qquad E_0 = \delta + \gamma E_1 + 72(1+\gamma)\epsilon.$$

If $2\max(E_0, E_1) < 1/25$*, time-only non-implementation and strict implementation by the fixed state-only witness persist.*

Proof. The finite-atom CVaR threshold representation bounds probability sensitivity by the nominal atom span times ϵ/α. Terminal nominal span is at most nine; root span is at most $9(1+\gamma)$. Since the smallest base level is $1/8$, these give the stated factors of 72. Monotonicity and translation equivariance bound reward and continuation perturbations by $\delta + \gamma E_{t+1}$. Mixtures preserve these uniform bounds, so each action margin changes by at most $2\max(E_0, E_1)$. Apply the margins of Theorem 1. □

For example, $\delta = 0.001$, $\epsilon = 0.00001$ give $E_0 = 0.0040868$. This proposition concerns inner implementation only. The outer penalty is discontinuous at zero; outer-optimality robustness would additionally require appropriate sign preservation and a separately controlled outer policy gap.

B Selection equations and implementation details

Our procedure selects a constrained risk map under the fixed outer utility. It uses offline candidate selection followed by neural fitting. The principal confirmation uses the horizon-two family of Theorem 1; earlier horizon-three experiments are reported separately.

Why separate selection from critic fitting? In development experiments, selecting maps through fitted neural policies produced high outer utility without faithful implementation of the selected map. On ten such cases, exact execution of the original selected maps restored fidelity but lost outer optimality in all ten. Reselecting maps through empirical nested backups recovered both properties in those cases. Because this correction used examined data, we report it as developmental diagnosis. The fresh evaluation in Section 4.1 freezes the correction before collection. This diagnosis motivates treating useful map selection and faithful critic execution as distinct requirements.

Outcome-based selection. Let Λ_k be the fixed candidate bank for sharing class k, and let D_{in} and D_{meta} contain disjoint trajectories. For each candidate, empirical conditional outcome laws from the inner data give

$$\widehat{Q}_t^{\lambda}(s,a) = \rho_{\lambda_{t,s}}\Big(R + \gamma \widehat{V}_{t+1}^{\lambda}(S') \mid s, a; D_{\mathrm{in}}\Big), \tag{9}$$

$$\widehat{V}_t^{\lambda}(s) = \max_a \widehat{Q}_t^{\lambda}(s,a), \qquad \widehat{\pi}_t^{\lambda}(s) = \arg\max_a \widehat{Q}_t^{\lambda}(s,a), \tag{10}$$

$$\widehat{\lambda}_k = \arg\max_{\lambda \in \Lambda_k} \widehat{J}_{D_{\mathrm{meta}}}(\widehat{\pi}^{\lambda}). \tag{11}$$

The terminal value is zero. Inner outcome donors use disjoint episode groups across stages; the stationary outcome law permits pooling donor transitions across stages. Meta evaluation estimates the fixed outer utility from its own empirical model. Ties use the frozen first-index rule, never exact-model labels. Observed outcomes and candidate maps supply all selection inputs. Exact model information is reserved for post-selection evaluation.

Neural deployment after selection. After candidate scores and the selected map are locked, we fit a backward quantile critic for each selected map using the same inner data. Each critic has two hidden layers of 32 Tanh units and 64 quantile heads, with 100 updates per stage at learning rate 0.003. Quantile regression includes a cross-fitted risk-score margin loss with weight one; query and outcome-donor episodes are disjoint. Targets use reward plus a frozen scalar nested-risk continuation, matching the nested valuation. The three sharing classes use identical architectures and update budgets but separate parameters. This requires three selected-map fits per unit. Direct planning and neural fitting retain their separately reported computational budgets.

Exact fitted loss (horizon two). The critic input is the three-dimensional one-hot physical state concatenated with $(2-t)/2$; its output has two actions and $N = 64$ heads. Write $\tau_i = (i + 1/2)/N$ for $i = 0, \ldots, N-1$, and $H(e) = e^2/2$ when $|e| \le 1$, otherwise $|e| - 1/2$. At stage t, query episodes satisfy $e \bmod 4 = 2t$ and contribute transitions at that stage; donor episodes satisfy $e \bmod 4 = 2t+1$. Donor transitions are pooled across stages conditional on (s,a), exploiting the stationary law. Each cell requires at least 16 distinct donor episodes. These sets are disjoint. Let $c_t(s')$ be the maximum spectral action score of the already frozen next-stage critic, or zero at the terminal stage. For each donor cell, define $d_t(s,a)$ as the exact atomic spectral valuation of its empirical samples $r + \gamma c_t(s')$, and $\Delta_t(s) = d_t(s,0) - d_t(s,1)$. For the fitted heads $z_i(s,a)$, define

$$\beta_i(\lambda) = \sum_j \lambda_j \frac{[\min\{(i+1)/N, \alpha_j\} - i/N]_+}{\alpha_j}, \tag{12}$$

$$S_t(s,a) = \sum_{i=0}^{N-1} \beta_i(\lambda_{t,s}) z_i(s,a), \tag{13}$$

$$L_t = \frac{1}{|Q_t|N} \sum_{(s,a,r,s') \in Q_t} \sum_i |\tau_i - \mathbf{1}\{y - z_i(s,a) < 0\}| H(y - z_i(s,a)) + \frac{1}{|Q_t|} \sum_{(s,a,r,s') \in Q_t} H\big(S_t(s,0) - S_t(s,1) - \Delta_t(s)\big), \qquad y = r + \gamma c_t(s'). \tag{14}$$

The indicator is detached in differentiation. Training uses full-batch Adam, float32, and separate backward-stage models. Heads are not sorted before the weighted sum: this loss does not guarantee ordered quantiles or calibrated values. Donor atomic valuations do sort their observed atoms. The reported action-fidelity test, rather than an assumed quantile guarantee, assesses execution.

Frozen candidate specification. The 80 per-unit bank files are the operative specification of the confirmation search, including candidate order and floating-point weights; they are included in the evidence package with hashes. Each bank has 109 maps per class. Shared banks include simplex-vertex and sampled five-base mixtures; the adaptive bank uses two-band state-only maps. These finite banks are not nested, and matching their sizes does not imply matched coverage. Their exact frozen files, rather than re-drawing candidates from a verbal distribution, define replay.

Deployment and evaluation are distinct. We retain the selected empirical policy and the fitted neural policy as separate endpoints. Every selection and neural policy is locked before exact evaluation. Outer optimality means maximizing J over all 64 deterministic policy tables; fidelity means matching the exact nested optimizer of the selected map at all six decision cells. Selection learns a useful map; different maps may implement the same policy.

Margin-based execution fidelity. The following standard action-gap argument links score accuracy to the fidelity check. Fix a selected risk map λ. Under the true outcome laws, define its optimal nested scores by

$$Q_t^{*,\lambda}(s,a) = \rho_{\lambda_{t,s}}\big(R + \gamma V_{t+1}^{*,\lambda}(S') \mid s,a\big), \qquad V_t^{*,\lambda}(s) = \max_a Q_t^{*,\lambda}(s,a), \quad V_H^{*,\lambda} = 0.$$

These optimize the selected inner valuation, rather than the outer objective. Let $S_t(s,a)$ be the deployed critic scores in equation 8.

Proposition 2 (Margin-based execution fidelity)**.** *At a decision cell (t,s) with at least two actions, suppose $a_{t,s}^*$ uniquely maximizes $Q_t^{*,\lambda}(s,\cdot)$. Define*

$$\Delta_{t,s} = Q_t^{*,\lambda}(s,a_{t,s}^*) - \max_{a \neq a_{t,s}^*} Q_t^{*,\lambda}(s,a) > 0, \qquad e_{t,s} = \max_a |S_t(s,a) - Q_t^{*,\lambda}(s,a)|.$$

If $2e_{t,s} < \Delta_{t,s}$, the deployed greedy action is uniquely $a_{t,s}^*$. If this condition holds at every cell, deployment is faithful everywhere. In particular, for a finite problem with unique maximizers at all cells, $\|S - Q^{*,\lambda}\|_\infty < \frac{1}{2}\min_{t,s}\Delta_{t,s}$ suffices.

Proof. For every $a \neq a_{t,s}^*$,

$$\begin{aligned} S_t(s,a_{t,s}^*) - S_t(s,a) &\geq Q_t^{*,\lambda}(s,a_{t,s}^*) - Q_t^{*,\lambda}(s,a) - 2e_{t,s} \\ &\geq \Delta_{t,s} - 2e_{t,s} > 0. \end{aligned}$$

Thus the exact maximizing action uniquely maximizes the deployed scores. Applying this argument at each cell proves the policy-level statement. □

This is a sufficient condition, not a necessary one: common score offsets, for example, preserve rankings. The reference scores use the true nested valuation; agreement with an empirical planner alone does not certify this bound. The proposition guarantees execution of the selected map, not outer optimality or a statistical error bound for the fitted critic. Our experiments check action fidelity directly rather than assuming its sufficient condition.

Controls and earlier selection procedure. State-only, global and time-only banks each contain 109 candidates. The adaptive bank uses the two-band subclass, whereas shared controls use five-base mixtures. Direct empirical utility planners use either inner data alone or pooled inner and meta data; their different data budgets are reported explicitly. In the earlier horizon-three studies, each candidate had its own fitted neural critic and meta selection scored its neural policy. Those studies therefore evaluate the original procedure and provide complementary evidence. Their state-time variants, data-efficiency comparisons and adverse results remain below.

C Final Currency Exchange protocol

This describes the stronger-control table in the main text; the later historical appendix records earlier nearest-neighbor-model studies, not this final fit. The environment permits fractional liquidation, but our planner compares only hold and sell-all. No equivalence to unrestricted fractional policies is claimed. This application deploys empirical model-based policies, not the neural critic used in the synthetic confirmation. It uses a single offline fit and developmental selection; episode-level intervals do not quantify refitting uncertainty.

Model and support. The inner fit uses episode IDs 0–511 and 576–8255 (8192 episodes); a disjoint meta fit uses IDs 8256–12351 (4096 episodes). The price model is a censored Gaussian AR(1), $P' = \mathrm{clip}(a + bP + \sigma\epsilon, 0, 5)$, $\epsilon \sim N(0,1)$, fitted by censored likelihood. Interior observations use Gaussian density; boundary observations use the corresponding Gaussian tail probability. L-BFGS-B starts from ordinary least squares and constrains $\log\sigma \in [-12, 3]$. A current price is supported when its 32nd-nearest training price is within one training-standard-deviation unit. Unsupported hold values are set to $-\infty$, except at the last stage.

Backup and selection. Backward planning uses 20 stages, no discount, a 201-point grid on $[0, 5]$, linear continuation interpolation and 64 midpoint Gaussian quantile nodes, clipped to $[0, 5]$. Terminal continuation is zero. The per-unit-inventory backup is $V_t(p) = \max\{p, \rho_{\lambda_{t,p}}(V_{t+1}(P'))\}$; at deployment both scores are multiplied by current inventory. Ties choose hold. The simulator uses continuous Gaussian innovations; it is not the finite quadrature model. Episodes terminate at 20 stages or inventory below 0.1, with no forced terminal liquidation. This accounting is retained for every evaluated policy.

Global and time-only banks each have 109 candidates: 17 constant two-band maps, three other simplex vertices and 89 Dirichlet mixtures (seed 54001). Time-only maps interpolate start and end spectra over the 20 stages. The state-only bank has 17 constants and 92 logistic price-dependent maps, with intercept and slope sampled uniformly on $[-3, 3]$ (seed 55001). The logistic input is price standardized using the training mean and standard deviation; its two-band weight is $w(p) = 1/2 + \mathrm{sigmoid}(b_0 + b_1\widetilde{p})/2$. Candidates are selected by lower-tail $\mathrm{CVaR}_{0.1}$ on 512 paired rollouts from the independent meta model (seed 55002), using its first 512 initial prices and first-index tie-breaking. The selection lock precedes reference evaluation.

Reference evaluation and controls. All methods use 4096 paired simulator episodes, with reset seed 55003 and noise seed 55004. Paired bootstrap intervals use 1000 episode resamples (seed 55005). Controls include the same planner with expectation and immediate liquidation. A 401-point-grid sensitivity check produced respectively 0, 1 and 1 action disagreements out of 8020 checks for global, time-only and state-only policies. The public environment revision and offline-data digest are recorded in the release provenance; published 1R2R scores are not inserted as matched baselines.

D Historical Bitcoin liquidation: protocol and robustness

We replay complete minute-quote windows from the Dryad dataset identified in the main text. The three venues share the Bitcoin market and are not independent market replications. Eligible hourly starts require 30 warmup minutes and a complete 60-minute future window; conditioning on quote completeness differs from deployment with missing feeds. No order-book depth or market impact is modeled. Normalized proceeds are $Y = (1 - f)B_\tau/B_0$, with $f = 0.0025$ and $\tau \in \{1, \ldots, 60\}$. The common fee is a controlled scenario, not a reconstruction of historical venue-specific fees.

Model and selection. Features are time remaining, log bid relative to arrival, relative spread, one-minute and trailing 30-minute log-mid changes, and trailing volatility. Training-only standardization supports stagewise nearest-anchor projection. Each of 64 anchors uses 32 nearest training outcome donors; unsupported states use the common sell fallback. Backward planning uses $\rho_w = (2w - 1)C_{1/2} + 2(1 - w)\mathbb{E}$, $w \in [1/2, 1]$. The shared bank has 109 equally spaced weights; the time/state banks add 96 logistic maps to all shared candidates. State maps exclude the clock. The selected GDAX map is $w(s) = 1/2 + \sigma(z_{30\mathrm{min}} - 1)/2$. The 512 fitting episodes occupy January–June 2018; 128 separate selection episodes occupy July–September, and 128 development episodes October–December. All selections were locked before opening 2019 returns and match the original 64-anchor development fit; no favorable bootstrap refit was substituted.

Evaluation and uncertainty. All eligible 2019 starts are included: 6,986 Bitstamp, 7,158 Bitfinex and 7,183 GDAX episodes. January–March and April–October are fixed reporting periods, with no intervening refit or tuning. We use 2,000 paired calendar-week bootstrap resamples, seed 95001, and percentile intervals. These condition on the selected policies, omit refitting uncertainty and carry no familywise significance claim. The combined interval is not an additional replication.

Table 4: Locked GDAX replay: net proceeds in basis points, including the common fee. Higher is better. Point estimates over 7,183 episodes.

Method	Mean	$\mathrm{CVaR}_{0.1}$
State-dependent	-24.637	-47.740
Shared	-24.684	-48.615
Time-dependent	-24.695	-48.971
Expectation-only	-24.628	-56.723
Immediate	-24.661	-47.456
Deadline	-23.552	-150.590

Adaptive and shared fills differ in 434 GDAX episodes: proceeds improve in 183, worsen in 186 and match in 6,814 (some different fills have equal quotes). Adaptive waits in 155 episodes versus shared's 513. Monthly tail differences are positive in eight months and negative in May and October; these slices are descriptive. Against immediate sale, the mean difference is 0.024 bp ($[-0.023, 0.068]$), while the tail difference is -0.284 bp ($[-0.563, -0.070]$). The earlier development advantage was concentrated in three changed decisions, with episode effects $-3.069, +24.932, -0.854$ bp.

Table 5: Five fitting-week resamples per venue, with selection repeated. Effects are adaptive minus shared risk in basis points. These refits share evaluation data; ranges are descriptive, not confidence intervals.

Venue	Varying map selected	Mean effect range	CVaR10 range
Bitstamp	0/5	$[0, 0]$	$[0, 0]$
Bitfinex	0/5	$[0, 0]$	$[0, 0]$
GDAX	1/5	$[0, 0.0444]$	$[0, 0]$

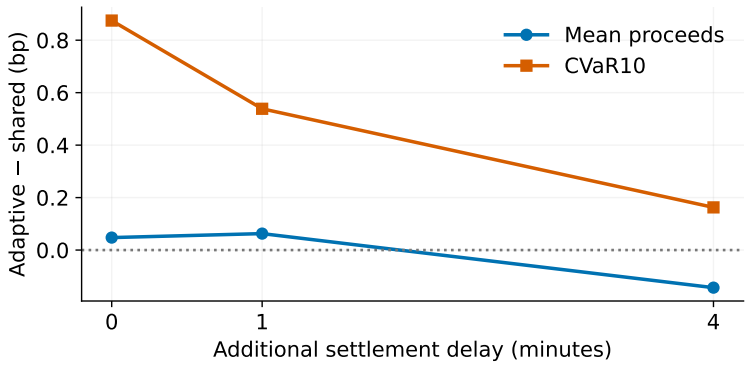


Figure 4: Settlement-delay sensitivity of the locked GDAX map. Point estimates compare adaptive with shared risk, holding decisions fixed and delaying fills up to the mandatory minute-60 deadline. This is a stress proxy, not latency-aware replanning or measured execution costs.

Estimation and execution boundaries. Training-only chronological folds fit on 256/384 episodes and predict the next 128. Conditional quantile scores improve over unconditional samples in all six venue/fold comparisons, but mean RMSE is slightly worse throughout. Nominal 90% conditional coverage ranges from 85.74% to 89.75%. GDAX 2019 coverage is 78.19%, and 36.63% of all queried decision states are outside fitted support (not the fraction of executed decisions). Five fitting-week resamples per venue preserve whole calendar-week blocks; all 15 adaptive-minus-shared tail differences are zero. They are resampling robustness, not independent data collection.

A post-hoc diagnostic of saved 2018 selection returns localizes part of this sensitivity. In 14 of 15 refits, the best varying map ties the best shared map on selection CVaR; shared candidates occur first and the frozen tie rule selects one. In the remaining GDAX refit, the varying map wins selection by 0.598 bp. Its two changed assessment returns leave the lowest 13 ordered observations and their

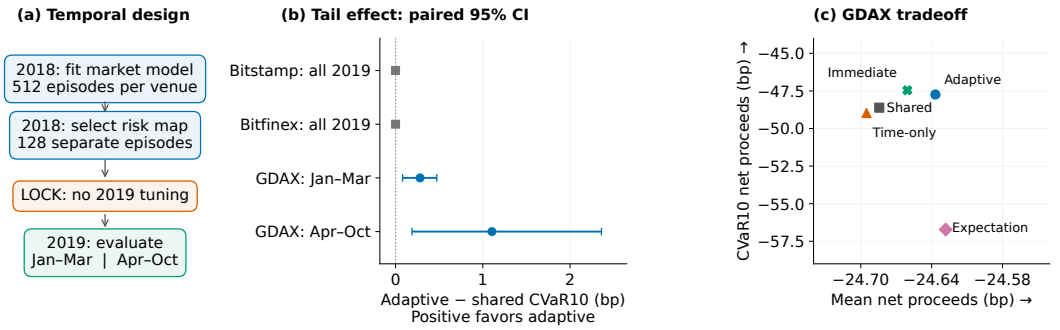


Figure 5: Protocol and complete venue context. The locked temporal design separates fitting, selection and evaluation. Paired weekly 95% intervals show GDAX tail effects and the other venues' identical selected policies. The GDAX mean–tail comparison displays the immediate-sale tradeoff.

episode identities unchanged, hence leave fractional CVaR at mass 12.8 unchanged. For the originally selected GDAX functional form, selection differences versus the best shared map across the five refits are $0, -2.682, +0.598, 0, 0$ bp. Thus both selection-score ties and a strict ranking reversal occur. These saved-return diagnostics retain every original selection; they neither evaluate alternative tied maps on assessment data nor establish that changing the tie rule would improve later performance.

Five-refit mechanism analysis. The five reconstructed GDAX models reveal distinct adaptive mean–downside trade-offs. We retain the functional form $w(s) = 1/2 + \sigma(z_{30\min} - 1)/2$, apply each refit's training normalization, and compare its policy with that refit's selected shared map on the same 128 assessment episodes from 2018. Mean proceeds improve in three reconstructions and match in one; the fifth exchanges mean proceeds for stronger downside protection. Empirical CVaR improves in two reconstructions and matches in three (Table 6). One reconstruction improves both metrics. These comparisons expose useful adaptive behaviours that a tail-only selection score can leave tied with shared risk.

Table 6: Five-refit mechanism analysis. Differences are fixed adaptive form minus the refit-specific selected shared controller, in basis points. The rows correspond to the five recorded training-block resamples, in order. The same assessment episodes are used throughout; these are descriptive model-reconstruction comparisons.

Refit	Mean difference	$\text{CVaR}_{0.1}$ difference
1	0	0
2	+0.3196	0
3	+0.0444	0
4	+0.0938	+0.0601
5	-1.1629	+1.0652

These reconstructions distinguish the behaviour of a fixed adaptive form from the map chosen by the selection procedure. All original selections remain recorded. The mean of the five mean differences is -0.1410 bp, while the mean of their CVaR differences is $+0.2251$ bp; the latter averages refit-specific tail metrics rather than pooling returns.

Complete twenty-refit context. The same training-block stream was extended to twenty GDAX refits, including the five above, with unchanged splits and selection. For the fixed adaptive form, mean proceeds are higher/equal/lower in $13/1/6$ refits and CVaR in $3/10/7$, averaging -0.0069 and -0.2273 bp respectively. Repeated selection yields counts $1/15/4$ and $1/17/2$, averaging -0.0235 and -0.0316 bp. Thus the five-refit tail pattern does not persist across the extension. These descriptive, overlapping-data resamples test fitting sensitivity; the separately locked 2019 test establishes temporal transfer for one controller, not repeatable superiority of refitting.

An extra 0.05 bp differential execution cost erases the observed 0.0475 bp mean gain over shared risk. One/four extra settlement minutes yield mean differences $+0.0627/-0.1433$ bp and tail

differences $+0.5386/+0.1627$ bp. Bid-side spread is already reflected in execution; differential costs are additional hypothetical costs, not measured impact. Scaling the mean estimate to \$1 billion/day gives approximately \$4,754/day conditionally, but neither the mean interval nor these quote data establish profit or executable capacity.

Reproducibility. Frozen source hashes, chronological manifests, selected indices, all outcomes and stress results are retained in the study's evidence record. A separate arithmetic audit reproduces the selected policies, fills, proceeds and fractional CVaR calculations; 34 Dryad tests pass. The 2019 periods are now evaluated data and cannot serve as untouched confirmation for subsequent method changes.

E Conditional scope and neighboring questions

The inner risk map is a designed controller component, not an identified human preference. A possible reason to require this representation is to expose local valuation rules for inspection; the present experiments do not establish that such inspection is easier or useful to practitioners. Direct outer-utility planning remains a legitimate alternative and matches optimal utility in the fresh synthetic confirmation. Thus the theorem concerns a restriction within one controller class, not the necessity of risk-map learning for general control.

What the interface makes testable. Fixing a map specifies a local distribution-to-score operator independently of the fitted executor. The same outer utility can therefore be accompanied by different action margins and different fidelity to the chosen operator. Our design exposes these two objects for separate evaluation: usefulness of the induced decisions, and agreement with their stated valuation rule. This is a representational and verification motivation; computational savings, human interpretability benefits and preference identification require additional evidence. The synthetic direct planners provide the utility reference rather than an implementation of the same restricted risk interface.

Closest formulations and algorithms. Moghimi & Ku (2025) optimize a static spectral measure of total return using a distributional method on an augmented state space. Their intermediate risk measures explain the resulting policy (Section 5); they are not the freely selected nested controller studied here. Our outer objective scores the policy induced by a local risk map, while their spectral objective scores the total-return distribution. This distinguishes both the optimized object and the role of evolving preferences. The distinction also matters for dynamic programming: Hau et al. (2023) identify invalid saddle-point assumptions in particular static CVaR/EVaR policy-optimization decompositions. Our recursion defines a nested valuation directly, rather than claiming such a decomposition of a static return criterion.

Wang & Xu (2023), Definition 2.2 and equation (2.3), aggregate state-conditional distortion measures over acts using subjective state probabilities. Their Section 5 formulates preference-robust decision making, with a weighted-CVaR representation in equation (5.2) and alternating updates of worst-case preferences and decisions in Algorithm 5.1. Our object is a Markov policy evaluated through nested continuation values, and our map is selected to serve a supplied outer objective. The separation concerns sharing that nested operator across decision states, rather than ambiguity about an agent's preferences or beliefs.

Schlosser (2020), equations (18)–(21) and Algorithm 5.1, calibrates local moment-based subproblems against a target return distribution and adjusts a weight to improve the resulting moment fit. This is direct precedent for tuning local valuation toward an overall distributional target. Our specific addition is a quantified obstruction for an entire time-only spectral class, a strict stationary state-dependent witness, and outcome-based selection followed by separate utility and execution checks. These are the objects of comparison, rather than state dependence or local calibration alone.

State-dependent risk aversion predates this work. In particular, Schlosser (2022) studies heuristic finite-horizon mean–variance and mean–semivariance optimization through state-dependent risk aversion and dynamic programming. Our claimed object is narrower: a specified spectral family's inability under time-only sharing to implement a target policy, a strict stationary witness, and separate

evaluation of selection and execution. The neighboring mean–variance result is not evidence that state dependence itself is new, nor do we claim a universal first result on adaptive preferences.

Preference elicitation is a different statistical task. Cheng et al. (2025) estimates spectral risk aversion using adaptive binary-choice questionnaires. Here the outer utility is supplied in advance, and maps are selected from outcome data without expert choices. Different maps can induce the same actions, so good outer utility or action fidelity alone does not identify a unique latent preference.

Contribution boundary. The new object is the sequential obstruction imposed by sharing a specified nested spectral operator, together with strict state-dependent implementation and separate selection/execution assessment. State dependence, quantile critics and the standard action-gap stability argument are established tools. The external studies assess decisions induced by empirical planning; they do not establish that a neural executor faithfully implements unknown market laws.

E.1 Comparison coverage and validated components

Study	Search and information access	Component and inference scope
Theorem	Entire specified five-base time-only simplex; exact target-policy continuations.	Sequential implementation separation; no finite-search or learning guarantee.
Fresh two-stage	109 maps/class; adaptive two-band, shared five-base; nonnested. 50,000 inner and 50,000 disjoint meta episodes/unit.	Outcome-based selection and neural fitting; exact utility/fidelity after locking. 80 units, conditional on eight instances.
Three-stage	Seven methods, three inner budgets with equal meta budgets; earlier candidate-wise neural selector. Pooled direct controls use both datasets.	Complementary sequential and transition-shift evidence, not a replication of the corrected selector. Budget prefixes within seeds are dependent.
Currency Exchange	109 maps/class, nonnested; 512 selection rollouts/map. Dynamics fitted on 8,192 inner and 4,096 meta episodes. Hold/sell-all actions.	Empirical planning and selection; 4,096 paired simulator evaluations conditional on one fit, not external neural-fidelity validation.
Historical Bitcoin	109 shared maps included in each 205-map state/time bank. 512 fitting and 128 selection episodes; common model and action restrictions.	Empirical planning and locked temporal replay. Weekly intervals condition on the selected fit; resampling studies separately assess fitting sensitivity.

Equal bank size controls candidate count, not policy coverage. In the fresh controls, time-only banks omit an optimal map in 20 units despite a global implementation existing; that deficit is a search-coverage result. The necessary-family obstruction instead covers the full specified shared class. Historical banks contain the shared candidates but have unequal search budgets, so their comparison is a full-system comparison rather than an isolated effect of state dependence at equal search cost.

An additional nested-bank study would need separately frozen candidates, selection budgets and evaluation data. None is reported here. The existing results are not relabelled as a matched-coverage ablation.

F Slippery CliffWalking: a shared-risk control

This supplementary experiment asks whether a bank permitting state dependence selects a nonconstant map on a different transition graph. It evaluates sampled-model planning, separately from the neural confirmation. In all five fits the state-dependent bank selects a constant map and exactly matches the nine-candidate shared comparator. This establishes fallback to shared risk within the tested bank, rather than equivalence of the underlying function classes.

Task and information access. Gymnasium 1.2.3 supplies the slippery `CliffWalking-v1` environment (`is_slippery=True`), with 48 state indices, four actions and three equiprobable transition outcomes per state/action. Native costs are one per ordinary step and 100 for a cliff fall, which returns the agent to the start. We fix a horizon of 100, no discount, and an additional cost of 100 for timeout without reaching the goal. This is a declared task adaptation, not the unmodified native benchmark score. The outer objective minimizes upper-tail episodic cost CVaR at mass 0.1.

Five fits use 256 independent outcome draws per state/action (seeds 1–5). Every planner receives the same fitted probabilities within a fit; outcome support and arbitrary-state generative access are known.

Banks and selection. Nested cost scores mix expectation and upper-tail CVaR at mass 0.5 with coefficients $2(1-w)$ and $2w-1$, respectively, for $w \in [0.5, 1]$. The shared bank contains nine equally spaced constants. State and time banks each add four maps $w = 0.5 + 0.5\,\mathrm{sigmoid}(bz)$, $b \in \{-2, -1, 1, 2\}$. For state maps, $z = \mathrm{row}/1.5 - 1$; for time maps, $z = 2t/99 - 1$. A budget-matched shared bank instead adds constants $\{0.53125, 0.65625, 0.78125, 0.90625\}$, also totaling thirteen. Thus shared-nine is nested in the adaptive banks; shared-thirteen controls candidate count. The four nonconstant state maps do not exhaust state dependence. Each candidate receives 2,000 paired native-law selection episodes (seeds 760000–761999). First-index ties and selected policies are locked before 10,000 fresh paired evaluation episodes (770000–779999). Scenarios are shared across methods and fits, not 50,000 independent evaluation scenarios.

Table 7: CliffWalking control, averaged over five fitted models. Costs are lower-is-better; goal rate is higher-is-better. No evaluated method incurred a cliff fall; adverse returns arose from timeout. These are empirical planning endpoints, not neural policies.

Method	Mean cost	$\mathrm{CVaR}_{0.1}$ cost	Goal rate (%)
Shared, nine candidates	72.26974	190.8314	90.910
Shared, thirteen candidates	72.24536	190.6092	90.932
State bank, thirteen candidates	72.26974	190.8314	90.910
Time bank, thirteen candidates	72.54294	191.4996	90.844
Expected-cost planner	72.60960	192.7162	90.688
Direct episodic CVaR planner	97.98294	192.4726	72.748

Selection and paired uncertainty. For fits 1–5, the state bank selects weights 0.5625, 0.6875, 0.6875, 0.8125 and 0.5, respectively, matching shared-nine. Exact equality of evaluation returns follows from selecting identical policies; it is not an independent statistical demonstration of equivalence. Against shared-thirteen, the state bank's CVaR cost difference is $+0.2222$, with an exploratory 95% paired bootstrap interval $[-0.2420, 1.0968]$. Time minus shared-thirteen is $+0.8904$ ($[-1.5958, 3.4425]$), and shared-thirteen minus expectation is -2.1070 ($[-5.1436, 0.1797]$). None resolves superiority. The 1,000 bootstrap replicates (seed 76076) resample fitted-model indices and common scenario indices across all methods. Intervals condition on the selection scenarios and task specification, use only five fits, and have no multiplicity adjustment.

Direct episodic baseline. The reference solves $\min_\eta\{\eta + \mathbb{E}[(C-\eta)_+]/0.1\}$ on each fitted model, using backward planning with cumulative-cost memory. Native costs are integers and a safe noncompletion policy costs 200, so integer thresholds 0–200 suffice. At three or more falls the accumulated cost exceeds every threshold; the remaining hinge objective is linear and an expected-cost continuation suffices. This baseline has a richer policy state than the risk maps and optimizes the outer criterion directly. It shares fitted data but does not use selection episodes as the banks do. In fit 3 it selects threshold 200 and a noncompletion policy costing 200 on every evaluation episode. The timeout penalty permits this behavior; it is not a completion constraint. This fit dominates the aggregate mean-cost gap, so that gap should not be interpreted as general superiority over direct CVaR.

Verification and interpretation. Vectorized rollouts reproduce native-environment costs exactly for 760 episodes each under expected-cost and direct-CVaR policies. Fixed-threshold Monte Carlo checks use 100,000 episodes on each fitted model and agree with the corresponding dynamic-programming objectives within the six-standard-error engineering tolerance. All five selection locks, return arrays and source hashes were checked. The result supplies a control for the conditional claim: the tested procedure can retain shared risk on a task where its nonconstant candidates do not improve selection. It does not establish that state dependence is unnecessary for every map or every CliffWalking configuration.

G FRESH CONFIRMATION PROTOCOL AND DIAGNOSTICS

The confirmation uses the stationary transition graph, $H = 2$, $\gamma = 0.97$, uniform initial states and outer utility in the main construction. For each of eight instances, q is drawn independently and uniformly from $\{500001, \ldots, 599999\}/10^6$, without rejection. The realized necessary values are 0.550113, 0.591574, 0.531800, 0.557147; control values are 0.524027, 0.577772, 0.562914, 0.507251. Safe rewards are respectively 0.9 and 2. Each instance has ten paired seeds. Inner and meta data contain 50,000 episodes each, or 100,000 transitions per role. This budget supersedes the unsuccessful 6,000-episode engineering coverage preflight; no preflight outcome is counted as a scientific result and no final unit receives a success-dependent top-up.

The 109-map banks are reused from the frozen development package to isolate new sampling, initialization and parameter effects; they are not newly optimized. PCG64 parameter streams are $239000000 + i$ for instance $i = 0, \ldots, 7$. For unit $j = 0, \ldots, 79$, let $b_j = 230000000 + 100000j$: inner environment and action streams are $b_j, b_j + 1$, meta streams $b_j + 10000, b_j + 10001$, and neural initialization uses $b_j + 50000$. Pooled data use identity $b_j + 90000$, not a new random stream. Donor episodes have indices modulo four equal to $2t + 1$ at stage t; query episodes use $2t$. Donor transitions are pooled across stages conditional on state and action under the stationary law. Meta evaluation uses its own stagewise empirical law and initial-state frequencies. All candidate policies, scores and meta values are retained, including ties. Exact model information enters only after selections and policies are locked across all units and verified.

For each deployment, confirmation requires all 40 units in each stratum, at least 36 jointly faithful and optimal in each, and positive necessary-stratum paired utility advantages with positive 95% bootstrap lower bounds against both shared controls. Missing units would make the outcome incomplete, not reduce the denominator. Each paired seed-block difference averages the four necessary instances at that seed index. The percentile bootstrap resamples ten blocks 10,000 times with PCG64 stream 239100036. No criterion was changed after results.

Development diagnosis and fresh confirmation. The selection correction was developed on ten previously examined fidelity failures. Exact execution of their original maps gave fidelity in 10/10 but optimality in 0/10; empirical reselection recovered both in 10/10. A subsequent reused-data check covered the remaining 70 cases. These diagnoses motivate the correction and are reported separately from the 80 fresh confirmation units. The fresh study evaluates the corrected procedure against shared-risk and direct-utility controls.

Ties, margins and value accuracy. For state-only, global and time-only selection respectively, faithful meta-top candidates number 1,918/1,944, 3,811/3,838 and 256/263 across the 80 banks. These are candidate counts, not independent trials. In unit 37, 15 of 17 meta-top candidates are faithful, but the first-index choice (candidate 24) is not. No guarantee of fidelity under arbitrary tie breaking follows. The minimum exact absolute action gap of a selected state-only map is 0.043910. Maximum neural score errors against exact selected-map values are 4.789116, 4.867585 and 5.799153 for the three methods. These errors distinguish successful action ranking from uniform value calibration.

Compute and reproducibility. There are 240 selected-map fits, each with 100 updates per stage, or 48,000 optimizer updates in total. Workers use one CPU, 8 GiB memory, no GPU, and at most two concurrent tasks. Aggregate scheduler runtime is 8,691 CPU-seconds; mean task runtime is 108.64 seconds. Worker versions are NumPy 2.2.6 and PyTorch 2.9.1+cpu; local analysis uses NumPy 1.26.4 and PyTorch 2.9.0+cpu. Hash inventories, outcome digests, stream identities, fitting bindings and reconstructed selections verify all 80 units. All 240 checkpoint replays preserve every action; maximum score roundoff is 4.77×10^{-7} and is not treated as a scientific failure. Independent arithmetic and lower-tail integration checks reproduce the reported utility and fidelity results. The local evidence package preserves the complete audit; preparation of the anonymous release remains a separate submission task.

H ADDITIONAL EXACT COVERAGE

H.1 FROZEN SAME-FAMILY CERTIFICATE COVERAGE

We conducted a separate exact-model validation of a sufficient certificate for time-only versus stationary state-only risk. This study evaluates exact implementation capacity. The finite horizon is two, with uniform initial distribution over physical states R, X, Z and stationary rewards and transitions. At R, Safe gives reward c and transitions to R with probability p, otherwise Z; Risky gives $-e$ and transitions to X with probability q, otherwise Z. At X, Growth gives an equiprobable reward in $\{0, b\}$ and Flat gives a; both return to X. At Z, Hold and Loss give 0 and -1, respectively, and return to Z. The fixed target chooses Safe, Growth and Hold at both stages. The outer criterion is $J = \mathbb{E}\sum_{t=0}^{1} \gamma^t [r_t - 5\mathbf{1}\{r_t < 0\}]$.

The population was specified after the discovery construction. Exactly 200 fresh parameter vectors were drawn, independently by coordinate and with replacement, uniformly from the following grids (step 0.01, endpoints included): $a \in [2, 4]$, $b \in [6, 10]$, $c \in [0.5, 2]$, $e \in [0.05, 0.50]$, $p \in [0.01, 0.30]$, $q \in [0.30, 0.80]$ and $\gamma \in [0.90, 0.99]$. No draws were rejected because of target optimality or certificate applicability. Three known construction/control instances were checked separately and are excluded from the scientific denominator. Portable hash-based sampling and all inputs are retained in the reproducibility record.

For each draw we enumerate all 64 deterministic Markov policy tables to check whether the fixed target uniquely maximizes J. Inner valuation recursively composes mixtures of lower-CVaR utilities at levels $(1/8, 1/4, 1/2, 3/4, 1)$. The fixed local witness uses CVaR at level $1/2$ at R and expectation at X, Z, shared across stages. Its six target-action margins are evaluated exactly. Time-only feasibility is tested using all 25 ordered pairs of pure basis functionals across the two stages. A successful witness proves feasibility; failure of this finite search does not prove impossibility.

For $0 < a < b/2$ and $0 < p \le 1/2$, set $m = 1/25$, $k = (a - m)/b$, $h = (0, 0, 0, 1/3, 1/2)$, and

$$\begin{aligned} r_j(q) &= \max\{0, (q + \alpha_j - 1)/\alpha_j\}, \\ L(q, k) &= \min_{\lambda \in \Delta^4 :\, h^\top \lambda \ge k} r(q)^\top \lambda, \\ F_q(m) &= -e + \gamma(a - m)L(q, k) - c(1 + \gamma p) - m. \end{aligned}$$

The sufficient obstruction requires $F_q(m) > 0$. To see its logic, if every target-action margin were at least $-m$, the two Growth comparisons would force $h^\top \lambda_t \ge k$ and terminal $V_1(X) \ge a - m$. Risky's root score would then exceed Safe's upper bound $c(1 + \gamma p)$ by more than m, a contradiction. Continuations here are evaluated under the target and the tested map, even when that target is not greedy. This is a sufficient family-specific argument; the interval construction and its proof appear in Section 2.2; the broader-grid test uses the sufficient certificate displayed here.

Table 8: Exact certificate coverage on all 200 frozen draws. These are model instances, not training seeds. No processing failures occurred.

Outcome	Count	Percentage
Certified separation	11	5.5%
Certified time-only feasibility	90	45.0%
Unknown	99	49.5%

Separation requires a uniquely outer-optimal target, strict local witness and positive obstruction certificate. There were 179 uniquely outer-optimal targets: 11 separations, 90 time-only witnesses and 78 unknowns. All 21 other draws remain in Table 8. Unknown reason flags overlap: 56 have a nonstrict local witness, 61 an inconclusive certificate, 21 fail target uniqueness or optimality, and 21 fall outside certificate assumptions. None was resampled. All outputs passed inventory, sampler and exact-result verification, including a second CVaR computation; model construction is shared between implementations.

These results recover the obstruction beyond the known controls and establish time-only sufficiency in many sampled instances. Additional separations may exist among unknowns, leaving prevalence

unresolved. The population retains the same transition graph and engineered reward structure, providing exact within-family evidence about implementation capacity. This descriptive study reports the complete collection under its frozen protocol.

I Complementary horizon-three evidence

The following complementary studies use the original neural-bank selector and a separate horizon-three model. They report global-risk comparisons, exact implementation certificates and the full direct-control comparison.

I.1 Completed exact model and certified separation

For this certificate, the global spectrum is a nonnegative five-coordinate simplex mixture of lower-tail CVaR utilities at levels $(1/8, 1/4, 1/2, 3/4, 1)$. The adaptive embedding is $\lambda(w) = (0, 0, 2w - 1, 0, 2(1-w))$, $w \in [1/2, 1]$. Both use the nested conditional recursion.

The controlled model has three observed states, two actions, $H = 3$, $\gamma = 0.97$, and initial probabilities $(0.8, 0.15, 0.05)$. Reward atoms are equiprobable, independently of the next state conditional on state and action.

State	Growth	Defend
Calm	$(0, 8)$	$(3, 3)$
Warning	$(-1, 9)$	$(2, 2)$
Fragile	$(-2, 4)$	$(0, 0)$

Growth and Defend transitions on the design law are respectively

$$P_G = \begin{pmatrix} .85 & .15 & 0 \\ .10 & .55 & .35 \\ .05 & .20 & .75 \end{pmatrix}, \quad P_D = \begin{pmatrix} .864 & .136 & 0 \\ .13 & .57 & .30 \\ .06 & .23 & .71 \end{pmatrix}.$$

On the shifted law they are

$$P'_G = \begin{pmatrix} .65 & .35 & 0 \\ .05 & .35 & .60 \\ .10 & .25 & .65 \end{pmatrix}, \quad P'_D = \begin{pmatrix} .696 & .304 & 0 \\ .08 & .40 & .52 \\ .11 & .28 & .61 \end{pmatrix}.$$

Declared decimals are interpreted as exact rationals for certification. This defines the mathematical model certified by rational arithmetic.

Proposition 3 (Global non-attainment)**.** *No global spectrum in the specified five-component simplex makes the policy Growth in Calm and Defend in Warning/Fragile at every stage Bellman-consistent, even allowing arbitrary favorable resolution of risk-action ties.*

Proof. At the terminal stage, the necessary Calm and Warning chosen-action advantages have coefficient rows $c = (-3, -3, -3, -1/3, 1)$ and $d = (3, 3, 3, -1/3, -2)$ in λ. Both must have nonnegative dot product with λ. However, $5c + 4d = (-3, -3, -3, -3, -3)$, so $0 \le (5c + 4d)^\top \lambda = -3$, a contradiction. Terminal rewards coincide across the two laws. No continuation or learned critic enters this proof. □

Exact finite-model certificate. Enumeration of all $2^9 = 512$ deterministic Markov policies per law finds the stated policy to be the unique outer optimum. Every state has positive initial mass and positive self-transition under both actions, so terminal restrictions cannot be discarded as unreachable. Three frozen learned state-only maps (one per development seed) make this policy exactly Bellman-consistent on BOTH laws. We verify all six map/law combinations with rational ordered-atom and independent threshold-formula CVaR calculations. No weights are adjusted.

Law	J^*	Global upper bound	Gap lower bound
Design	9.484265038	9.401648792	0.082616246
Shifted	8.416317057	8.335215660	0.081101397

The upper bound is the best utility among all policies other than the unique optimum. Since every Bellman-consistent global policy belongs to that remaining finite set, the bound is conservative. The adaptive witnesses attain J^*, yielding the displayed separation. This certificate establishes separation for deterministic Markov policies that are Bellman-consistent at every decision in this finite model.

I.2 Richer-global comparison and execution diagnostics

The earlier richer-global study uses three development seeds, 12,000 inner and 12,000 meta episodes per seed, 64 quantile heads, two 32-unit hidden layers, and 100 optimizer updates per stage with learning rate 0.003. The primary reporting budget is 109; all 13/37/109 prefixes are retained. Each policy/law is evaluated on 1,000 episodes. The richer global bank contains 109 spectra, including 13 reused anchors; a 205-candidate union is an explicitly unequal-budget stress test. Across seeds there are 1,605 risk fits, three neural utility fits and three empirical planners. Candidate counts are matched across risk classes; single-fit controls have distinct computational budgets.

Adaptive variant	Comparator	Design mean ΔJ	Shifted mean ΔJ
State-time	Rich global / union	0.6894681	1.1180484
State-only	Rich global / union	0.7244592	1.1720903
State-time	Direct planner	-0.0349911	-0.0540419
State-only	Direct planner	0	0

Original global, rich global and union select identical Always-Growth policies; their matching effects reflect that shared selected policy. State-only matches the empirical planner, direct neural utility control and direct affine policy in all three seeds. State-time matches in seeds 1 and 3 but loses in seed 2, where no planner-matching policy exists in its fitted candidate bank. The selector correctly chooses its meta maximum. This diagnoses missing realized policy coverage within the fitted bank.

The selected seed-2 state-time critic disagrees with its exact map at two cells. One disagreement actually favors the direct utility-optimal action: fixing risk fidelity and outer utility therefore require separate checks. State-only matches direct control, while the state-time deficit localizes a candidate-coverage issue. Within-seed paired bootstrap intervals use 999 resamples and describe episode uncertainty conditional on each fitted model, without multiplicity adjustment. This is development evidence on the horizon-three model; the fresh horizon-two confirmation above evaluates the corrected selector.

I.3 Ten-seed data-efficiency evaluation

The subsequent frozen campaign evaluates seven methods at 2,000, 6,000 and 12,000 inner episodes with equally sized, disjoint meta datasets. Within each of ten training seeds, lower budgets are nested prefixes of the largest datasets; the budget comparisons therefore share data. The environment is the earlier fully observed three-state, two-action, horizon-three model with discount 0.97. Policies are locked before exact-model evaluation on the design and shifted transition laws. The complete matrix contains $10 \times 3 \times 7 \times 2 = 420$ evaluations.

State-only, state-time and rich-global banks each contain 109 candidates. Candidate-specific backward critics use 64 quantiles and 100 updates per stage at learning rate 0.003. Direct neural utility controls use two hidden layers of 70 units and the same per-stage update count and learning rate. Direct controls are trained either on the inner data alone or on the pooled inner-plus-meta data; the latter account for the adaptive method's selection data. Optimization compute is reported separately from the shared data budget. Total work comprises 9,810 risk fits, 60 neural utility fits and 60 empirical planners, with 2,961,000 optimizer updates.

The primary statistic is the state-only minus comparator utility difference, averaged over the two lower budgets within each seed on the design law, then averaged across seeds. The predeclared bootstrap resamples ten paired seeds 10,000 times; percentile endpoints are 0.0125 and 0.9875 (97.5% intervals). These approximate intervals quantify training-seed uncertainty within this model.

Table 9: Mean exact outer utility over ten training seeds. D and S denote design and shifted laws. Columns are inner episode budgets; an equal meta budget is available for selection or pooled direct training. These descriptive means summarize dependent budget and law evaluations.

	D			S		
Method	2,000	6,000	12,000	2,000	6,000	12,000
State-only	9.445504	9.484265	9.484265	8.356083	8.416317	8.416317
State-time	9.472896	9.484265	9.484265	8.404558	8.416317	8.416317
Rich global	8.734924	8.714900	8.714900	7.285643	7.245190	7.245190
Planner, inner	9.393088	9.449632	9.460970	8.293259	8.370732	8.382435
Neural, inner	9.401350	9.460970	9.460970	8.301369	8.382435	8.382435
Planner, pooled	9.461248	9.484265	9.484265	8.387616	8.416317	8.416317
Neural, pooled	9.472618	9.484265	9.484265	8.399376	8.416317	8.416317

Comparator	Mean difference	97.5% bootstrap interval
Pooled planner	-0.007872	$[-0.034192,\ 0.017471]$
Pooled neural utility	-0.013557	$[-0.051012,\ 0.017471]$

The promotion criterion required both lower interval bounds and both per-budget mean differences against each pooled baseline to be positive. It was not met. Planner differences at 2,000 and 6,000 episodes were -0.015745 and zero; neural differences were -0.027114 and zero. The predeclared stop rule therefore returned `stop_route`, retaining only the narrower positive comparison against global-risk controls. At larger budgets the observed utilities tie. Direct methods retain headroom below the exact optimum at some lower-budget seed combinations; the lower-budget evidence leaves relative sample efficiency unresolved. All ten seed result records were verified in the completed campaign; no unsuccessful seed was dropped.

J Frozen horizon-three adaptive witnesses

The state-only witnesses list weights in Calm, Warning, Fragile order. Finite decimals are interpreted exactly as rationals and held constant across stages.

$$w^{(1)} = (0.5718838835280007,\ 0.9624182863377314,\ 0.7783916201390906),$$
$$w^{(2)} = (0.5507546549864419,\ 0.7430258162303255,\ 0.9509500099373207),$$
$$w^{(3)} = (0.5642653262284236,\ 0.8981879513246769,\ 0.9337317701691928).$$

The candidate indices are 8, 6 and 27 at budget 109. Conservative utility-gap bounds are $41308123/500000000$ and $81101397/1000000000$; these bound all non-optimal policy tables rather than exactly optimizing over spectra.

K Preliminary validation on Currency Exchange

This appendix preserves the initial estimator-validation study that preceded the learned risk-map comparison in Section 4.3. Its diagnostic policies test offline outcome estimation. These earlier results and limitations are retained chronologically.

Data and objective. We use the public Currency Exchange environment and random-policy dataset released with 1R2R (Rigter et al., 2023).[2] The dataset contains 499,999 transitions in 25,945 complete episodes. We verified its recorded transitions and checked adapter parity on 1,000 transitions and five resets. The proposed external selection target is lower-tail $\mathrm{CVaR}_{0.1}(G)$ of the complete undiscounted revenue $G = \sum_t r_t$, without terminal liquidation. This static episodic objective is distinct from the nested inner valuation and the additive outer utility used in the controlled experiments. This extension is evaluated empirically.

Estimator and information access. A first prototype resampled complete transition tuples from nearby state–action pairs. Such resampling can assign a neighbouring sale's revenue to a no-sale action. We therefore preserve the public inventory, reward and termination equations and estimate only the stochastic price transition. The estimator uses the public exogenous, Markov, time-homogeneous price structure, but not its drift or volatility parameters. It samples next prices from 32 nearest observed prices, pooling across time, inventory and action. Price distances are standardized using training data; queries whose 32nd neighbour lies farther than one training standard deviation are unsupported. The guard defines a heuristic support diagnostic. Any future model-based comparator must receive the same structural information.

We fit on the first 512 complete episodes (9,925 transitions). The next 64 episodes serve as development checks. All 1,247 development queries meet the price-distance guard. Observed coverage of the predicted 10th, 50th and 90th next-price quantiles is 12.03%, 51.24% and 88.21%, respectively. These dependent observations assess one-step price coverage; episode-return tails receive a separate diagnostic below.

Fixed-policy return check. Before inspecting results, we fixed four diagnostic policies, 2,048 fresh episodes per policy and a budget of at most 163,840 reference transitions. Policies hold throughout, sell all immediately, sell 10% of remaining inventory each step, or sell all at time 19. Model and simulator share initial states but use independent transition streams; all probes share the reference price innovations. Reset, model, reference and bootstrap seeds are 44001, 44002, 44003 and 44004. The reference uses the parity-checked adapter. Its 124,928 transition calls are diagnostic only: no resulting samples enter training, tuning or risk-map selection.

Immediate-sale returns agree exactly, as required by accounting and shared initial states: mean revenue is 100.0563 and empirical $\mathrm{CVaR}_{0.1}$ is 90.9311. Each other probe completes 2,046 of 2,048 model episodes. The same two trajectories cross the distance guard after 14 and 17 steps; their neighbour distance is 1.02571. We retain these incomplete trajectories and withhold the full-distribution point comparison, rather than remove them or change the guard. The predeclared 500-replicate paired bootstrap therefore applies only to immediate sale, where both difference intervals are exactly zero. These intervals concern simulation error conditional on one fitted model.

[2] Environment: `https://github.com/marc-rigter/1R2R`; dataset: `https://huggingface.co/datasets/marcrigter/1R2R-datasets`. Source revision `58d6cad`; dataset revision `18afea3`.

Descriptive bounds for missing returns. As a supplementary, post hoc calculation, we bound rather than impute the missing full returns. For the sampled initial states, public accounting and price bounds imply revenue in $[0, 500]$. Replacing missing returns by either endpoint bounds empirical lower-tail CVaR by monotonicity. Table 10 reports these bounds; they are not confidence intervals or a promotion criterion.

Diagnostic policy	Model CVaR interval	Reference CVaR
Hold	$[0, 0]$	0.0000
Sell 10% each step	$[55.9829, 56.6538]$	54.7845
Sell at time 19	$[31.8625, 32.4144]$	31.9022

Table 10: Preliminary empirical return calibration at tail mass 0.1. Each interval allows either missing full return to range over $[0, 500]$; reference values use all 2,048 simulator episodes. The rows are diagnostic probes of the preliminary estimator.

The gradual-sale model sample is optimistic even under these loose bounds, but simulation and training-sample uncertainty remain unquantified for that difference. The final-step-sale reference lies inside its descriptive interval. This preliminary study validates accounting and identifies return-estimation questions. Section 4.3 reports the subsequent parametric-model risk-map comparison, including its evaluation uncertainty.