

X-Moe-S: Explainable Mixture-of-Experts Strategies to Disentangle Risk Premia in Financial Assets

Riccardo Giacomelli

`riccardo.giacomelli@nyoratech.com`

Abstract

We propose X-Moe-S, a modified mixture-of-experts (MoE) architecture to disentangle cross-sectional and term-structure risk premia into explainable tradable strategies. The architectural innovation of per-head routing into experts (in contrast to per-token routing), coupled with a diversity penalty and differentiated learning rate improves both the explainability of individual experts and exploration in strategy space (low and high dimensional). The resulting experts are a small set of interpretable base strategies, and the mixing gate allocates expert weights—regime dependent—among them. We also provide bounds implied by the implicit regularizations for non stationary optimal policy shifting. We report results on low-dimensional term-structure data (commodities and FX), a non-linear mid-dimensional cross-sectional application (Equity, FX options), and a high-dimensional application (equities stocks).

1 Introduction

Deep hedging and portfolio construction have shown that neural networks can learn trading policies directly from reward signals. While effective, these models are commonly black boxes, which limits their adoption in scientific and risk-managed settings [1]. Mixture-of-experts (MoE) models provide a principled route to modularity by combining specialist components through a gating mechanism: classical formulations cast routing as latent assignment and modern large-scale variants emphasizing sparse conditional computation [7, 2, 3]. In many financial applications, however, experts still share most of the representation pipeline (e.g., shared attention layers), leaving limited separation at the de-

cision level.

We take a different approach: experts are instantiated as independent strategy networks, while only a shared market data processing backbone is shared. This design improves debuggability and supports a stronger notion of *understandability*, because each expert can be examined as a distinct trading skill rather than as a minor perturbation inside a monolithic policy; see Section 2.1 and the orthogonal-noise SNR intuition in Section 3.4.

Our setting is naturally formulated as a partially observable Markov decision process (POMDP) [14]. For the Heath–Jarrow–Morton (HJM) term-structure model we observe only a subset of latent states; for Heston options the volatility factors are latent; and for cross-sectional equity factor models we observe prices but not the driving factors [6]. This latent-state structure makes the problem closely related to three neighboring literatures. First, it resembles *unsupervised skill discovery*, where diverse skills are learned without explicit labels and only later composed or selected [10]. Second, it is closely connected to *hierarchical reinforcement learning* and the *options* framework, in which a higher-level controller selects temporally extended skills and decides when to persist or switch [8, 9]. Third, it admits a probabilistic interpretation in which the gate represents a belief over latent market states, while experts represent reward-relevant conditional policies associated with those states [7]. In this sense, the architecture may be viewed as learning beliefs over latent regimes together with reward-specialized experts.

This perspective is particularly relevant in finance because profitable structure may be weak or even invisible after averaging across regimes: a strategy can be profitable only on a subset of latent states and neutral or adverse elsewhere. In such cases,

a single policy is naturally pulled toward an averaged behavior, whereas a modular architecture can in principle separate conditional premia across assets by coupling latent-state inference, routing, and specialist decision rules. We train these policies with PPO and Soft Actor-Critic (SAC), focusing on interpretability and on the ability to disentangle reward-bearing behaviors across latent market conditions [4, 5].

2 Architecture

Figure 1 summarizes the model.¹ Each asset is processed by a shared 1D CNN over its lookback window. The CNN uses two temporal convolution layers (e.g., $C \rightarrow 32$ with a larger kernel and stride, followed by $32 \rightarrow 64$ with a small kernel), then a linear projection maps the flattened output to a D -dimensional token. These per-asset tokens are passed through a shared multi-head attention (MHA) block with H heads that computes cross-sectional signals across assets, followed by a feed-forward layer and normalization. We then reshape the output from per-token positions to per-head representations, keeping each trading signal confined within a single head; this makes it straightforward to connect experts to specific heads.

We use two routing stages. The first is an *inverted routing* that connects MHA heads to experts, analogous in spirit to expert-choice routing where experts select inputs rather than inputs selecting experts [15]. Concretely, the MHA output is split by head; each head (or a learned slot-blend of heads) is routed to a dedicated expert MLP ($d \rightarrow 2d \rightarrow d$), producing expert-specific features in disjoint subspaces. This structural routing preserves expert separation at the representation level and avoids collapse. In standard soft MoE, the routing (dispatch) and combining weights are tied, so the same routing matrix that assigns tokens to expert slots also determines the expert mixture. We decouple these roles to improve explainability and give the gate freedom to set final long/short expert weights. The second routing stage is an *expert mixing gate* that blends expert portfolios into the final allocation. The gate takes the full cross-sectional state (flattened MHA output) and outputs soft mixing

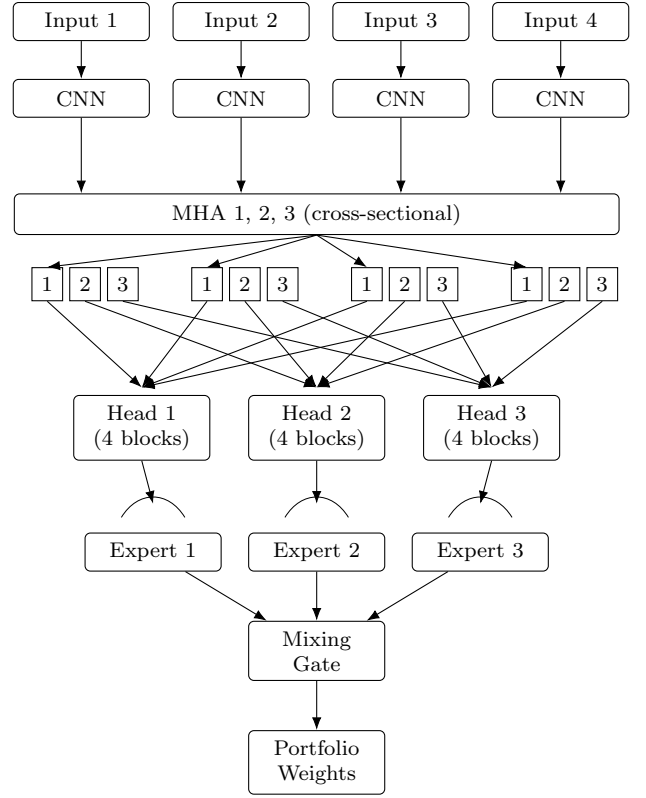


Figure 1: Overview of the architecture: per-asset CNNs, shared MHA, per-head reordering, routing into experts, and a mixing gate for portfolio construction.

weights; with a signed gate, the same expert can be flipped between long and short.

The reward is the portfolio P&L with a risk penalty,

$$R_t = \text{P\&L}_t - \lambda \text{P\&L}_t^2, \quad (1)$$

where $\lambda > 0$ controls risk aversion. The orthogonalization penalty is applied to expert outputs to promote diversity and interpretability.

2.1 Regularisation

In the network there are multiple working regularisations described below.

2.1.1 Expert Induced Regularisation

We regularize the MoE representation so that each expert corresponds to an interpretable strategy and experts do not collapse to the same trade. Formally, let $z_t \in \{1, \dots, K\}$ denote an unobserved

¹Code repository: <https://github.com/>.

latent market state and b_t a learned belief state under partial observability. The gate is interpreted as a belief over latent states,

$$g_t^{(k)} = p_\phi(z_t = k \mid b_t),$$

and expert k is a state-conditional strategy

$$w_t^{(k)} = \pi_{\theta_k}(b_t).$$

The resulting portfolio is

$$w_t = \sum_{k=1}^K g_t^{(k)} w_t^{(k)}.$$

Hence the gate performs soft regime assignment, while experts act as reward-specialized skills. Diversity regularization encourages experts to align with distinct risk and P&L drivers, improving basis coverage and allowing the gate to express a broader and more interpretable family of exposures.

Concretely, we penalize similarity between expert portfolios via pairwise cosine similarity (using a smooth-max to focus on the most collinear pair). All penalty definitions are deferred to Appendix A.1.

2.1.2 Injected Noise Regularisation

We inject explicit noise in the low-dimensional MoE coordinates (e.g., expert outputs, gate logits/weights, or the blended mean action). This turns optimization into a locally *smoothed* objective: instead of fitting a single point policy, training favors solutions that perform well in a neighborhood, reducing brittleness under small perturbations or mild regime shifts. Details and assumptions are given in Appendix A.3.

Bounded tube neighborhood (action space).

At state s , the blended mean action is

$$\mu_\theta(s) = \sum_{k=1}^K w_k(s; \theta) a_k(s; \theta) \in \mathbb{R}^m, \quad (2)$$

where m is the action dimension, $a_k(s; \theta) \in \mathbb{R}^m$ are expert actions, and $w(s; \theta) \in \Delta_K$ with $\Delta_K := \{w \in \mathbb{R}_+^K : \sum_{k=1}^K w_k = 1\}$ is the K -simplex of mixing

weights. Define $x_t := \mu_{\theta_t}(s) - \mu_\star(s)$. Under expert regularization and injected Gaussian noise (scale σ),

$$\limsup_{t \rightarrow \infty} \mathbb{E} \|x_t\|^2 \lesssim \frac{\sigma^2 m}{\mu}. \quad (3)$$

Intuition: stronger regularization (larger μ) tightens the tube, while higher noise (larger σ) widens it.

2.1.3 PPO Clipping (Update Regularisation)

PPO clipping regularizes *updates* on the policy-distribution manifold by limiting how much the action likelihood can change on rollout samples. The clipped ratio

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \quad (4)$$

stops encouraging further changes once $r_t(\theta) \notin [1 - \epsilon, 1 + \epsilon]$, which acts as an *approximate* local trust region and stabilizes learning. In contrast, expert regularization shapes *representation* by biasing the policy toward a diversified, interpretable expert dictionary. Suppose an update satisfies

$$\mathbb{E}_s[\text{KL}(\pi_{\theta_{\text{old}}}(\cdot \mid s) \parallel \pi_\theta(\cdot \mid s))] \leq \delta, \quad (5)$$

(as encouraged by clipping and optionally enforced by KL control). Then for any measurable f with $\|f\|_\infty \leq 1$,

$$\left| \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)}[f(a)] - \mathbb{E}_{a \sim \pi_{\theta_{\text{old}}}(\cdot \mid s)}[f(a)] \right| \leq \sqrt{2\delta}, \quad (6)$$

so each update induces only a small distributional change in bounded action statistics.

Why low-rank (experts) makes PPO learnable in high dimension. A key difficulty in portfolio-style rewards is that the policy-gradient signal is typically *diffuse* across many action coordinates. If the action lives in $\mathbb{R}^m$ and the update has a roughly fixed KL budget, then the typical per-coordinate mean shift scales as $\Delta\mu_i/\sigma_i = O(\sqrt{\text{KL}/m})$, i.e. it vanishes as m grows. MoE/factorized policies change the scaling by restricting actions to a K -dimensional subspace $U(s) = \text{span}\{v_k(s)\}_{k=1}^K$ with $K \ll m$: the update is

concentrated in K coordinates (gate/expert coefficients), so the same KL budget yields $\Delta\alpha_j/\sigma_{\alpha,j} = O(\sqrt{\text{KL}/K})$. Intuitively, this moves training from a regime where *data requirements scale with m* to one where they scale with *the effective dimension K* , making previously noise-dominated learning (update-SNR $\rightarrow 0$ as $m \uparrow$) feasible with the same simulator/batch size.

2.1.4 Two-Timescale Learning Rates for Gate and Experts

We utilize a two-timescale update scheme: the gate is trained with a higher learning rate, while expert networks are trained with a lower learning rate. The slow expert timescale stabilizes the expert dictionary across regime changes, so each expert retains a persistent trading identity instead of drifting with short-horizon noise. The fast gate timescale then adapts routing weights quickly as market conditions change, improving regime responsiveness without re-learning the underlying expert behaviors at each shift.

This separation is related to an EM-style division of roles only at the conceptual level. A faster gate update plays a responsibility-like role (state-dependent assignment over experts), while slower expert updates refine the underlying components under a quasi-stationary routing distribution. No exact likelihood-maximization guarantee is implied [17, 18, 19].

2.1.5 How the Three Regularizers Differ

- **Expert regularization (representation)** increases the pull-back strength μ by discouraging expert collapse and improving coverage/conditioning of the strategy dictionary.
- **Injected coordinate noise (robustness)** sets the diffusion scale σ by explicitly perturbing MoE coordinates, yielding a stationary “tube” of thickness $\limsup_t \mathbb{E}\|x_t\|^2 \lesssim \sigma^2 m / (\eta\mu)$.
- **PPO clipping (updates)** bounds per-update drift (approx. KL trust region), keeping optimization stable so the local pull-back/diffusion picture remains valid.

Together they provide coverage (experts), safe steps (PPO), and local thickness (injected noise).

2.1.6 Unforeseen Market Regimes

When the optimal policy shifts due to non-stationarity, performance depends on (i) what the expert span can represent, (ii) how stable the learning path is, and (iii) how robust the solution is within a local tubular neighborhood. Intuitively, broader shift coverage comes with a bias cost (under isotropic noise, worst-case bias scales like $d\sigma^2$), and coverage risk is controlled by overage radius: a large regime-shift loss is bounded in probability within the coverage range. Formal statements and bounds are given in Appendix A.4.

3 Results

We evaluate on three regimes to assess generality and interpretability.

3.1 RL Environment

All simulations are wrapped as Gymnasium environments and trained with Stable-Baselines3 for PPO baselines, alongside custom SAC code [11, 12]. The actor maps observations to portfolio actions, while the critic estimates the value of states (or state-action pairs) to guide policy updates; this yields stable learning under stochastic rewards. Unless noted otherwise, experiments are executed on a machine with 128 CPU cores and RTX 4090/RTX 5090 GPUs. The actor has approximately 1.1M parameters and the critic about 500k.

PPO in Financial Time Series PPO is an on-policy method that updates the policy by maximizing a clipped surrogate objective, which limits destructive policy shifts between updates [4]. Each iteration collects trajectories with the current policy and performs multiple epochs of stochastic gradient updates to both the policy and value function [13]. We choose PPO as a stable, data-efficient baseline for non-stationary financial time series, where conservative updates help avoid overfitting short-lived

regimes.

3.2 Risk-Neutral vs Objective Dynamics

All experiments compare simulation under the risk-neutral (arbitrage-free) measure $\mathbb{Q}$ with an objective measure $\mathbb{P}$ where risk premia are introduced. The learning task is to infer and exploit these premia from observable price dynamics alone (partial observability), rather than from latent factor states. Concretely, we simulate a baseline under $\mathbb{Q}$ and an alternative under $\mathbb{P}$, and evaluate whether the policy can identify and time the premia embedded in prices.

3.3 Numerical Results

We summarize out-of-sample performance with annualized return, annualized volatility, Sharpe ratio, maximum drawdown (MDD), and turnover. Results are reported for both simulation measures ($\mathbb{Q}$ and $\mathbb{P}$), with mean and standard deviation across seeds.

3.3.1 Ablation studies

We report targeted ablations to isolate the contribution of gate design choices (learned gate, frozen gate, and average gate) and compare their effect on return and Sharpe ratio across HJM and Heston.

3.4 Applications

3.4.1 Term-Structure Portfolios (HJM)

We model the curve with an affine HJM specification,

$$\begin{aligned} \log f(t, \tau) = & \log F_0(t + \tau) \\ & + \sum_k \psi_k(t, \tau) \\ & + \sum_k g_k(t) \\ & + \sum_k u_k(\tau) f_k(t), \end{aligned} \quad (7)$$

where $u_k(\tau) = A_k e^{-\alpha_k \tau} + B_k$ are loading shapes and $f_k(t)$ are mean-reverting latent factors. For fixed-

maturity contracts T , the futures price $F(t, T) = f(t, T - t)$ satisfies analogous dynamics. Under $\mathbb{Q}$,

$$\begin{aligned} d \log f(t, \tau) = & \mu^{\mathbb{Q}}(t, \tau) dt \\ & + \sigma_f(t, \tau) dW_t^{\mathbb{Q}}, \end{aligned} \quad (8)$$

and

$$\begin{aligned} d \log F(t, T) = & \mu_F^{\mathbb{Q}}(t, T) dt \\ & + \sigma_F(t, T) dW_t^{\mathbb{Q}}. \end{aligned} \quad (9)$$

Under $\mathbb{P}$ we add a state-dependent risk premium,

$$\begin{aligned} d \log f(t, \tau) = & (\mu^{\mathbb{Q}}(t, \tau) + \sigma_f(t, \tau) \lambda(t)) dt \\ & + \sigma_f(t, \tau) dW_t^{\mathbb{P}}, \end{aligned} \quad (10)$$

and

$$\begin{aligned} d \log F(t, T) = & (\mu_F^{\mathbb{Q}}(t, T) + \sigma_F(t, T) \lambda(t)) dt \\ & + \sigma_F(t, T) dW_t^{\mathbb{P}}, \end{aligned} \quad (11)$$

with a nonlinear specification such as $\lambda_k(t) = \lambda_k f_k(t)$ or indicator/magnitude-based variants. The agent observes only the curve (futures prices on a maturity grid), so the state is partially observed. Head-wise routing yields distinct curve-shape experts (level, slope, curvature). The diversity penalty separates these experts, and the gate switches among them as the curve regime changes.

The simulator uses a three-factor setup on an equally spaced grid of eight zero-coupon futures with maturities from 0.25y to 2.0y (step 0.25y). Under $\mathbb{Q}$, each latent factor follows an OU process,

$$df_k(t) = -\alpha_k f_k(t) dt + \sigma_k dW_k(t), \quad (12)$$

and under $\mathbb{P}$ we add a sign-sensitive risk-premium drift through the same loading basis, $drp_k(t) = \sigma_k \lambda_k f_k(t) dt$. This creates a mean-reversion trading signal: long when $f_k > 0$, short when $f_k < 0$, with magnitude increasing in $|f_k|$. Factor 1 (level) is deliberately non-exploitable ($\lambda_1 = 0$), Factor 2 (slope) loads monotonically across tenor, and Factor 3 (curvature) concentrates around the belly of the curve.

In training, combining continuous weight decay with a lower learning rate for experts than for the gate yields stable expert identities and state-dependent routing. The three experts remain consistent across regimes (Table 2): E0 is short directional and short-butterfly, E1 is long directional and long-butterfly, and E2 is belly-focused long. Hence

Table 1: HJM simulation parameters (three-factor setup).

Parameter	F1	F2	F3
OU speed α_k	5.0	5.0	9.0
Volatility σ_k	0.05	0.25	0.10
Risk premium λ_k	0	5	20
Loading decay β_k	0.05	0.50	0.60
Stationary sd $\sigma_k/\sqrt{2\alpha_k}$	0.016	0.079	0.024
Half-life days $\ln(2)/\alpha_k$	35	35	19

experts form a fixed basis of trades, while adaptation is primarily performed by the gate.

The gate switches mainly across curvature states f_3 . In low- f_3 states it loads on the long experts ($E1, E2$), producing a strongly long blended portfolio; in high- f_3 states it shifts toward $E0$ and away from ($E1, E2$), producing a strongly short portfolio. Across slope states f_2 , the gate induces a weaker short-to-long rotation, again dominated by directional reweighting rather than a pure spread trade. Overall, this confirms the intended mechanism: experts stay stable and the gate performs regime switching across both direction and curve shape. Table 3 summarizes the performance gains.

Table 2: HJM expert identities.

Expert	Main identity
E0	Short directional, short butterfly
E1	Long directional, long butterfly
E2	Belly-focused long

Table 3: HJM main results.

Metric	Value
Annualized Sharpe	3.60
Gain vs frozen gate	+3.41
Gain vs equal-weight baseline	+3.74

3.4.2 Cross-Sectional Options (Heston)

On mid-dimensional options portfolios, we simulate prices under $\mathbb{Q}$ with a two-factor Heston model. The risk premium is introduced under $\mathbb{P}$ by shifting the variance dynamics with a time- and state-dependent market price of volatility risk. A com-

pact form is

$$\begin{aligned} dV_i &= \kappa_i(\theta_i - V_i) dt + \sigma_i \sqrt{V_i} dW_t^{V,i,\mathbb{Q}}, \\ dV_i &= \kappa_i(\theta_i - V_i) dt + \sigma_i \sqrt{V_i} (dW_t^{V,i,\mathbb{P}} + \lambda_i(t, V_i) dt), \end{aligned} \quad (13)$$

with $\lambda_i(t, V_i)$ possibly proportional to V_i or other state variables. The model observes a bucketed option menu plus a spot token; for each bucket we include moneyness, delta, gamma, theta, implied volatility, price/spot, and surface factors (level, skew, curvature), while we omit the latent-factor vegas. We also feed levered Greeks (delta, gamma, theta) scaled by a target notional so the network sees per-contract sensitivities in dollar terms; this makes the inputs comparable across strikes and maturities and aligns the state with risk limits and P&L attribution. The factor-space diversity penalty uses a whitened sensitivity matrix $\tilde{S}$ built from per-bucket Greeks (excluding $vega_{V1}, vega_{V2}$) and penalizes similarity among expert portfolios:

$$\mathcal{L}_{\text{div}} \propto \log \sum_{i < j} \exp \left(\frac{\cos^2(\tilde{S}w_i, \tilde{S}w_j)}{\tau} \right). \quad (14)$$

Experts output bucket weights and a spot weight; a scale head controls gross exposure while a separate spot head adjusts net delta. The spot weight lets experts intentionally over- or under-hedge within a bounded range, which is critical for learning volatility-risk premia without taking unlimited directional exposure. We enforce hard caps on gross and net delta via a two-stage projection, so $\sum_b |w_b| \omega_b \leq \Delta_{\text{max}}^{\text{gross}}$ and $|\Delta_{\text{net}}| \leq \Delta_{\text{max}}^{\text{net}}$, where ω_b is the delta leverage per bucket. This yields stable long-short portfolios with interpretable surface exposures.

Table 4 summarizes the calibrated two-factor Heston channels under $\mathbb{P}$ and $\mathbb{Q}$. The dominant shift appears in the first-factor skew channel $\rho_1 \sigma_1$, which is $4.6\times$ larger under $\mathbb{Q}$; this concentrates the volatility risk premium (VRP) in downside skew and, operationally, in OTM put exposures.

Table 6 reports expert-level exposures. E0 is an upside-wing buyer (OTM-call convexity), while E3 is a broad convexity buyer (long ATM calls and long put wing, with trimmed call wing). E1 is a wing seller (strangle) and E2 is a risk-reversal sleeve. The blend remains close to delta-neutral while preserving positive gamma. For readability in two-column

Table 4: Two-factor Heston channels under physical and risk-neutral measures.

Parameter	$\mathbb{P}$	$\mathbb{Q}$	Ratio ($\mathbb{Q}/\mathbb{P}$)
θ_1	0.040	0.060	1.50
σ_1	0.55	0.80	1.45
ρ_1	-0.35	-0.62	1.77
$\rho_1\sigma_1$	-0.193	-0.496	2.57
θ_2	0.020	0.024	1.20
σ_2	0.30	0.33	1.10
ρ_2	-0.20	-0.22	1.10

format, the full table is reported in Appendix Table 6.

Gate allocation is regime aware (Table 7): after spot crashes, routing tilts strongly to E1 (wing seller) with a secondary increase in E3 (broad convexity buyer), while E0 and E2 are reduced. In calm markets, E0 and E2 increase as OTM-call convexity and directional risk-reversal exposure are reloaded. Time-to-expiry adds an orthogonal dimension: Month 1 is E3-led (with E1 second), whereas Month 3 rotates to E2 dominance with a higher E0 weight. The full routing table is reported in Appendix Table 7.

Finally, Table 5 reports the Greek decomposition used in our backtest accounting. The profile is consistent with VRP capture through realized-variance convexity (large gamma), offset by correlation/hedging frictions (vanna and spot), while net theta remains close to zero.

Table 5: Greek-level P&L attribution (contribution to option strategy P&L, %).

Component	Contribution (%)
Gamma	+106
Vanna (v_1)	-31
Vega	-6
Theta	-1.5
Delta	+3

3.4.3 Cross-Sectional Equities (Factor Model)

On high-dimensional equity portfolios, we simulate returns under $\mathbb{Q}$ with zero factor premia and add factor risk premia under $\mathbb{P}$ through a nonzero mean of factor returns. Orthogonalization in weight space prevents redundant experts and encourages special-

ization tied to latent factors. The gate dynamically reallocates among experts as factor regimes shift, providing interpretable decompositions while preserving overall portfolio quality.

4 Conclusion

We introduce Moe-XS, a modular MoE architecture for cross-sectional RL that separates experts to improve interpretability while retaining performance. Our training design—continuous weight decay together with a lower learning rate for experts than for the gate—stabilizes expert identities across regime changes. Furthermore diversity penalisation avoids experts collapse and push them to span different aspect of prices dynamic. The gate adaptively reweights these base experts as market states evolve. Adaptation is primarily realized through routing over stable, interpretable expert strategies, enabling the model to learn profitable policies without sacrificing transparency.

5 Future Work

Future directions include dynamic expert creation, causal attribution of expert signals, and tighter links between orthogonalization and risk-factor decomposition. Extending to transaction costs and multi-objective rewards is also of interest.

A Appendix: Additional Heston Tables

Table 6: Expert strategy weights and portfolio Greeks (per unit notional).

Expert	$P:OTM$	$P:ATM$	$C:ATM$	$C:OTM$	γ	θ	Vega	Strategy
E0	-0.02	+0.07	+0.05	+0.31	+2.6	-6.7	+4.5	Upside wing buyer (OTM calls)
E1	-0.28	+0.15	+0.09	-0.07	-1.7	+3.6	-2.3	Wing seller (strangle)
E2	-0.16	-0.20	+0.19	+0.01	+0.2	+1.0	-1.8	ATM risk reversal
E3	+0.21	+0.15	+0.24	-0.03	+2.1	-6.2	+5.8	Broad convexity buyer (ex OTM Call)
Blend	-0.26	-0.02	+0.21	+0.04	+0.5	≈ 0	-0.7	Short skew, long gamma

Table 7: Gate behavior by regime and horizon.

Slice	E0	E1	E2	E3	Interpretation
Crash regime	0.057	0.525	0.187	0.232	E1 up on rich wings; E3 up as broad convexity sleeve
Calm regime	0.280	0.090	0.510	0.120	E0/E2 reload: cheap OTM-call convexity and directional RR
Month 1 buckets	0.18	0.29	0.18	0.35	Early tenor: E3 dominates, E1 second
Month 3 buckets	0.22	0.23	0.40	0.15	Late tenor: E2 dominates, E0 rises

References

- [1] H. Buehler, L. Gonon, J. Teichmann, and B. Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019.
- [2] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.
- [3] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120):1–39, 2022.
- [4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.
- [5] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.
- [6] R. C. Grinold and R. N. Kahn. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*. McGraw-Hill, 1999.
- [7] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.
- [8] R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999.
- [9] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017.
- [10] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019.
- [11] E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *IROS*, 2012.
- [12] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.
- [13] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press, 2018.
- [14] M. J. Kochenderfer, T. A. Wheeler, and K. H. Wray. *Algorithms for Decision Making*. MIT Press, 2022.
- [15] Y. Zhou, S. Li, S. Li, C. Huo, and D. Chen. Expert choice routing: Solving load balancing in Mixture-of-Experts without auxiliary loss. In *NeurIPS*, 2022.
- [16] T. M. Cover and J. A. Thomas. *Elements of Information Theory* (2nd ed.). Wiley-Interscience, 2006.
- [17] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.
- [18] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.
- [19] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Springer, 1998.

A Model Details

A.1 Experts as Manifold Regularization: Proof

We provide the details behind Section 2.1.1. The argument formalizes how expert diversity regularizes the parameter dynamics around a low-dimensional manifold.

Setup and assumptions. Let $\theta \in \mathbb{R}^p$ and let $\mathcal{M} \subset \mathbb{R}^p$ be a k -dimensional, smooth embedded manifold of interpretable expert configurations. In a tubular neighborhood of $\mathcal{M}$, choose local coordinates $(\theta_{\parallel}, \theta_{\perp})$ such that $\theta_{\perp} = 0$ characterizes $\mathcal{M}$. Assume:

1. The expert-diversity penalty is twice differentiable and satisfies

$$R(\theta) = \frac{\lambda}{2} \|\theta_{\perp}\|^2 + o(\|\theta_{\perp}\|^2), \quad (15)$$

for some $\lambda > 0$ in a neighborhood of $\mathcal{M}$.

2. The learning dynamics are approximated by the continuous-time SDE

$$d\theta_t = -\nabla L(\theta_t) dt - \nabla R(\theta_t) dt + \sqrt{2T} dW_t, \quad (16)$$

where L is the task loss, $T > 0$ is an effective noise temperature, and W_t is standard Brownian motion.

Normal dynamics. Let $P_\perp$ denote the orthogonal projector onto the normal space of $\mathcal{M}$. Applying $P_\perp$ to (16) and using $\nabla R(\theta) = \lambda \theta_\perp + o(\|\theta_\perp\|)$ yields

$$\begin{aligned} d\theta_{\perp,t} &= -\lambda \theta_{\perp,t} dt \\ &\quad - P_\perp \nabla L(\theta_t) dt \\ &\quad + \sqrt{2T} dW_t^\perp + o(\|\theta_{\perp,t}\|) dt. \end{aligned} \quad (17)$$

The $P_\perp \nabla L(\theta_t)$ term couples tangential and normal dynamics but is typically bounded on compact sets. Over a local time window (or under weak coupling), the dominant behavior in the normal directions is therefore the linear drift $-\lambda \theta_{\perp,t}$.

Stationary variance bound. Ignoring the coupling term in (17) yields the Ornstein–Uhlenbeck process

$$d\theta_{\perp,t} = -\lambda \theta_{\perp,t} dt + \sqrt{2T} dW_t^\perp, \quad (18)$$

whose invariant distribution is $\mathcal{N}(0, (T/\lambda)I)$. Consequently,

$$\mathbb{E}\|\theta_\perp\|^2 = \frac{T}{\lambda} \dim(\theta_\perp). \quad (19)$$

With the coupling term included, the same scaling persists up to constants depending on the local Lipschitz bound of ∇L in the normal directions. This establishes that increasing λ contracts the parameter trajectory toward $\mathcal{M}$, i.e., the experts act as a manifold regularizer.

Discrete-time form For the OU approximation, the second moment satisfies

$$\frac{d}{dt} \mathbb{E}\|\theta_{\perp,t}\|^2 = -2\lambda \mathbb{E}\|\theta_{\perp,t}\|^2 + 2T d_\perp, \quad (20)$$

so the deterministic pull-back is exponential with rate 2λ . Using an Euler step with step size η ,

$$\theta_{\perp,t+1} \approx (I - \eta\lambda)\theta_{\perp,t} + \sqrt{2T\eta}\varepsilon_t, \quad (21)$$

which implies

$$\mathbb{E}\|\theta_{\perp,t+1}\|^2 \leq (1 - 2\eta\lambda) \mathbb{E}\|\theta_{\perp,t}\|^2 + C(T\eta) d_\perp. \quad (22)$$

This matches the contraction-plus-noise template used in the main text, with effective contraction $\mu \approx 2\lambda$ (up to discretization constants) and diffusion scale $\sigma^2 \sim T\eta$.

A.2 PPO Update Regularization: Sketch

PPO regularizes policy updates by clipping the likelihood ratio [4]:

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, \quad (23)$$

with clipped surrogate

$$\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t). \quad (24)$$

When $r_t(\theta)$ exits $[1 - \epsilon, 1 + \epsilon]$, the objective stops rewarding further movement on those samples.

Clipping does not impose an explicit KL constraint, but it encourages updates with small average KL, and many implementations optionally add KL-based early stopping or penalties. For small parameter steps $\Delta\theta = \theta - \theta_{\text{old}}$,

$$\mathbb{E}_s[\text{KL}(\pi_{\theta_{\text{old}}}(\cdot | s) \parallel \pi_{\theta_{\text{old}} + \Delta\theta}(\cdot | s))] \approx \frac{1}{2} \Delta\theta^\top F(\theta_{\text{old}}) \Delta\theta. \quad (25)$$

Hence a small-KL update stays in a local neighborhood around $\pi_{\theta_{\text{old}}}$ (approximately $\Delta\theta^\top F(\theta_{\text{old}}) \Delta\theta \leq 2\delta$ when the average KL is bounded by δ). This is the only role needed here: PPO stabilizes optimization drift, while expert regularization shapes representation.

A.3 Noise on the Manifold: Details

We formalize noise as a regularizer that smooths the objective and induces local robustness. To make the gate explicit, write the policy mean as

$$\mu(s) = V(s)w(s), \quad w(s) \in \Delta_K, \quad (26)$$

where $V(s) = [v_1(s) \cdots v_K(s)] \in \mathbb{R}^{m \times K}$ and Δ_K is the K -simplex. Let $\alpha(s)$ denote a low-dimensional gate coordinate with decoder $w(s) = h(s, \alpha(s))$. Inject noise in coordinate space,

$$\tilde{\alpha}(s) = \alpha(s) + \xi, \quad \xi \sim \mathcal{N}(0, \Sigma_\alpha), \quad (27)$$

and act with $\pi_\theta(\cdot | s)$ whose mean is $\mu(s) = V(s)h(s, \tilde{\alpha}(s))$. This perturbs the gate/expert mixing itself (stronger than adding only action-level noise after blending/projection) and corresponds to training on a tubular neighborhood

$$T_\varepsilon(\alpha) = \{\alpha + \Delta : \|\Delta\| \leq \varepsilon\}, \quad \varepsilon^2 \approx \text{tr}(\Sigma_\alpha). \quad (28)$$

Let $J_z(\alpha)$ denote the return objective in regime z , viewed as a function of the manifold coordinate. With noise, the optimized objective becomes the smoothed functional

$$J_z^{\text{smooth}}(\alpha) = \mathbb{E}_\xi[J_z(\alpha + \xi)]. \quad (29)$$

More generally, if the per-regime objective is $J(\theta) = \mathbb{E}[\text{return under } \pi_\theta]$ and noise is injected in parameters or coordinates, then

$$J^{\text{noise}}(\theta) = \mathbb{E}_\xi[J(\theta, \xi)], \quad (30)$$

so noise regularizes by smoothing sharp spikes in J and favoring flatter solutions.

If J_z is L -smooth in α over the tubular neighborhood, then

$$J_z(\alpha) - \mathbb{E}_\xi[J_z(\alpha + \xi)] \leq \frac{L}{2} \mathbb{E}\|\xi\|^2 = \frac{L}{2} \text{tr}(\Sigma_\alpha), \quad (31)$$

and for any perturbation $\Delta\alpha$ in the tubular neighborhood,

$$J_z(\alpha) - J_z(\alpha + \Delta\alpha) \lesssim \frac{L}{2} \|\Delta\alpha\|^2. \quad (32)$$

Equivalently, a Taylor bound gives

$$J_z(\alpha + \Delta\alpha) \geq J_z(\alpha) + \nabla J_z(\alpha)^\top \Delta\alpha - \frac{L}{2} \|\Delta\alpha\|^2, \quad (33)$$

so near a stationary point the degradation is at most quadratic in $\|\Delta\alpha\|$. If J is twice differentiable, a Taylor expansion gives the second-order view:

$$\mathbb{E}_\xi[J(\alpha + \xi)] \approx J(\alpha) + \frac{1}{2} \text{tr}(\nabla^2 J(\alpha) \Sigma), \quad (34)$$

so optimizing the noisy objective adds an implicit curvature penalty that discourages overly sharp solutions. Thus nonzero noise (or a variance floor) forces learning to be good on average in a neighborhood, discouraging brittle point solutions and smoothing routing boundaries.

A.4 Unforeseen Regimes and Non-Stationarity

We formalize how regime shifts affect performance and how expert diversity, manifold noise, and PPO regularization enter the bounds.

Setup. Let regimes be $z \in \mathcal{Z}$. At state s , define a target signature $v^*(s, z) \in \mathbb{R}^m$ (e.g., weights, exposures, or P&L signature). The MoE produces feasible signatures in the expert span

$$\begin{aligned} U(s) &= \text{span}\{v_1(s), \dots, v_K(s)\}, \\ V(s) &= [v_1(s) \cdots v_K(s)] \in \mathbb{R}^{m \times K}, \\ v_\pi(s) &= V(s)w(s), \quad w(s) \in \Delta_K. \end{aligned} \quad (35)$$

Assume the regime- z utility is locally quadratic around v^* :

$$J_z(s, v^*(s, z)) - J_z(s, v) \leq \frac{L}{2} \|v - v^*(s, z)\|^2. \quad (36)$$

Define the regime-shift loss

$$B(z \rightarrow z') := \mathbb{E}_s[J_{z'}(s, v^*(s, z')) - J_{z'}(s, v_\pi(s))]. \quad (37)$$

Proposition 1 (Unavoidable orthogonal miss). Define the irreducible component

$$\epsilon_\perp(s, z')^2 := \inf_{v \in U(s)} \|v - v^*(s, z')\|^2 = \|P_{U(s)}^\perp v^*(s, z')\|^2. \quad (38)$$

Then any policy with $v_\pi(s) \in U(s)$ satisfies

$$B(z \rightarrow z') \geq \frac{L}{2} \mathbb{E}_s \epsilon_\perp(s, z')^2. \quad (39)$$

Sketch. Apply (A1) and the definition of projection distance.

Proposition 2 (Diversity improves conditioning of the expert dictionary). Fix a state s and define $V(s) = [v_1(s) \cdots v_K(s)] \in \mathbb{R}^{m \times K}$. Assume the diversity regularizer enforces

$$\max_{i \neq j} \frac{\langle v_i(s), v_j(s) \rangle}{\|v_i(s)\| \|v_j(s)\|} \leq \rho, \quad \rho < 1, \quad (40)$$

and $\|v_k(s)\| \geq c > 0$ for all k . Then

$$\sigma_{\min}(V(s)) \geq \underline{\sigma}(c, \rho, K, m) > 0, \quad (41)$$

and any constant with scaling $C_\parallel \propto \sigma_{\min}(V(s))^{-2}$ satisfies

$$C_\parallel^{\text{MoE-div}}(s) \leq C_\parallel^{\text{no-div}}(s), \quad (42)$$

whenever the non-diverse baseline has higher coherence (larger ρ) or smaller $\sigma_{\min}(V(s))$. *Interpretation.* Diversity prevents expert collapse, improves conditioning, and lowers sensitivity of the trackable component.

Proposition 3a (Noise yields a tubular neighborhood robustness bound). Let $\alpha(s) \in \mathbb{R}^d$ be the manifold coordinate with decoder $v = h(s, \alpha)$. Inject noise $\tilde{\alpha} = \alpha + \xi$, $\xi \sim \mathcal{N}(0, \Sigma)$, and define the smoothed objective

$$\bar{J}_z(\alpha) := \mathbb{E}_\xi \mathbb{E}_s [J_z(s, h(s, \alpha + \xi))]. \quad (43)$$

If $J_z(s, h(s, \alpha))$ is L_α -smooth in α , and the optimal coordinate shifts by $\Delta^* = \alpha^*(z') - \alpha^*(z)$, then locally

$$\bar{J}_{z'}(\alpha^*(z')) - \bar{J}_{z'}(\alpha^*(z)) \leq \frac{L_\alpha}{2} \|\Delta^*\|^2. \quad (44)$$

Sketch. Apply a Taylor bound to $\bar{J}_{z'}$; smoothing preserves differentiability and curvature bounds.

Lemma (Smoothing bias). Let $f_z(\alpha) := \mathbb{E}_s [J_z(s, h(s, \alpha))]$ and $\bar{f}_z(\alpha) := \mathbb{E}_\xi [f_z(\alpha + \xi)]$. If f_z is twice differentiable and $\|\nabla^2 f_z(\alpha)\| \leq M$ locally, then for any α ,

$$0 \leq \bar{f}_z(\alpha) - f_z(\alpha) \leq \frac{M}{2} \text{tr}(\Sigma). \quad (45)$$

Sketch. Use a second-order Taylor expansion with $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi\xi^\top] = \Sigma$.

Proposition 3b (Tubular neighborhood radius from drift-diffusion dynamics). Let $\alpha_t \in \mathbb{R}^d$ denote the stochastic manifold coordinate used by the policy/gate during training (or online adaptation). Assume updates around a reference $\bar{\alpha}_t$ satisfy

$$\alpha_{t+1} - \bar{\alpha}_{t+1} = (I - \eta H)(\alpha_t - \bar{\alpha}_t) + \sigma \varepsilon_t + \eta \zeta_t, \quad \varepsilon_t \sim \mathcal{N}(0, I_d), \quad (46)$$

where $H \succeq \mu I_d$ with contraction rate $\mu > 0$, σ is the injected noise scale, and ζ_t is a bounded coupling perturbation with $\mathbb{E}\|\zeta_t\|^2 \leq G^2$. For $\eta \leq 1/\|H\|$,

$$\mathbb{E}\|\alpha_{t+1} - \bar{\alpha}_{t+1}\|^2 \leq (1 - \eta\mu) \mathbb{E}\|\alpha_t - \bar{\alpha}_t\|^2 + \sigma^2 d + \eta^2 G^2. \quad (47)$$

Hence the stationary tubular neighborhood width is bounded by

$$r^2 := \limsup_{t \rightarrow \infty} \mathbb{E}\|\alpha_t - \bar{\alpha}_t\|^2 \leq \frac{\sigma^2 d + \eta^2 G^2}{\eta\mu}. \quad (48)$$

Interpretation. Larger noise increases the explored neighborhood radius r , while stronger restoring drift (μ) shrinks it; r gives an explicit coverage scale for regime shifts in α -space.

Proposition 4 (PPO KL bounds learning drift). PPO approximately enforces

$$\mathbb{E}_s [\text{KL}(\pi_{\theta_{\text{old}}}(\cdot | s) \parallel \pi_\theta(\cdot | s))] \leq \delta. \quad (49)$$

and for small steps,

$$\mathbb{E}[\text{KL}] \approx \frac{1}{2} \Delta\theta^\top F(\theta) \Delta\theta. \quad (50)$$

Thus each update stays within an ellipsoid in the Fisher metric. PPO stabilizes optimization drift, while expert regularization shapes representation.

Proposition 5 (Regime-shift bound). Fix a regime shift $z \rightarrow z'$ and a state distribution over s . Let $U(s) = \text{span}\{v_1(s), \dots, v_K(s)\}$ and decompose

$$\begin{aligned} v^*(s, z') &= v_{\parallel}^*(s, z') + v_{\perp}^*(s, z'), \\ v_{\parallel}^*(s, z') &\in U(s), \quad v_{\perp}^*(s, z') \perp U(s), \end{aligned} \quad (51)$$

with $\epsilon_{\perp}(s, z') := \|v_{\perp}^*(s, z')\|$. Assume the local quadratic upper bound (as in Proposition A.4)

$$J_{z'}(s, v^*(s, z')) - J_{z'}(s, v) \leq \frac{L}{2} \|v - v^*(s, z')\|^2. \quad (52)$$

Then for any policy with $v_\pi(s) \in U(s)$,

$$\begin{aligned} B(z \rightarrow z') &\leq \frac{L}{2} \mathbb{E}_s \|v_\pi(s) - v_{\parallel}^*(s, z')\|^2 \\ &\quad + \frac{L}{2} \mathbb{E}_s \epsilon_{\perp}(s, z')^2. \end{aligned} \quad (53)$$

Moreover, under Proposition A.4 (conditioning in the expert span) and Proposition A.4 (smooth shift in gate-coordinate space), there exist constants $C_{\text{gate}}, C_{\text{dict}}, C_{\text{tub}} > 0$ such that

$$\begin{aligned} \mathbb{E}_s \|v_\pi(s) - v_{\parallel}^*(s, z')\|^2 &\leq C_{\text{gate}} \mathbb{E}_s \|\Delta w^*(s)\|^2 \\ &\quad + C_{\text{dict}} \mathbb{E}_s \|\Delta V(s) w^*(s, z)\|^2 \\ &\quad + C_{\text{tub}} \|\Delta^*\|^2, \end{aligned} \quad (54)$$

where $w^*(s, z) \in \arg \min_{w \in \Delta_K} \|V(s)w - v^*(s, z)\|$, $\Delta w^*(s) := w^*(s, z') - w^*(s, z)$, $\Delta V(s) := V_{z'}(s) - V_z(s)$, and $\Delta^* := \alpha^*(z') - \alpha^*(z)$. Combining (53) and (54) yields

$$\begin{aligned} B(z \rightarrow z') &\leq \underbrace{C_{\text{gate}} \mathbb{E}_s \|\Delta w^*(s)\|^2}_{\text{gate shift}} \\ &\quad + \underbrace{C_{\text{dict}} \mathbb{E}_s \|\Delta V(s) w^*(s, z)\|^2}_{\text{expert-dictionary shift}} \\ &\quad + \underbrace{C_{\text{tub}} \|\Delta^*\|^2}_{\text{tube/noise term}} \\ &\quad + \underbrace{C_{\perp} \mathbb{E}_s \epsilon_{\perp}(s, z')^2}_{\text{irreducible orthogonal miss}}, \end{aligned} \quad (55)$$

with $C_\perp = L/2$. Finally, Proposition A.4 together with Lemma ?? implies small per-update distributional drift,

$$\sup_{\|f\|_\infty \leq 1} \left| \mathbb{E}_{\pi_{\theta_{t+1}}}[f] - \mathbb{E}_{\pi_{\theta_t}}[f] \right| \leq \sqrt{2\delta}, \quad (56)$$

which stabilizes the coupled evolution of $V(s)$ (experts) and $\alpha(s)$ (gate coordinates), but does not remove the orthogonal term.

Define

$$R(z \rightarrow z') := C_{\text{gate}} \mathbb{E}_s \|\Delta w^*(s)\|^2 + C_{\text{dict}} \mathbb{E}_s \|\Delta V(s) w^*(s, z)\|^2 + C_\perp \mathbb{E}_s \epsilon_\perp(s, z)^2. \quad (57)$$

Then (55) rewrites as

$$B(z \rightarrow z') \leq C_{\text{tub}} \|\Delta^*\|^2 + R(z \rightarrow z'). \quad (58)$$

Corollary 5.1 Covered-shift bound. If coverage ensures $\|\Delta^*\|^2 \leq r^2$ (cf. Proposition A.4), then

$$B(z \rightarrow z') \leq C_{\text{tub}} r^2 + R(z \rightarrow z'). \quad (59)$$

If Δ^* is random under the test-regime distribution and r_ϵ satisfies $\mathbb{P}(\|\Delta^*\| \leq r_\epsilon) \geq 1 - \epsilon$, then

$$\mathbb{P}(B(z \rightarrow z') \leq C_{\text{tub}} r_\epsilon^2 + R(z \rightarrow z')) \geq 1 - \epsilon. \quad (60)$$

Moreover, for any $r > 0$,

$$\mathbb{P}(B(z \rightarrow z') > C_{\text{tub}} r^2 + R(z \rightarrow z')) \leq \mathbb{P}(\|\Delta^*\| > r). \quad (61)$$

Thus extreme-shift risk is quantified by the tail $\mathbb{P}(\|\Delta^*\| > r)$ (or its empirical estimate).

Proposition 6 (Diversity-tightened regime-shift bound). Consider two dictionaries evaluated under the same shift ($z \rightarrow z'$): a diverse MoE dictionary and a non-diverse baseline. Under Proposition A.4,

$$C_{\parallel}^{\text{MoE-div}}(s) \leq C_{\parallel}^{\text{no-div}}(s). \quad (62)$$

If the tube term and orthogonal term are matched (same $C_{\text{tub}}, \Delta^*, \epsilon_\perp$), then from (55)

$$B^{\text{MoE-div}}(z \rightarrow z') \leq B^{\text{no-div}}(z \rightarrow z'). \quad (63)$$

Equivalently, in the high-probability form of Corollary 5.1,

$$\begin{aligned} \mathbb{P}(B^{\text{MoE-div}}(z \rightarrow z') \leq C_{\text{tub}} r_\epsilon^2 + R^{\text{MoE-div}}(z \rightarrow z')) \\ \geq \mathbb{P}(B^{\text{no-div}}(z \rightarrow z') \\ \leq C_{\text{tub}} r_\epsilon^2 + R^{\text{no-div}}(z \rightarrow z')), \end{aligned} \quad (64)$$

with $R^{\text{MoE-div}} \leq R^{\text{no-div}}$ when diversity improves the gate-conditioning constant (i.e., the C_{gate} term).

Injected coordinate noise (Appendix A.3) smooths the objective and introduces a bias gap of order $\mathcal{O}(\text{tr } \Sigma)$ (Lemma “Smoothing bias”), trading peak performance for robustness.

A.5 HJM Term Structure (Affine Form)

We use an affine HJM specification for the constant-maturity futures curve,

$$\begin{aligned} \log f(t, \tau) = & \log F_0(t + \tau) \\ & + \sum_k \psi_k(t, \tau) \\ & + \sum_k g_k(t) \\ & + \sum_k u_k(\tau) f_k(t), \end{aligned} \quad (65)$$

where $\log f(t, \tau)$ is the log forward/futures curve at time t and time-to-maturity τ , $F_0(\cdot)$ is the initial curve, $\psi_k(t, \tau)$ are deterministic convexity-adjustment terms, $g_k(t)$ are deterministic drift components, $u_k(\tau)$ are loading shapes, and $f_k(t)$ are latent factors. The affine structure keeps the dynamics Markovian in a small set of state variables. We use loading shapes $u_k(\tau) = A_k e^{-\alpha_k \tau} + B_k$. For fixed-maturity contracts T , $F(t, T) = f(t, T - t)$ and

$$\begin{aligned} \log F(t, T) = & \log F_0(T) + \sum_k \psi_k(t, T - t) \\ & + \sum_k g_k(t) + \sum_k u_k(T - t) f_k(t). \end{aligned} \quad (66)$$

The factors $f_k(t)$ follow OU dynamics under the objective measure,

$$df_k = (\sigma_k \lambda_k - \alpha_k f_k) dt + \sigma_k dW_k, \quad (67)$$

so f_k mean-reverts to $\theta_k = \sigma_k \lambda_k / \alpha_k$ when $\lambda_k \neq 0$. The drift-compensation terms evolve as

$$\begin{aligned} dg_k = & B_k \alpha_k f_k dt, \\ d\psi_k(t, \tau) = & -\frac{1}{2} \sigma_k^2 u_k(t + \tau)^2 dt. \end{aligned} \quad (68)$$

Main parameters are $(A_k, B_k, \alpha_k, \sigma_k, \lambda_k)$ for each factor and the initial curve $F_0(\cdot)$. Observations are prices on a discrete maturity grid; the latent factors f_k (and g_k, ψ_k) are not observed directly, so the problem is partially observable.

A.6 Heston Stochastic Volatility

The options model is a two-factor Heston system,

$$dS_t = \mu S_t dt + \sqrt{v_t} S_t dW_t^S, \quad (69)$$

$$dv_t = \kappa(\theta - v_t) dt + \xi \sqrt{v_t} dW_t^v, \quad (70)$$

with $dW_t^S dW_t^v = \rho dt$. For two factors, we use

$$\frac{dS_t}{S_t} = (r - q) dt + \sqrt{V_1} dW_t^{S,1} + \sqrt{V_2} dW_t^{S,2}, \quad (71)$$

$$dV_i = \kappa_i(\theta_i - V_i) dt + \sigma_i \sqrt{V_i} dW_t^{V,i}, \quad i = 1, 2, \quad (72)$$

with $\text{corr}(dW_t^{S,i}, dW_t^{V,i}) = \rho_i$ and the two factors independent. Under $\mathbb{P}$ we introduce a time- and state-dependent volatility risk premium by shifting the variance shocks,

$$dV_i = \kappa_i(\theta_i - V_i) dt + \sigma_i \sqrt{V_i} (dW_t^{V,i,\mathbb{P}} + \lambda_i(t, V_i) dt), \quad (73)$$

so the implied-volatility premium depends on (t, V_i) . The variance factors V_i are latent to the agent; only prices and option sensitivities are observed.

A.7 Equity Factor Model

We use a cross-sectional risk model in the spirit of Barra, combining style and sector exposures with a time-series factor process. For asset i ,

$$r_{i,t} = \alpha_i + \beta_i^\top f_t + s_i^\top g_t + \epsilon_{i,t}, \quad (74)$$

where β_i are exposures to style factors (e.g., value, momentum, size, quality, low-volatility, liquidity), s_i are sector or industry dummy exposures, and f_t, g_t are the corresponding factor returns. The factor covariance matrix is

$$\Sigma_f = \text{Cov} \left(\begin{bmatrix} f_t \\ g_t \end{bmatrix} \right), \quad (75)$$

and the idiosyncratic covariance is typically diagonal, $\Sigma_\epsilon = \text{diag}(\sigma_{\epsilon,i}^2)$.

To capture persistence and mean reversion, factor returns follow a VAR(1) or OU-type dynamics,

$$f_{t+1} = \Phi f_t + \eta_{t+1}, \quad \rho(\Phi) < 1, \quad (76)$$

with analogous dynamics for g_t . Only prices and returns are observed; factors and their covariance structure are latent and must be inferred.