MEAN-PRESERVING EXPLORATION IN CONTINUOUS-TIME REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study a fundamental question in continuous-time reinforcement learning: how can we introduce exploration while ensuring that the mean exploratory action exactly recovers the target optimal policy? Using unconstrained continuous-time portfolio selection as a canonical setting, we investigate two exploration mechanisms. First, we solve entropy-regularized exploratory control with state-independent temperatures and characterize a compatibility condition under which state-dependent temperatures preserve the optimal mean. We give an explicit state-dependent construction and show why arbitrary temperature schedules need not have this property. Second, we develop a direct Gaussian exploration formulation with prescribed state-independent and state-dependent covariances, including their mixtures, and establish exact recovery of the target policy through the exploratory mean. The common mechanism is preservation of a derivative ratio of the value function. We derive a trajectory-based actor–critic algorithm and provide a closed-form analytical example that separates the cost of exploration from the performance of the extracted policy.

1 INTRODUCTION

Exploration is essential when an agent must learn from interaction with unknown dynamics. In continuous action spaces, a common practice is to add Gaussian noise to a deterministic policy and deploy its mean after training. This practice raises a control-theoretic question: does optimizing the randomized policy preserve the deterministic decision that would have been optimal without exploration? Zero-mean noise alone does not answer this question. Although it leaves the instantaneous drift unchanged for a fixed mean action, it changes the state distribution and hence the objective being optimized.

Continuous-time portfolio selection offers a tractable setting for answering this question. The terminal quadratic-loss problem underlying mean–variance optimization has a linear feedback solution when the investment opportunity set is deterministic (Zhou and Li, 2000). Relaxed controls replace actions by probability distributions and average their infinitesimal generators. Entropy regularization then produces Gaussian policies for linear–quadratic problems (Wang et al., 2020). In continuous-time mean–variance portfolio selection, a constant entropy temperature yields an exploratory policy whose mean agrees with the classical optimal feedback (Wang and Zhou, 2020). These results motivate a more precise design problem: which changes in exploration dispersion preserve that agreement?

We study the unconstrained action space $\mathbb{R}^d$, allowing both long and short risky positions. The target is a fixed-center terminal quadratic-loss problem; the connection to mean–variance calibration is given separately in Appendix C. We retain the discounted portfolio dynamics and Gaussian policy-learning structure of the constrained formulation of Li et al. (2026), but dispense with cone constraints, nonnegative means, piecewise feedbacks, and projected optimization. This exposes a common mechanism behind two exploration formulations.

Contributions. We first characterize mean preservation through the ratio V_x/V_{xx} and show that, within a smooth strictly convex class, it requires a quadratic value function centered at the target. For ordinary differential-entropy regularization, we recover the state-independent-temperature solution and derive a compatibility condition for state-dependent temperatures, together with an explicit con-

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.

struction. We also explain why a temperature proportional to the squared wealth gap generally fails. For direct Gaussian exploration, we prove mean preservation whenever the covariance contributes a constant-plus-quadratic term to the wealth variance; this includes additive, multiplicative, and mixed exploration. Finally, we specialize the trajectory-based actor–critic approach to unconstrained policies and give a one-asset analytical example. No numerical convergence or empirical-performance claim is made.

Related work and scope. The Gaussian optimizer of an entropy-regularized quadratic Hamiltonian and the constant-temperature portfolio solution are established results (Wang et al., 2020; Wang and Zhou, 2020); we use them as starting points. Direct policy randomization can also recover optimal deterministic controls in Merton's problem (Dai et al., 2026), while Li et al. (2026) study a Gaussian auxiliary formulation with nonnegative means. Martingale policy evaluation and continuous-time policy gradients provide the learning tools used here (Jia and Zhou, 2022a;b). Recent work develops portfolio learning algorithms with regret analysis and empirical evaluation (Huang et al., 2024). Our focus is the structural condition for exact mean recovery under state-dependent dispersion, rather than a new general-purpose policy-gradient method or a claim that Gaussian exploration is new.

2 Unconstrained target problem and exploratory dynamics

Consider d risky assets with appreciation vector b_t, volatility matrix $\sigma_t \in \mathbb{R}^{d\times d}$, and a riskless rate r_t. Set

$$\beta_t = b_t - r_t\mathbf{1}, \qquad Q_t = \sigma_t\sigma_t^{\mathsf{T}}, \qquad k_t = Q_t^{-1}\beta_t, \qquad q_t = \beta_t^{\mathsf{T}}Q_t^{-1}\beta_t. \tag{1}$$

For discounted risky dollar holdings $u_t \in \mathbb{R}^d$, discounted wealth satisfies

$$\mathrm{d}X_t = \beta_t^{\mathsf{T}}u_t\,\mathrm{d}t + u_t^{\mathsf{T}}\sigma_t\,\mathrm{d}W_t. \tag{2}$$

There are no sign, leverage, or wealth constraints. Admissibility requires progressive measurability and $\mathbb{E}\int_0^T \|u_t\|^2\mathrm{d}t < \infty$.

Assumption 2.1. *The coefficients are deterministic, bounded, and continuous on $[0,T]$, and $Q_t \succeq \delta I_d$ for some $\delta > 0$. The center $w \in \mathbb{R}$ is fixed. Exploratory policies have finite integrated second moments. In the entropy formulation, their running entropy cost is absolutely integrable in expectation. Localized Itô identities below are used for policies with integrable state suprema of order two; the displayed optimal feedbacks satisfy these conditions.*

Continuity is imposed to state classical HJB identities. The learning procedure does not require access to these coefficient functions. For $X_t = x$, define

$$V^0(t,x;w) = \inf_u \mathbb{E}_{t,x}(X_T - w)^2. \tag{3}$$

Write $y = x - w$ and suppress w when unambiguous. Completing the square in the HJB equation gives

$$V^0(t,x) = P_0(t)y^2, \qquad P_0(t) = \exp\left(-\int_t^T q_s\mathrm{d}s\right), \qquad u^*(t,x) = -k_t y. \tag{4}$$

Indeed, $V_t^0 - q_t(V_x^0)^2/(2V_{xx}^0) = 0$, with $V^0(T,x) = y^2$.

A randomized feedback $\pi(\mathrm{d}u \mid t,x)$ has mean $m(t,x)$ and covariance $C(t,x)$. Its relaxed wealth generator is

$$\mathcal{L}^\pi f = \beta_t^{\mathsf{T}}m f_x + \frac{1}{2}\left[m^{\mathsf{T}}Q_t m + \mathrm{tr}(Q_t C)\right]f_{xx}. \tag{5}$$

Thus an equivalent scalar diffusion has drift $\beta_t^{\mathsf{T}}m$ and diffusion coefficient $\sqrt{m^{\mathsf{T}}Q_t m + \mathrm{tr}(Q_t C)}$. Crucially, the second moment of the action enters the diffusion; substituting the mean into (2) would omit exploration risk. Sample-and-hold randomized actions approximate this generator as the decision interval vanishes. They are not an exact finite-step implementation of the relaxed diffusion.

3 THE MECHANISM OF MEAN PRESERVATION

For a smooth candidate value V with $V_{xx} > 0$, the action-dependent Hamiltonian is

$$H_V(u) = V_x \beta_t^\top u + \frac{1}{2} V_{xx} u^\top Q_t u. \tag{6}$$

Both entropy regularization over all action densities and optimization over Gaussian means with a fixed covariance yield the same mean minimizer:

$$m_V(t,x) = -Q_t^{-1} \beta_t \frac{V_x(t,x)}{V_{xx}(t,x)}. \tag{7}$$

The two formulations generally have different value functions. Their mean policies agree with (4) precisely when their derivative ratios agree with y (at times with $\beta_t \neq 0$).

Proposition 3.1 (Quadratic closure)**.** *At a fixed time with $\beta_t \neq 0$, let $V(t,\cdot) \in C^2(\mathbb{R})$ and $V_{xx} > 0$. The mean in (7) equals $-k_t(x-w)$ for every x if and only if*

$$V(t,x) = P(t)(x-w)^2 + B(t), \qquad P(t) > 0. \tag{8}$$

Proof. Equality of the means implies $V_x = yV_{xx}$. On each half-line, the differential equation for $g(y) = V_x(t, w+y)$ gives $g(y) = c_\pm y$. Continuity of V_{xx} at zero forces $c_+ = c_-$; continuity of V matches the additive constants. Strict convexity gives $P = c_\pm/2 > 0$. The converse is immediate. □

This observation distinguishes Hamiltonian minimization from dynamic optimization. Holding V fixed, the covariance does not appear in (7). Nevertheless, changing exploration changes V through the HJB equation. Exact recovery requires preservation of quadratic closure, not merely a Gaussian policy or zero-mean perturbations. If $\beta_t = 0$, the optimal mean at that time is zero regardless of the derivative ratio, so the necessity assertion is not applicable there.

4 ENTROPY-REGULARIZED EXPLORATION

Fix an exogenous positive temperature $\lambda(t,x)$, independent of the policy being optimized, and consider

$$V^\lambda(t,x) = \inf_\pi \mathbb{E}_{t,x}\left[(X_T^\pi - w)^2 + \int_t^T \lambda(s, X_s^\pi) \int_{\mathbb{R}^d} \pi_s(u) \log \pi_s(u) \mathrm{d}u \, \mathrm{d}s\right]. \tag{9}$$

Here π_s is a Lebesgue density, and actions are expressed in fixed discounted-dollar units. This is ordinary differential entropy; no state-dependent reference measure or entropy normalization is implicit. The Gibbs variational identity yields

$$\pi_V^\lambda(\cdot \mid t,x) = \mathcal{N}\left(-k_t \frac{V_x}{V_{xx}}, \frac{\lambda(t,x)}{V_{xx}} Q_t^{-1}\right), \tag{10}$$

$$0 = V_t - \frac{q_t V_x^2}{2V_{xx}} - \frac{\lambda(t,x)}{2} \log \frac{(2\pi\lambda(t,x)/V_{xx})^d}{\det Q_t}. \tag{11}$$

Appendix A supplies a verification identity rather than relying on formal HJB substitution alone.

4.1 STATE-INDEPENDENT TEMPERATURE

Theorem 4.1. *Let $\lambda(t,x) = \lambda_t$ be positive and continuous. Then*

$$V^\lambda(t,x) = P_0(t) y^2 + B_\lambda(t), \tag{12}$$

$$B_\lambda(t) = -\frac{1}{2} \int_t^T \lambda_s \log \frac{(\pi\lambda_s / P_0(s))^d}{\det Q_s} \, \mathrm{d}s, \tag{13}$$

$$\pi^{\lambda,*}(\cdot \mid t,x) = \mathcal{N}\left(-k_t y, \frac{\lambda_t}{2P_0(t)} Q_t^{-1}\right). \tag{14}$$

In particular, the exploratory mean equals u^* for every positive temperature schedule.

Proof. Substitute (12) into (11). The quadratic coefficient gives $P_0' = qP_0$, and the constant coefficient gives $B_\lambda' = \frac{\lambda_t}{2}\log[(\pi\lambda_t/P_0)^d/\det Q_t]$ with terminal value zero. The Gaussian formula follows from (10); verification is given in Appendix A. □

The temperature changes the covariance and the regularized value, but does not change the coefficient P_0. This is the familiar mean–variance separation in entropy-regularized portfolio control, extended here to a deterministic temperature schedule.

4.2 State-dependent temperature: a compatibility condition

Define the minimized entropy contribution for a quadratic value by

$$F_\lambda(t,y;P) = -\frac{\lambda(t,w+y)}{2}\log\frac{(\pi\lambda(t,w+y)/P(t))^d}{\det Q_t}. \tag{15}$$

Theorem 4.2 (Compatible state dependence)**.** *Suppose a positive C^1 function P, with $P(T)=1$, and continuous functions f_0, f_2 satisfy*

$$F_\lambda(t,y;P) = f_2(t)y^2 + f_0(t), \qquad P' = qP - f_2. \tag{16}$$

Under the integrability conditions of Assumption 2.1, and provided the feedback below is admissible, the solution is

$$V^\lambda(t,x) = P(t)y^2 + \int_t^T f_0(s)\mathrm{d}s, \qquad \pi^{\lambda,*} = \mathcal{N}\left(-k_t y, \frac{\lambda(t,x)}{2P(t)}Q_t^{-1}\right). \tag{17}$$

Conversely, a $C^{1,2}$ strictly convex value with exact mean preservation must satisfy (16) on any time interval where $\beta_t \neq 0$.

Proof. Quadratic substitution in (11) gives $(P' - qP)y^2 + B' + F_\lambda = 0$. Condition (16) sets both coefficients to zero. For necessity, apply Proposition 3.1 and the same HJB identity. Verification follows from the Gibbs gap identity in Appendix A. □

The condition is constructive. Let $D_t = (\det Q_t)^{1/d}$, choose a bounded continuous $\eta_t \geq 0$, and let W_0 be the nonnegative branch of the Lambert function, defined by $W_0(z)e^{W_0(z)} = z$ for $z \geq 0$. Set

$$P_\eta(t) = \exp\left[-\int_t^T (q_s + \eta_s)\mathrm{d}s\right], \tag{18}$$

$$\lambda_\eta(t,x) = \frac{P_\eta(t)D_t}{\pi}\exp\left[W_0\left(\frac{2\pi\eta_t(x-w)^2}{dD_t}\right)\right]. \tag{19}$$

Then $F_{\lambda_\eta} = -\eta_t P_\eta(t)y^2$, so Theorem 4.2 applies with $f_2 = -\eta P_\eta$ and $f_0 = 0$. This temperature is strictly positive even at $x = w$ and depends on the state whenever $\eta_t > 0$. Its optimal covariance is

$$C_\eta(t,x) = \frac{D_t}{2\pi}\exp\left[W_0\left(\frac{2\pi\eta_t y^2}{dD_t}\right)\right]Q_t^{-1}. \tag{20}$$

The identity $e^{W_0(z)} \leq 1 + z$ ensures at most quadratic covariance growth. The Gaussian feedback has finite moments and integrable entropy cost on a finite horizon, as detailed in Appendix A. This construction establishes existence; it depends on market coefficients and is not claimed to be a coefficient-free temperature-selection algorithm.

Why an arbitrary state-dependent temperature fails. For $\lambda(t,x) = \theta_t y^2$ with $\theta_t > 0$ and $y \neq 0$, (15) becomes

$$F_\lambda = -\frac{\theta_t y^2}{2}\left[d\log\frac{\pi\theta_t}{P(t)} - \log\det Q_t + d\log y^2\right]. \tag{21}$$

The $y^2\log y^2$ term cannot be absorbed into a quadratic value. Thus this natural temperature does not preserve the optimal mean in the smooth class of Theorem 4.2. A state-dependent normalization of entropy can cancel this term, but changes the objective. In particular, the direct Gaussian result below cannot be transferred to ordinary entropy regularization by simply identifying temperature with covariance.

5 DIRECT GAUSSIAN EXPLORATION

Instead of optimizing entropy, prescribe a covariance $C(t,x)$ and optimize only the Gaussian mean:

$$V^C(t,x) = \inf_m \mathbb{E}_{t,x}(X_T^{m,C} - w)^2, \qquad \pi^{m,C}(\cdot \mid t,x) = \mathcal{N}(m(t,x), C(t,x)). \tag{22}$$

Positive semidefinite covariances, including Dirac laws, are allowed in this formulation. There is no entropy term. Its HJB equation is

$$0 = V_t - \frac{q_t V_x^2}{2V_{xx}} + \frac{1}{2} V_{xx} \operatorname{tr}(Q_t C(t,x)), \qquad V(T,x) = y^2. \tag{23}$$

Theorem 5.1 (Additive and multiplicative exploration)**.** *Let $C_0(t), C_2(t)$ be bounded continuous positive semidefinite matrices and set*

$$C(t,x) = C_0(t) + (x-w)^2 C_2(t), \qquad s_j(t) = \operatorname{tr}(Q_t C_j(t)), \quad j = 0, 2. \tag{24}$$

Then the optimal exploratory policy and value are

$$\pi^{C,*}(\cdot \mid t,x) = \mathcal{N}(-k_t(x-w), C(t,x)), \tag{25}$$

$$V^C(t,x) = P_C(t)(x-w)^2 + B_C(t), \tag{26}$$

$$P_C(t) = \exp\left[\int_t^T (s_2(s) - q_s)\mathrm{d}s\right], \qquad B_C(t) = \int_t^T P_C(s) s_0(s)\mathrm{d}s. \tag{27}$$

Consequently, $\mathbb{E}_{\pi^{C,*}}[u \mid t,x] = u^*(t,x)$ for every exploration level covered by (24).

Proof. Substitution in (23) yields $P_C' + (s_2 - q)P_C = 0$ and $B_C' + s_0 P_C = 0$, with $P_C(T) = 1$ and $B_C(T) = 0$. Completing the square in the mean gives (25). More explicitly, for any admissible mean m,

$$\mathbb{E}_{t,x}(X_T^{m,C} - w)^2 = V^C(t,x) + \mathbb{E}_{t,x} \int_t^T P_C(s) \|m(s, X_s) + k_s(X_s - w)\|_{Q_s}^2 \mathrm{d}s, \tag{28}$$

where $\|v\|_Q^2 = v^\top Q v$. Itô's formula and localization justify the identity. The additional term is nonnegative and vanishes under the displayed optimal mean. □

Interpretation. State-independent covariance ($C_2 = 0$) adds a constant to the value. Purely multiplicative covariance ($C_0 = 0$) rescales the quadratic coefficient. Their mixture does both, while the derivative ratio remains $x - w$. For pure multiplicative exploration, both mean and covariance vanish at $x = w$; the Gaussian then denotes the Dirac law at zero, not a positive-density distribution. Adding $C_0 \succ 0$ avoids this degeneracy and is convenient for score-based learning.

Only $\operatorname{tr}(Q_t C)$ enters the generator. Thus the same conclusion holds for other covariance fields with the same constant-plus-quadratic trace, subject to well-posedness and integrability. Conversely, when $\beta_t \neq 0$, quadratic closure in (23) requires that trace to have this form. Positive semidefiniteness for every x forces its constant and quadratic coefficients to be nonnegative.

Mean recovery does not remove exploration risk. The unregularized exploration cost is explicit:

$$V^C(t,x) - V^0(t,x) = \big[P_C(t) - P_0(t)\big](x-w)^2 + B_C(t) \geq 0. \tag{29}$$

After training, deploying the mean feedback under (2) recovers V^0, whereas continuing to randomize incurs (29). Equality of feedback maps does not mean equality of realized actions along different state trajectories. Moreover, entropy-regularized values include a running reward and cannot be compared directly with unregularized terminal losses.

6 LEARNING THE UNCONSTRAINED POLICY FROM TRAJECTORIES

Theorem 5.1 suggests a model-free implementation with designer-chosen matrices $C_0 = \varepsilon I_d$ and $C_2 = \psi I_d$, where $\varepsilon > 0$ and $\psi \geq 0$. No market covariance is needed to specify these matrices. Use

$$m^\phi(t,x) = -\phi(t)(x-w), \quad \pi^\phi = \mathcal{N}(m^\phi, C), \quad V^\theta(t,x) = e^{a_\theta(t)}(x-w)^2 + b_\theta(t), \tag{30}$$

Algorithm 1 Unconstrained actor–critic with mean-preserving Gaussian exploration

Input: Center w, grid $\{t_i\}$, batch size M, actor ϕ, critic θ, $C(t,x) = \varepsilon I_d + \psi(x-w)^2 I_d$, and learning rates.

1. For each training iteration, hold ϕ fixed and collect M on-policy trajectories with sampled actions.
2. Form terminal targets $G^{(j)} = (X_N^{(j)} - w)^2$.
3. Update θ using (31), enforcing the terminal critic condition.
4. Collect a fresh on-policy batch with the same actor; freeze the fitted critic.
5. Compute scores (32) and temporal differences δ_i^θ.
6. Update every actor vector by (33), without projection; repeat.

Output: Deterministic feedback $\widehat{u}(t,x) = -\phi(t)(x-w)$.

with $a_\theta(T) = b_\theta(T) = 0$. The actor $\phi(t) \in \mathbb{R}^d$ is unrestricted. A time grid or a differentiable time network can represent these functions. For deterministic ϕ, the exact relaxed-policy value lies in this critic class; its coefficients are given in Appendix B.

Let $0 = t_0 < \cdots < t_N = T$, $\Delta_i = t_{i+1} - t_i$. Collect on-policy trajectories (X_i, U_i, X_{i+1}) by sampling $U_i \sim \pi^\phi(\cdot \mid t_i, X_i)$ and holding the action over the next interval. With $G = (X_N - w)^2$, fit the critic by the martingale-inspired regression loss

$$\widehat{L}(\theta) = \frac{1}{M} \sum_{j=1}^{M} \sum_{i=0}^{N-1} \Delta_i \big[V^\theta(t_i, X_i^{(j)}) - G^{(j)} \big]^2. \tag{31}$$

At each state, the population squared-loss minimizer is the conditional expected terminal cost. This is the terminal-reward specialization of martingale policy evaluation (Jia and Zhou, 2022a). An additive critic term is required whenever $C_0 \neq 0$.

For separate actor vectors ϕ_i at grid points, the Gaussian score is

$$S_i = \nabla_{\phi_i} \log \pi^\phi(U_i \mid t_i, X_i) = -(X_i - w) C(t_i, X_i)^{-1} \big[U_i + \phi_i (X_i - w) \big]. \tag{32}$$

Set $\delta_i^\theta = V^\theta(t_{i+1}, X_{i+1}) - V^\theta(t_i, X_i)$. The actor estimate is

$$\widehat{g}_{\phi_i} = \frac{1}{M} \sum_{j=1}^{M} S_i^{(j)} \delta_i^{\theta,(j)}, \qquad \phi_i \leftarrow \phi_i - \alpha_\phi \widehat{g}_{\phi_i}. \tag{33}$$

For shared actor parameters, apply the chain rule and sum over i. No additional factor Δ_i multiplies this temporal-difference estimator. With an exact value function for the sampled finite-step problem, it is an unbiased likelihood-ratio gradient for that problem (Appendix B). The continuous-time interpretation is consistent with the policy-gradient approach of Jia and Zhou (2022b). A learned critic introduces approximation error; a finite decision interval also changes the optimal control problem.

Algorithm 1 requires observed states, sampled actions, and terminal losses, but no estimates of β or Q. It is a learning procedure motivated by the exact control result; that result does not establish convergence of noisy actor–critic iterates. Fresh batches keep the fitted critic independent of the actor-estimation batch conditional on the actor. If an unbiased finite-grid gradient is preferred to bootstrapping, replace δ_i^θ by $G - V^\theta(t_i, X_i)$; this generally has higher variance but does not require an exact critic. The covariance is held fixed during each actor update. Optimizing it jointly without entropy or a prescribed exploration level would favor vanishing exploration.

7 A SIMPLE ANALYTICAL EXAMPLE

Consider one risky asset with constant excess return $\beta = 0.06$, volatility $\sigma = 0.20$, horizon $T = 1$, initial wealth $x_0 = 1$, and fixed center $w = 1.2$. Then $Q = 0.04$, $k = 1.5$, $q = 0.09$, and

$$u^*(t,x) = -1.5(x - 1.2), \qquad V^0(t,x) = e^{-0.09(1-t)}(x-1.2)^2. \tag{34}$$

The feedback is positive below 1.2 and negative above 1.2, with no truncation. For direct covariance $C(t,x) = c_0 + c_2(x-1.2)^2$, set $a = 0.04c_2 - 0.09$. The exact exploratory value is

$$V^C(t,x) = e^{a(1-t)}(x-1.2)^2 + 0.04c_0\frac{e^{a(1-t)}-1}{a}, \tag{35}$$

where the fraction is $1-t$ when $a=0$. Table 1 evaluates this formula directly; its entries are not simulation results. For a constant entropy temperature $\lambda > 0$, the optimal covariance is $C^\lambda(t) =$

Table 1: Exact terminal quadratic losses at $(t,x) = (0,1)$. Every extracted mean has initial action 0.3 and deployed loss 0.03655725.

Exploration	c_0	c_2	Exploratory loss $V^C(0,1)$
None	0	0	0.03655725
State-independent	0.04	0	0.03808736
State-dependent	0	0.25	0.03692465

$\lambda e^{0.09(1-t)}/0.08$, while the mean remains (34). Its unregularized terminal loss exceeds $V^0(t,x)$ by $\lambda(1-t)/2$, obtained by integrating $P_0(s)QC^\lambda(s) = \lambda/2$. This is distinct from its entropy-regularized value in Theorem 4.1.

For an explicit state-dependent temperature, take $\eta = 0.02$ in (19). Then

$$\lambda_\eta(t,x) = \frac{0.04}{\pi}e^{-0.11(1-t)}\exp\left[W_0(\pi(x-1.2)^2)\right], \tag{36}$$

$$V^{\lambda_\eta}(t,x) = e^{-0.11(1-t)}(x-1.2)^2, \qquad C_\eta(t,x) = \frac{1}{2\pi}\exp\left[W_0(\pi(x-1.2)^2)\right]. \tag{37}$$

Again the optimal mean is (34). The smaller regularized value reflects the entropy reward, not a reduction of unregularized portfolio risk. These closed-form cases illustrate the distinction between preserving exploitation and changing the training objective or state dispersion.

8 Discussion and Conclusion

Mean-preserving exploration is a property of the optimized control problem. In the deterministic-coefficient portfolio model, its mechanism is preservation of the ratio $V_x/V_{xx} = x - w$. Direct Gaussian covariances with constant-plus-quadratic contributions to wealth variance preserve this ratio automatically. Ordinary entropy regularization preserves it for state-independent temperatures and for compatible state-dependent designs, but not for arbitrary state dependence.

The results are restricted to scalar discounted wealth, deterministic market coefficients, and terminal quadratic loss. Random investment opportunities introduce additional state variables and hedging terms; nonlinear objectives need not admit quadratic closure. The state-dependent entropy construction uses model coefficients, whereas the direct-covariance algorithm does not. No finite-sample complexity, actor–critic convergence theorem, or numerical experiment is claimed. The analytical example establishes exact identities and provides a benchmark for future tests of estimation error and time discretization. Within the stated setting, the results identify precisely how exploration dispersion can change while the optimal deployed feedback remains fixed.

REPRODUCIBILITY STATEMENT

All control identities are derived in the main text and Appendices A–C. The one-asset example specifies every parameter and uses closed-form formulas. Algorithm 1 states the sampling, evaluation, and actor updates; it has not been empirically evaluated in this draft.

AI USE STATEMENT

An OpenAI language model was used to assist in restructuring a source manuscript, drafting text and mathematical derivations, preparing LaTeX, and checking analytical formulas and document formatting. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Min Dai, Yuchao Dong, Yanwei Jia, and Xun Yu Zhou. Data-driven Merton's strategies via policy randomization. *arXiv preprint arXiv:2312.11797*, version 3, 2026.

Yilie Huang, Yanwei Jia, and Xun Yu Zhou. Mean–variance portfolio selection by continuous-time reinforcement learning: Algorithms, regret analysis, and empirical study. *arXiv preprint arXiv:2412.16175*, 2024.

Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55, 2022a.

Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275):1–50, 2022b.

Xun Li, Yutian Wang, and Xun Yu Zhou. Dynamic mean–variance portfolio selection with no-shorting constraints and unknown investment opportunity sets. *arXiv preprint arXiv:2607.16625*, 2026.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.

Haoran Wang and Xun Yu Zhou. Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30:1273–1308, 2020.

Xun Yu Zhou and Duan Li. Continuous-time mean–variance portfolio selection: A stochastic LQ framework. *Applied Mathematics & Optimization*, 42:19–33, 2000.