Connecting Martingale Optimal Transport, Reinforcement Learning, and Dynamic Hedging

Abstract

This document explores the connection between Martingale Optimal Transport (MOT) problems, reinforcement learning frameworks, and dynamic hedging strategies. Specifically, we analyze how the actor-critic reinforcement learning framework provides a natural solution method for dynamic MOT problems and relate this to the Schrödinger Bridge Problem (SBP) and entropically regularized optimal transport.

1 Martingale Optimal Transport (MOT)

MOT involves finding the coupling $\pi(x, y)$ that minimizes a cost function $c(x, y)$ under martingale constraints:

$$\inf_{\pi \in \Pi(\mu_0, \mu_T)} \int c(x, y)\, d\pi(x, y), \tag{1}$$

where $\Pi(\mu_0, \mu_T)$ is the set of couplings between μ_0 and μ_T such that the martingale property holds:

$$\mathbb{E}[Y \mid X] = X. \tag{2}$$

The dynamic formulation of MOT introduces time evolution via measures ρ_t and velocity fields v_t constrained by the continuity equation:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0. \tag{3}$$

To make this problem computationally tractable, we use entropic regularization, which modifies the objective:

$$\inf_{\pi \in \Pi(\mu_0, \mu_T)} \int c(x, y)\, d\pi(x, y) + \frac{1}{\epsilon} H(\pi \| \pi_0), \tag{4}$$

where $H(\pi \| \pi_0)$ is the Kullback-Leibler (KL) divergence relative to a prior measure π_0.

2 Reinforcement Learning Framework

Dynamic hedging can be framed as a Markov Decision Process (MDP) with the following components:

- **States (s_t)**: Represent market conditions and portfolio states.

- **Actions (a_t)**: Trading decisions to hedge risk.
- **Rewards (r_t)**: Cashflows and penalties based on transaction costs and risk aversion.

The value function $V(s_t)$, which captures the risk-adjusted expected return, satisfies the Bellman equation:

$$V(s_t) = \sup_{a_t} \mathbb{E}\left[r_t + \gamma V(s_{t+1}) \mid s_t, a_t\right], \tag{5}$$

where γ is a discount factor.

For risk-averse reinforcement learning, we use a risk-sensitive value function:

$$V(s_t) = -\frac{1}{\lambda} \log \mathbb{E}\left[\exp\left(-\lambda \sum_{i=t}^{T} r(s_i, a_i)\right)\right], \tag{6}$$

where λ controls the degree of risk aversion.

3 Actor-Critic Framework for MOT

The actor-critic method provides a solution framework for MOT problems by splitting optimization into two parts:

- **Actor**: Parameterizes the policy $\pi_\theta(a \mid s)$ and updates it to maximize the value function.
- **Critic**: Estimates the value function $V(s)$ or the action-value function $Q(s, a)$ to guide the Actor.

The policy gradient is computed as:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim\pi, a\sim\pi}\left[\nabla_\theta \log \pi_\theta(a \mid s) Q(s, a)\right]. \tag{7}$$

For MOT, the actor parameterizes the control variables h_t and q, while the critic estimates the cost-to-go function, providing feedback for policy improvement.

4 Connection to Schrödinger Bridge Problem

The Schrödinger Bridge Problem (SBP) seeks the most likely stochastic process Q connecting two distributions μ_0 and μ_T, minimizing the relative entropy with respect to a reference process R:

$$\inf_{Q\in\mathcal{P}(\mu_0, \mu_T)} H(Q\|R). \tag{8}$$

The entropic regularization in OT is equivalent to solving SBP dynamically, where the reference measure R corresponds to the prior π_0 in OT.

Using the Pythagorean theorem for relative entropy, the solution can be decomposed as:

$$H(P\|R) = H(P\|Q^*) + H(Q^*\|R), \tag{9}$$

where Q^* is the intermediate calibration measure satisfying the marginal constraints.

5 Unified Framework

By introducing a calibration measure Q^* (or π^*), we unify MOT, SBP, and reinforcement learning under the same inequality:

$$H(\text{Solution}\|\text{Reference}) = H(\text{Solution}\|\text{Calibration}) + H(\text{Calibration}\|\text{Reference}). \quad (10)$$

This decomposition highlights the dual role of regularization: smoothing the solution and introducing a bias toward the reference measure.

6 Conclusion

The connection between MOT, reinforcement learning, and SBP provides a powerful framework for solving dynamic problems in finance and beyond. The actor-critic method offers a computationally efficient way to tackle MOT problems, while the equivalence to SBP unifies the treatment of entropic regularization across dynamic and static settings.