Recent Advances in RL in finance

Optimal Execution

- The Almgren-Chriss Model
    - Setup
        - Trader sells an amount q_0 of an asset with price S_0
        - Time period [0,T] with trading decision at discrete time points
        - Final inventory $q_T = 0$
    - Goal
        - Determine the liquidation strategy $u_1, u_2, \ldots, u_T$ (u_t : the amount of the asset to sell at t)
    - Price impact
        - Temporary impact : 매도에 따른 수요 공급의 일시적인 불일치
        - Permanent impact: 장기적으로 평형가격에 영향을 미침

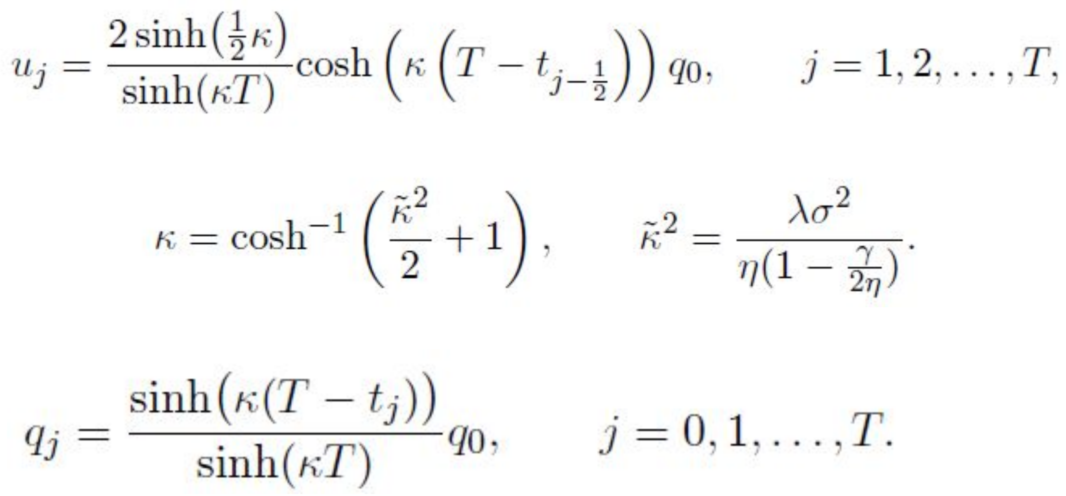
Optimal Execution

- The Almgren-Chriss Model
  - Discrete arithmetic random walk $S_t = S_{t-1} + \sigma\xi_t - g(u_t), \qquad t = 1, 2, \ldots, T$
  - Inventory process $q_t = q_{t-1} - u_t.$
  - Price per share $\widetilde{S}_t = S_{t-1} - h(u_t)$
  - Cost of the trading trajectory $C = q_0 S_0 - \sum_{t=0}^{T} q_t \widetilde{S}_t \qquad \mathbb{E}[C] = \sum_{t=1}^{T} \left(q_t g(u_t) + u_t h(u_t)\right), \qquad \mathrm{Var}(C) = \sigma^2 \sum_{t=1}^{T} q_t^2$
  - Minimize expected cost and variance of the cost

$$\min_{\{u_t\}_{t=1}^{T}} \left(\mathbb{E}[C] + \lambda \mathrm{Var}(C)\right)$$

Optimal Execution

- The Almgren-Chriss Model
  - General solution



$$u_j = \frac{2\sinh(\frac{1}{2}\kappa)}{\sinh(\kappa T)}\cosh\left(\kappa\left(T - t_{j-\frac{1}{2}}\right)\right) q_0, \qquad j = 1, 2, \ldots, T,$$

$$\kappa = \cosh^{-1}\left(\frac{\tilde{\kappa}^2}{2} + 1\right), \qquad \tilde{\kappa}^2 = \frac{\lambda\sigma^2}{\eta(1 - \frac{\gamma}{2\eta})}.$$

$$q_j = \frac{\sinh\big(\kappa(T - t_j)\big)}{\sinh(\kappa T)} q_0, \qquad j = 0, 1, \ldots, T.$$

Optimal Execution

- Evaluation Criteria
    - the Profit and Loss (PnL), Implementation Shortfall, and the Sharpe ratio
        - The PnL : the final profit or loss induced by a given execution algorithm
        - The Implementation Shortfall : the difference between the PnL of the algorithm and the PnL received by trading the entire amount of the asset instantly
        - The Sharpe ratio : the ratio of expected return to standard deviation of the return
            - the differential Sharpe ratio
            - the Sortino ratio
    - Time-Weighted Average Price (TWAP)
    - Volume-Weighted Average Price (VWAP)
    - Submit and Leave (SnL) policy
        - where a trader places a sell order for all shares at a fixed limit order price, and goes to the market with any unexecuted shares remaining at time T.
    -

RL Approach : Optimal Execution

- a brief overview of the existing literature on RL for optimal execution.
  - The most popular types of RL methods
    - Q-learning algorithms and (double) DQN
    - Policy-based algorithms
    - (deep) policy gradient methods, A2C, PPO, DDPG
- The state variables
  - time stamp
  - the market attributes including (mid-)price of the asset and/or the spread
  - the inventory process
  - Past returns
- The control variables
  - the amount of asset (using market orders) to trade and/or the relative price level (using limit orders) at each time point.
- Examples of reward functions : cash inflow or outflow
  - implementation shortfall, profit, Sharpe ratio, return, and PnL

RL Approach : Optimal Execution

- Popular choices of performance measure
    - Implementation Shortfall, PnL, trading cost, profit, Sharpe ratio, Sortino ratio, return
- value-based algorithms
    - provided the first large scale empirical analysis of a RL method applied to optimal execution problems.
    - focused on a modified Q-learning algorithm to select price levels for limit order trading, which leads to significant improvements over simpler forms of optimization such as the SnL policies in terms of trading costs.
- policy-based algorithms : combined deep learning with RL to determine whether to sell, hold, or buy at each time point
    - In the first step of their model: neural networks are used to summarize the market features
    - in the second step the RL part makes trading decisions

Portfolio Optimization

- Markowitz Model : mean-variance model
    - Investor seeks a portfolio to maximize the expected total return for any given level of risk measured by variance
    - Time-inconsistency problem : optimal strategy is no longer optimal at time s>t
    - Setting
        - n risky assets with initial wealth x_0
    - Goal
        - Find an optimal strategy such that the portfolio return is maximized while minimizing the risk of the investment

$$\max_{\{\boldsymbol{u}_t\}_{t=0}^{T-1}} \mathbb{E}[x_T] - \phi \mathrm{Var}(x_T),$$

$$x_{t+1} = \sum_{i=1}^{n-1} e_t^i u_t^i + \left(x_t - \sum_{i=1}^{n-1} u_t^i \right) e_t^n, \qquad t = 0, 1, \ldots, T-1,$$

RL Approach : Portfolio Optimization

- Methods
  - value-based methods: Q-learning, SARSA, DQN
  - policy-based algorithms : DPG and DDPG
- 요소:
  - The state variables : time, asset prices, asset past returns, current holdings of assets, and remaining balance.
  - The control variables : the amount/proportion of wealth invested in each component of the portfolio.
  - Examples of reward signals : portfolio return, (differential) Sharpe ratio, profit
  - The benchmark strategies : Constantly Rebalanced Portfolio (CRP)
  - The performance measures : the Sharpe ratio, the Sortino ratio, portfolio returns, portfolio values, cumulative profits
  - Some models incorporate the transaction costs and investments in the risk-free asset

RL Approach : Portfolio Optimization

- For value-based algorithms
  - considered the portfolio optimization problems of a risky asset and a risk-free asset
  - compared the performance of the Q-learning algorithm and a Recurrent RL (RRL) algorithm
    - Under three value functions : the Sharpe ratio, differential Sharpe ratio, and profit
  - The RRL algorithm : a policy-based method which uses the last action as an input.
  - concluded
    - Q-learning algorithm is more sensitive to the choice of value function and has less stable performance than the RRL algorithm
    - suggested that the (differential) Sharpe ratio is preferred rather than the profit as the reward function.

RL Approach : Portfolio Optimization

- For policy-based algorithms
  - [110] proposed a framework combining neural networks with DPG.
    - a so-called Ensemble of Identical Independent Evaluators (EIIE) topology
      - to predict the potential growth of the assets in the immediate future using historical data which includes the highest, lowest, and closing prices of portfolio components.
    - The experiments using real cryptocurrency market data showed that their framework achieves higher Sharpe ratio and cumulative portfolio values compared with three benchmarks including CRP and several published RL models
  - [224] explored the DDPG algorithm for the portfolio selection of 30 stocks
    - at each time point, the agent can choose to buy, sell, or hold each stock
    - The DDPG algorithm was shown to outperform two classical strategies including the min-variance portfolio allocation method [231] in terms of several performance measures including final portfolio values, annualized return, and Sharpe ratio, using historical daily prices of the 30 stocks

RL Approach : Portfolio Optimization

- For policy-based algorithms
  - [137] considered the DDPG, PPO, and policy gradient method with an adversarial learning scheme which learns the execution strategy using noisy market data.
  - [236] proposed a model using DDPG, which includes prediction of the assets future price movements based on historical prices and synthetic market data generation using a Generative Adversarial Network (GAN) [84]
- Using Actor-Critic methods
  - [3] combined the mean-variance framework (the actor determines the policy using the mean-variance framework) and the Kelly Criterion framework (the critic evaluates the policy using their growth rate).
  - studied eight policy-based algorithms including DPG, DDPG, and PPO, among which DPG was shown to achieve the best performance.

Option pricing and Hedging

- Black-Scholes Model
  - Geometric Brownian motion $dS_t = \mu S_t dt + \sigma S_t dW_t,$

$$\frac{\partial V}{\partial t}(S_t, t) + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 V}{\partial S^2}(S_t, t) + r S_t \frac{\partial V}{\partial S}(S_t, t) - rV(S_t, t) = 0,$$

  - Call option $P(S_T) = \max(S_T - K, 0)$

$$V(S_t, t) = N(d_1)S_t - N(d_2)Ke^{-r(T-t)}$$

$$d_1 = \frac{1}{\sigma\sqrt{T-t}}\left(\ln\left(\frac{S_t}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)\right), \qquad d_2 = d_1 - \sigma\sqrt{T-t}.$$

Option pricing and Hedging

- hedge a given derivative contract by buying and selling the underlying asset
  - to eliminate risk
  - Black-Scholes analysis delta hedging $\Delta_t := \frac{\partial V}{\partial S}(S_t, t)$
  - to use financial derivatives to hedge against the volatility of given positions in the underlying assets
  - can only rebalance portfolios at discrete time points and frequent transactions may incur high costs.
  - an optimal hedging strategy depends on the tradeoff between the hedging error and the transaction costs.
    - in a similar spirit to the mean-variance portfolio optimization framework

Option pricing and Hedging

- In practice some assumptions of the BSM model are not realistic
  - transaction costs due to commissions, market impact, and non-zero bid-ask spread exist in the real market;
  - the volatility is not constant
  - short term returns typically have a heavy tailed distribution
- The resulting prices and hedges may suffer from model mis-specification when the real asset dynamics is not exactly as assumed and the transaction costs are difficult to model
  - focus on a model-free RL approach that can address some of these issues.
-

Robo Advising

- Robo-advisors, or automated investment managers
  - a class of financial advisers that provide online financial advice or investment management with minimal human intervention
  - Digital financial advice based on mathematical rules or algorithms
- History
  - The first robo-advisors were launched after the 2008 financial crisis
  - Examples of pioneering robo-advising firms include Betterment and Wealthfront.
  - As of 2020, the value of assets under robo-management is highest in the United States and exceeded $650 billion
- Stochastic control framework
  - a regime switching model of market returns
  - A mechanism of interaction between the client and the robo-advisor
  - a dynamic model (i.e., risk aversion process) for the client’s risk preferences
  - an optimal investment criterion

Robo Advising

- The robo-advisor doesn’t know the client’s risk preference in advance; learns it while interacting with the client
- several challenges in the application of robo-advising
  - Firstly, the client’s risk preference may change over-time and may depend on the market returns and economic conditions
    - the robo-advisor needs to determine a frequency of interaction with the client that ensures a high level of consistency in the risk preference when adjusting portfolio allocations.
  - Secondly, the robo-advisor usually faces a dilemma
    - investing according to the client’s risk preference
    - going against the client’s wishes in order to seek better investment performance
  - Finally, there is also a subtle trade-off between the rate of information acquisition from the client and the accuracy of the acquired information.
    - the robo-advisor may not always have access to up-to-date information about the client’s profile
    - information communicated to the robo-advisor may not be representative of the client’s true risk aversion as the client is subject to behavioral biases

RL Approach

- only a few references on robo-advising with an RL approach since this is still a relatively new topic
  - The first RL algorithm for a robo advisor : **Robo-advising: Learning investors’ risk preferences via portfolio choices 2021**
    - an exploration-exploitation algorithm to learn the investor’s risk appetite over time by observing her portfolio choices in different market environments.
    - The investor interacts with the robo-advisor by portfolio selection choices, and such interactions are used to update the robo-advisor’s estimate of the investor’s risk profile.
    - the proposed exploration-exploitation algorithm performs near optimally with the number of time steps depending polynomially on various model parameters.
  - proposed an investment robo-advising framework consisting of two agents
    Robo-advising: Enhancing investment with inverse optimization and deep reinforcement learning 2021
    - The first agent, an inverse portfolio optimization agent, infers an investor’s risk preference and expected return directly from historical allocation data using online inverse optimization.
    - The second agent, a deep RL agent, aggregates the inferred sequence of expected returns to formulate a new multi-period mean-variance portfolio optimization problem that can be solved using a deep RL approach based on the DDPG method