Deep Reinforcement Learning for Market Making Under a Hawkes Process-Based Limit Order Book Model

Sergio Charles, Tanya Otsetarova, Jake Kaplan, Rafael Abreu

Department of Mathematics
Stanford University

May 23, 2023

Table of Contents

Market Making

- In the **Avellaneda-Stoikov framework**:
  - the mid price S_t^m follows Brownian motion.
  - the arrival of buy/sell market orders (MO) hitting a limit order (LO) at distance d from the mid price, is an independent Poisson process with intensity $\lambda(d) = A\exp(-kd)$ where $A > 0$, $k > 0$.
- The **market maker's (MM) objective** is to maximize risk-adjusted wealth at the end of the trading period by controlling their bid price S_t^b and ask price S_t^a at different times, under the dynamics of the mid price S_t^m, their cash X_t, inventory Q_t, and market order arrivals on the bid and ask sides N_t^b, N_t^a.

Stochastic Optimal Control Problem

- The optimal stochastic control problem is:

$$
\begin{aligned}
&\max_{S_t^b, S_t^a \in \mathcal{A}} \mathbb{E}[U(X_T + Q_T S_T^m)] \\
&dS_t^m = \sigma dW_t \\
&dX_t = S_t^a dN_t^a - S_t^b dN_t^b \\
&dQ_t = dN_t^b - dN_t^a \\
&\lambda_b = A \exp(-k(S_t^m - S_t^b)) \\
&\lambda_a = A \exp(-k(S_t^a - S_t^m))
\end{aligned}
$$

for N_t^b, N_t^a Piosson processes with intensity λ_b, λ_a, σ is instantaneous volatility, and $U(\cdot)$ a concave utility function. Here, $\mathcal{A}$ is the set of admissible strategies δ_t^a, δ_t^b; namely, $\mathcal{F}_t$-adapted and bounded from below.

MM Control Problem

- The MM caps their inventory to be bounded above by $\bar{q} > 0$ and below by $\underline{q} < 0$.
- At time T, the MM liquidates terminal inventory Q_T using a MO at a price worse than the midprice to account for "liquidity taking fees" and the MO walking the LOB.
- The performance criterion is:

$$H^\delta(t, x, S, q) = \mathbb{E}_{t,x,q,S}\left[X_T + Q_T^\delta(S_T^\delta - \alpha Q_T^\delta) - \phi \int_t^T (Q_u)^2 du\right]$$

where $\alpha \geq 0$ is a fee for the MM taking liquidity (i.e. using a MO) and the impact of the MO walking the LOB, and $\phi \geq 0$ is the inventory penalty.

Hamilton-Jacobi-Bellman

- The MM's value function is:

$$H(t,x,S,q) = \sup_{\delta^\pm \in \mathcal{A}} H^\delta(t,x,S,q)$$

  for $\mathcal{A}$ the set of admissible strategies $\delta_t^\pm$; namely, $\mathcal{F}_t$-adapted and bounded from below.

- The optimal control problem satisfies the following Hamilton-Jacobi-Bellman equation:

$$\begin{aligned} 0 =& \partial_t H + \frac{1}{2}\sigma^2 \partial_{SS} H - \phi q^2 \\ &+ \lambda^+ \sup_{\delta^+} \{ e^{-\kappa^+ \delta^+} (H(t, x + (S + \delta^+), q - 1, S) - H) \} \mathbb{1}_{q > \underline{q}} \\ &+ \lambda^- \sup_{\delta^-} \{ e^{-\kappa^- \delta^-} (H(t, x + (S + \delta^-), q + 1, S) - H) \} \mathbb{1}_{q < \bar{q}} \end{aligned}$$

  with terminal condition $H(T,x,S,q) = x + q(S - \alpha q)$.

Interpretation of DPE

- Terms in the DPE equation represent (1) the arrival of MOs that may be filled by LOs, (2) the diffusion of the asset price through the term $\frac{1}{2}\partial_{SS}H$, and (3) the effect of penalizing deviations of inventories from zero along the entire path of the strategy, described by the ϕq^2 term.
- The sup over δ^+ contain the terms due to the arrival of the market buy order (which is filled by a limit sell order)
- Represents the change in the value function H due to the arrival of the MO which fills the LO, so that cash increases by $(S + \delta^+)$ and inventory decreases by one unit. (Analogous terms for the market sell orders which are filled by limit buy orders.)
- However, the AS framework is inconsistent with respect to many important LOB features.

Inconsistencies

- **Price Consistency**: Price and order arrivals are assumed to be independent, so price can rise on a large sell market order; this can generate large phantom gains for MM, since they are usually on the wrong side of the trade.
- **Price-Time Priority**: AS framework assumes there's no cost in changing the bid/ask prices, as the model was originally designed for a quote-driven market.
- **Price Ticks**: Prices are only allowed on a fixed price grid (0.01); thus, price is a pure-jump process with two dimensions: jump times and magnitudes.
- **Execution Probability**: AS model uses a rate function $\lambda(d) = A\exp(-kd)$, which affects execution probability of LO in a given interval. Price is continuous so d is continuous. Because of the discrete price grid, the rate function is truly a step function.
- **Order Size**: AS assumes all MOs and LOs are of the same size; usually MM will instead place LOs at many different price levels to continuously maintain priority in LOB.

LOB Consistency

- **Notation**:
  - Let $(S_t^b, S_t^a, S_t^m = (S_t^b + S_t^a)/2, S_t = S_t^a - S_t^b)$ denote the bid price, ask price, mid price and bid-ask spread respectively.
  - Let $\tau_m^b, \tau_m^a, \tau_l^b, \tau_l^a, \tau_c^b, \tau_c^a$ denote the arrival times of any market sell, market buy, limit buy, limit sell, limit buy cancellation, and limit sell cancellation orders. Let the corresponding volume and price (LO only) be represented by ν.
- A LOB is called **consistent** if it satisfies direction, timing and volume consistency.

Direction Consistency

- **Direction Consistency**: On arrival of a marketable sell/buy order (LO or MO), the bid/ask price can't move up/down while the ask/bid price can only stay unchanged:

$$\mathbb{P}\left[\{S^a_{\tau^a_m} \geq S^a_{\tau^{a-}_m}\} \cap \{S^b_{\tau^a_m} = S^b_{\tau^{a-}_m}\}\right] = \mathbb{P}\left[\{S^b_{\tau^b_m} \leq S^b_{\tau^{b-}_m}\} \cap \{S^a_{\tau^b_m} = S^a_{\tau^{b-}_m}\}\right] = 1$$

- On arrival of limit sell/buy order with price falling inside the bid-ask spread, the ask/bid price can only move down/up while the bid price can only stay unchanged. If the limit order is outside the bid-ask spread, the ask and bid prices are unchanged.

Direction Consistency Cont'd.

- When direction consistency is violated, MM profit can be significantly exaggerated. E.g. when price plunges after a sequence of sell MOs, the MM will suffer a major loss because it has net long inventory by taking opposite sides of the trades. If price violates direction consistency and goes up, the MM will instead enjoy a profit.

Timing Consistency

- **Timing Consistency**: The bid/ask price moves only at the instants of orders arrivals/cancellations:

$$\mathbb{P}(\{S_t^b = S_{t-}^b\} \cap \{S_t^a = S_{t-}^a\} | t \notin \Gamma) = 1$$

where Γ is the set of all stopping times of market and limit orders.

Volume Consistency

- **Volume Consistency**: If the volume of the marketable buy/sell order is equal to or larger than the depth of the best ask/bid queue (Q_t^a, Q_t^b), the ask/bid price moves up/down; otherwise it stays unchanged:
- If the volume of the limit buy/sell cancellation is equal to the depth of the best ask/bid queue (Q_t^a, Q_t^b), the ask/bid price moves up/down; otherwise, it stays unchanged.

Inconsistent models

In the Avellaneda and Stoikov model, mid prices are independent Brownian motions:

$$dS_t^m = \sigma dW_t$$

- When a buy MO arrives, half of the time the mid price will go down since it is an independent BM, thus overstating a MM's profit.
- This is a continuous time process, so it will move even without orders.

Towards Consistency

Observing the total independence of price and order arrivals, Cartea et al. [Buy Low Sell High] bifurcate the buy and sell MOs into influential $(\bar{M}_t^+, \bar{M}_t^-)$ and non-influential $(\tilde{M}_t^+, \tilde{M}_t^-)$ where $(\bar{M}_t^+, \bar{M}_t^-, \tilde{M}_t^+, \tilde{M}_t^-)$ is a multivariate Hawkes process. The midprice is a diffusion coupled with the MOs via an unobservable mean-reverting process α_t as follows:

$$dS_t^m = (\nu + \alpha_t)dt + \sigma dW_t$$
$$d\alpha_t = -\zeta\alpha_t dt + \sigma_\alpha dB_t + \epsilon^+ d\bar{M}_t{}^+ - \epsilon^- d\bar{M}_t{}^-$$

where W_t and B_t and independent BMs and $\nu \in \mathbb{R}$, $\zeta, \sigma, \sigma_\alpha, \epsilon^+, \epsilon^-$ are strictly positive.

- When an influential buy MO $\bar{M}_t{}^+$ arrives, α_t jumps so the midprice S_t^m has a larger drift. However, the W_t term can still have an even larger negative change that causes overall downward price movement.

Weakly Consistent LOB

- We call a LOB **weakly consistent** if the model only complies with direction and timing consistency.
- *Market Making under a Weakly Consistent Limit Order Book Model* [Law & Viens, 2020] presents a weakly-consistent pure-jump market model; however, it assumes constant order arrival intensities. Thus, self- or mutual-excitation and inhibition between many types of LOB order arrivals are unaccounted for.
- The introduction of self-exciting arrival intensities in a weakly-consistent LOB model makes the HJB equation analytically intractable.
- *Deep Reinforcement Learning for Market Making Under a Hawkes Process-Based Limit Order Book Model* [Gasperov Konstanjcar, 2022] presents an approach to finding approximate optimal controls.

Multivariate Hawkes processes

Definition

A p-**dimensional linear Hawkes process** is a in-homogeneous p-dimensional Poisson counting process $N(t) = (N_k(t); k = 1, \ldots, p)$ with intensity of N_k given by:

$$\lambda_k(t) = \mu_k + \sum_{\ell=1}^{p} \int_0^{t-} f_{k,\ell}(t-s)dN_\ell(s) \tag{1}$$

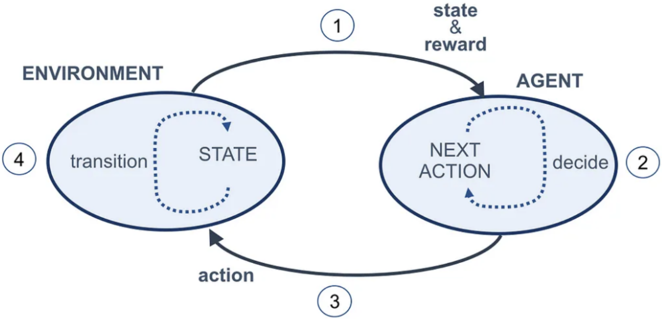
where

- $\mu_k \geq 0$ are the baseline intensities
- $N_\ell(t)$ is the number of arrivals within $[0, t]$ corresponding to the ℓ-th component
- arrivals in dimension ℓ perturb the intensity of arrivals in dimension k at time t by $f_{k,\ell}(t-s)$ for $t > s$
- Generally, one uses an exponential kernel: $f_{k,\ell}(t) = \alpha_{k,\ell} e^{-\beta_{k,\ell}(t)}$

Table of Contents

Reinforcement Learning

- Optimal controls for the weakly consistent model presented by [Law & Viens, 2020] could be solved for using the HJB equation
- Instead, [Gašperov and Kostanjčar, 2022] train a deep reinforcement learning controller on a simulation of the LOB model they study to learn locally optimal controls
- The RL process looks like [Torres, 2020]



Event types

$$E_{all} = \{M_a^b, M_a^s, L_a^b, L_a^s, C_a^b, C_a^s, M_n^b, M_n^s\}$$

Event types

$$E_{all} = \{M_a^b, M_a^s, L_a^b, L_a^s, C_a^b, C_a^s, M_n^b, M_n^s\}$$

- Aggressive Market Order - one which completely depletes the best bid or ask queue;

Event types

$$E_{all} = \{M_a^b, M_a^s, L_a^b, L_a^s, C_a^b, C_a^s, M_n^b, M_n^s\}$$

- Aggressive Market Order - one which completely depletes the best bid or ask queue;
- Aggressive Limit Order - one which has a price inside the bid-ask spread;

Event types

$$E_{all} = \{M_a^b, M_a^s, L_a^b, L_a^s, C_a^b, C_a^s, M_n^b, M_n^s\}$$

- Aggressive Market Order - one which completely depletes the best bid or ask queue;
- Aggressive Limit Order - one which has a price inside the bid-ask spread;
- Aggressive Cancellation - one which cancels the last remaining order in the best bid or best ask queue

Event effects on price

Type of order	Mid-price	Bid-price	Ask-price
1 - Aggressive market buy	+	no effect	+
2 - Aggressive market sell	-	-	no effect
3 - Aggressive limit buy	+	+	no effect
4 - Aggressive limit sell	-	no effect	-
5 - Aggressive limit buy cancelation	-	-	no effect
6 - Aggressive limit sell cancelation	+	no effect	+
7 - Non-aggressive market buy	no effect	no effect	no effect
8 - Non-aggressive market sell	no effect	no effect	no effect
9 - Non-aggressive limit buy	no effect	no effect	no effect
10 - Non-aggressive limit sell	no effect	no effect	no effect
11 - Non-aggressive limit buy cancelation	no effect	no effect	no effect
12 - Non-aggressive limit sell cancelation	no effect	no effect	no effect

Events

Let $N(t) = \left(N_{M_b^a}(t), ..., N_{M_n^s}(t)\right)$ be the multivariate point process of the number of orders in each type up to and including time t.

The associated intensity vector is $\lambda(t) = \left(\lambda_{M_b^a}(t), ..., \lambda_{M_n^s}(t)\right)$.

Let $N(t) = \left(N_{M_b^a}(t), ..., N_{M_n^s}(t)\right)$ be the multivariate point process of the number of orders in each type up to and including time t.
The associated intensity vector is $\lambda(t) = \left(\lambda_{M_b^a}(t), ..., \lambda_{M_n^s}(t)\right)$.
The mid-price P_t is given by

$$P_t = P_0 + \left(\sum_{\mathrm{e, T(e) \in E_{inc}}} \mathrm{J_e} - \sum_{\mathrm{e, T(e) \in E_{dec}}} \mathrm{J_e} \right) \frac{\delta}{2}$$

where P_0 is the initial price, δ is the tick size, $T(e)$ is the type of event e, $J(e)$ is the associated jump with an event e, $\mathrm{E_{inc}} = \{\mathrm{M_a^b}, \mathrm{L_a^b}, \mathrm{C_a^s}\}$, and $\mathrm{E_{dec}} = \{\mathrm{M_a^s}, \mathrm{L_a^s}, \mathrm{C_a^b}\}$.

Market Making Procedure

- At the start of each time-step, we eliminate all outstanding limit orders and observe the state of the environment.

Market Making Procedure

- At the start of each time-step, we eliminate all outstanding limit orders and observe the state of the environment.
- The agent decides whether to post limit orders (and at which prices) or market orders.

Market Making Procedure

- At the start of each time-step, we eliminate all outstanding limit orders and observe the state of the environment.
- The agent decides whether to post limit orders (and at which prices) or market orders.
- If the absolute value of the agent's inventory is equal to the inventory constraint c, $c \in \mathbb{N}$, the order on the corresponding side is ignored.

Market Making Procedure

- At the start of each time-step, we eliminate all outstanding limit orders and observe the state of the environment.
- The agent decides whether to post limit orders (and at which prices) or market orders.
- If the absolute value of the agent's inventory is equal to the inventory constraint c, $c \in \mathbb{N}$, the order on the corresponding side is ignored.
- Update all variables such as bid, ask, mid-price, spread, agent's inventory, and cash.

Market Making Procedure

- At the start of each time-step, we eliminate all outstanding limit orders and observe the state of the environment.
- The agent decides whether to post limit orders (and at which prices) or market orders.
- If the absolute value of the agent's inventory is equal to the inventory constraint c, $c \in \mathbb{N}$, the order on the corresponding side is ignored.
- Update all variables such as bid, ask, mid-price, spread, agent's inventory, and cash.
- All LOB events generated by the simulation procedure are processed sequantially, and each is followed by an update of the variables.

Market Making Procedure

- At the start of each time-step, we eliminate all outstanding limit orders and observe the state of the environment.
- The agent decides whether to post limit orders (and at which prices) or market orders.
- If the absolute value of the agent's inventory is equal to the inventory constraint c, $c \in \mathbb{N}$, the order on the corresponding side is ignored.
- Update all variables such as bid, ask, mid-price, spread, agent's inventory, and cash.
- All LOB events generated by the simulation procedure are processed sequantially, and each is followed by an update of the variables.
- Executed limit orders are not replaced by new ones until the next time-step.

Market Making Procedure

- At the end of the time-step and receives the reward $R_{t+\triangle t}$.

Market Making Procedure

- At the end of the time-step and receives the reward $R_{t+\triangle t}$.
- Unexecuted orders are cancelled once time $t + \triangle t$ is reached and the procedure iterates until terminal time T.

Market Making Procedure: Inventory

$$dI_t = dN_t^b - \ dN_t^a + \ dN_t^{mb} - \ dN_t^{ms}$$

- N_t^b - limit order buys of the MM
- N_t^a - limit order sells of the MM
- N_t^{mb} - market order buys of the MM
- N_t^{ms} - market order sells of the MM

Market Making Procedure: Inventory

$$dI_t = dN_t^b - dN_t^a + dN_t^{mb} - dN_t^{ms}$$

- N_t^b - limit order buys of the MM
- N_t^a - limit order sells of the MM
- N_t^{mb} - market order buys of the MM
- N_t^{ms} - market order sells of the MM

$$dN_t^b = dN_{M_s^a} \mathbb{1}_{\mathrm{fill}, M_s^a} + dN_{M_s^n} \mathbb{1}_{\mathrm{fill}, M_s^n}$$

- $\mathbb{1}_{fill, M_s^n}$ is the indicator function whether the incoming (non-)aggressive market order fulfils the market-maker's limit order.

Market Making Procedure: Cash

$$\mathrm{d}X_t = Q_t^a \,\mathrm{d}N_t^{\mathrm{a}} - Q_t^b \,\mathrm{d}N_t^{\mathrm{b}} - (P_t^{\mathrm{a}} + \epsilon_t)\,\mathrm{d}N_t^{\mathrm{mb}} + \left(P_t^{\mathrm{b}} - \epsilon_t\right)\mathrm{d}N_t^{\mathrm{ms}}$$

- $Q_t^a (Q_t^b)$ is price at which the agent's ask(bid) quote is posted,
- $P_t^a (P_t^b)$ is the best ask (bid) price,
- ϵ is the the additional costs due to fees and market impact

Simplifying assumptions

- MM orders are aggressive with probability Z_1.

Simplifying assumptions

- MM orders are aggressive with probability Z_1.
- Limit order cancellations are aggressive with probability Z_2.

Simplifying assumptions

- MM orders are aggressive with probability Z_1.
- Limit order cancellations are aggressive with probability Z_2.
- The jumps J_e associated with the LOB events e are modeled by exponential distribution with density $f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta}\right)$ where μ is the location and β is scale parameter.

Simplifying assumptions

- MM orders are aggressive with probability Z_1.
- Limit order cancellations are aggressive with probability Z_2.
- The jumps J_e associated with the LOB events e are modeled by exponential distribution with density $f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta}\right)$ where μ is the location and β is scale parameter.
- Jump sizes are independent of jump times.

Table of Contents

Controller

- State space (what the agent sees)
- Action space (what the agent can and should do)
- Reward function
- DRL model architecture and optimization

State Space (S_t)

- Inventory (I_t)
  - $\{-c, ..., c\}$ due to inventory constraints
  - Min-max normalization
- Spread (Δ_t)
  - Also integer (measured in ticks) and strictly positive
  - z-score normalization using mean/variance from controller with random actions
- Trend Variable (α_t)
  - Describes market's "net buying pressure", i.e. expected buy minus sell density
  - $\lambda_{M_b}(t) - \lambda_{M_a}(t)$
    - Where $\lambda_{M_x}(t) = \lambda_{M_x^a}(t) + \lambda_{M_x^n}(t)$
  - Also z-score normalization as above
  - Notably, can only be approximated experimentally by MM
- Does not include volume (weakly consistent)!

Action Space (A_t)

- Essentially, how aggressive to be on each of the bid and the ask
  - Specifically, how much to penny each by (i.e. beat the BBO)
- Let (P_t^b, P_t^a) denote best (bid,ask) prices
- Let (Q_t^b, Q_t^a) denote agent's quoted market (any number)

Action Space (A_t) Cont.

- Ask Offset ($Q_t^a - P_t^a$)
- Bid Offset ($P_t^b - Q_t^b$)
  - Crossed markets ignored
  - Markets crossing best bid/ask treated as market orders
  - All orders unit size and rounded to nearest tick
  - Note that offsets are described as in the paper, but will typically be nonpositive since orders outside the BBO are never executed
- Still no volume!

Incentives for zero (or less aggressive) offsets

- Maintain small inventory by lowering chance of execution in a direction
  - Less aggressive $\rightarrow$ no chance of execution
  - Zero offset $\rightarrow$ $\frac{1}{4}$ chance of execution
  - Essentially, orders sometimes exhaust the BBO, but never go deeper into the book

Incentives for small offsets (pennying)

- Larger spread $\rightarrow$ more profit
- The rest of the market is *price agnostic*, i.e. decides event not price
  - Market order events execute against agent regardless of price
  - Limit order events penny the BBO (i.e. agent) regardless of price
  - Unrealistic model of execution probability, which should decay exponentially with spread

Incentives for large offsets (crossing the spread)

- Maintain small inventory by guaranteeing execution in a direction
  - Crossing the spread $\rightarrow$ market order
  - Never any reason to post an aggressive limit order!
    - Aggressive orders only serve a purpose when they cross the spread (market orders)
    - Otherwise, aggressive (limit) orders are strictly bad

Optimal Actions

- When position is relatively flat
  - If expecting high order density, join the BBO
  - Otherwise, penny the BBO
    - Use trend variable to model order density
- When position is far from zero
  - If spread is small, cross the spread to neutralize inventory
  - Otherwise, penny to neutralize inventory and be less aggressive in opposite direction
- Observe that this strategy is highly nonlinear...

Reward Function (R_t)

- Want to maximize

$$E_{\pi_\theta}\left[W_T - \phi \int_0^T |I_t| dt\right]$$

  - Optimize over $\pi_\theta : S \to P(A)$, i.e. policies mapping state to distribution of action
  - $W_t = I_t P_t + X_t$ is total wealth
  - $\phi \geq 0$ punishes nonzero inventories
    - Note that this punishment is already partially baked in by not allowing the agent to execute orders that exceed its inventory constraints

Reward Function (R_t) Cont.

- Implies reward function

$$R_{t+\Delta t} = \Delta W_{t+\Delta t} - \phi \int_t^{t+\Delta t} |I_s| ds$$

  - Each timestep rewards wealth gains, punishes nonzero inventory
  - Integral piecewise constant $\rightarrow$ trivial computation
  - Absolute inventory (vs quadratic inventory) has convenient VaR (value at risk) interpretation
    - Not elaborated on in the paper...

Controller Design (NN)

- 2 fully-connected hidden layers of 64 neurons
- ReLU activation
- Simple controller design common in DRL
  - Empirically comparable or better than sophisticated models for LOB

Training Algorithm (SAC: Soft Actor-Critic)

- Entropy maximization
  - Balances explore and exploit
- Learns 2 Q-functions
  - Mapping (state,action) to value
  - Considers min value between the two
- SAC generally robust, multiple modes of near-optimal behavior
- Empirically beats DQN, TD3
- 10^6 training timesteps

Table of Contents

Experiments

- Compare the performance of their approach against some benchmarks
- Use Monte Carlo to generate synthetic data.
- Standard MM benchmarks like the (Avellaneda, Stoikov; 2008) approximations are ill-suited since they don't take into account neither the existence of the bid-ask spread nor the discrete nature of the underlying LOB.

Benchmarks

- Consider a class of MM strategies linear in inventory and including inventory constraints. best performing: LIN strategy.

$$Q_t^i - P_t^i = \alpha^i + \beta^i I_t$$

Benchmarks

- Consider a class of MM strategies linear in inventory and including inventory constraints. best performing: LIN strategy.

$$Q_t^i - P_t^i = \alpha^i + \beta^i I_t$$

- Simple (SYM) strategy: always places limit orders precisely at the best bid and the best ask.

$$Q_t^a = P_t^a, \quad Q_t^b = P_t^b$$

Risk and performance metrics

- Profit and Loss (PnL) distribution statistics (of the terminal wealth)
- Mean episode return ($PnL -$ discounted inventories)
- Mean Absolute Position (MAP)

$$\mathrm{MAP} = \frac{1}{N} \sum_{k=1}^{N} |I_{k\Delta t}| ,$$

  where N is the number of time-steps in an episode.
- Sharpe ratio

$$SR = \frac{\mu_{W_T}}{\sigma_{W_T}},$$

  where μ_{W_T} (σ_{W_T}) denotes the mean (standard deviation) of the terminal wealth (PnL).
- (Mean PnL)/MAP - variant of (Gasperov and Kostanjcar, 2021)

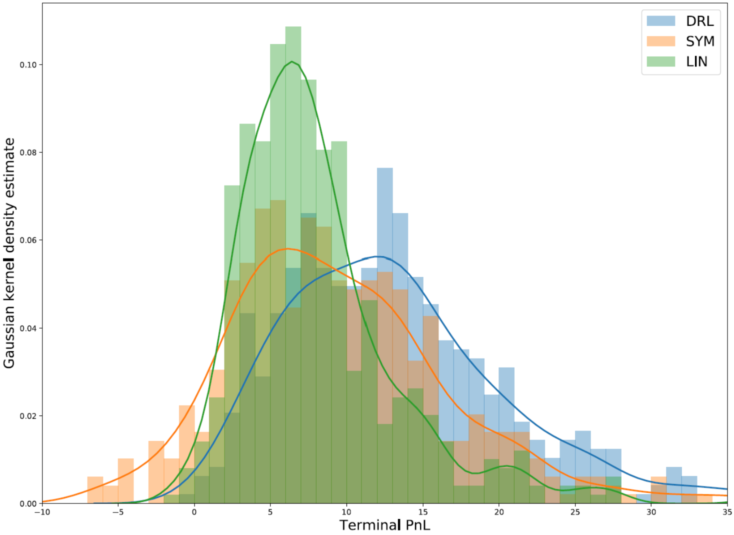
Risk and performance metrics

TABLE I

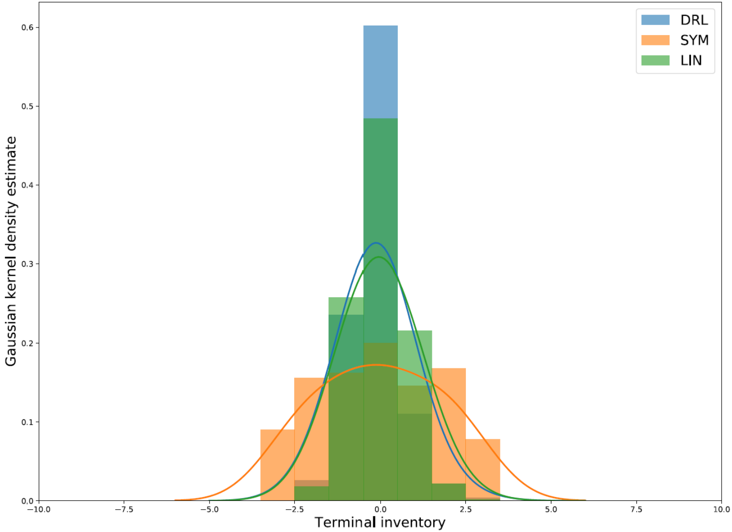
DISTRIBUTIONAL STATISTICS — DRL CONTROLLER VS BENCHMARKS

Metric	DRL	SYM	LIN
Mean episode return	13.3378	8.3913	7.6772
Mean PnL	13.7567	9.8686	8.2105
Std PnL	8.4952	8.2443	5.7884
Kurtosis PnL	4.2083	6.0552	10.5585
Skew PnL	1.5870	1.6375	2.3863
Jarque Bera PnL	578.8301	987.3005	2797.0786
Jarque Bera PnL p-value (5%)	0	0	0
10th percentile PnL	4.9445	1.6775	2.7400
20th percentile PnL	7.0480	3.6220	3.9800
80th percentile PnL	19.0340	15.0000	11.4360
90th percentile PnL	24.4310	19.6150	14.7170
Sharpe Ratio	1.6193	1.1970	1.4184
Abs. mean terminal inv.	0.454	1.46	0.56
Mean terminal inv.	-0.122	-0.028	-0.028
Std terminal inv.	0.7477	1.7650	0.8070
Kurtosis inv.	1.6838	-1.0332	0.0277
Skew inv.	0.3742	0.0032	0.1879
Jarque Bera inv.	70.7342	22.2388	2.9569
Jarque Bera inv. p-value (5%)	0	0	0.2280
10th percentile inv.	-1	-2	-1
20th percentile inv.	-1	-2	-1
80th percentile inv.	0	2	1
90th percentile inv.	1	2	1
Mean Absolute Position (MAP)	0.4232	1.4659	0.5478
(Mean PnL)/MAP	32.5064	6.7321	14.9881

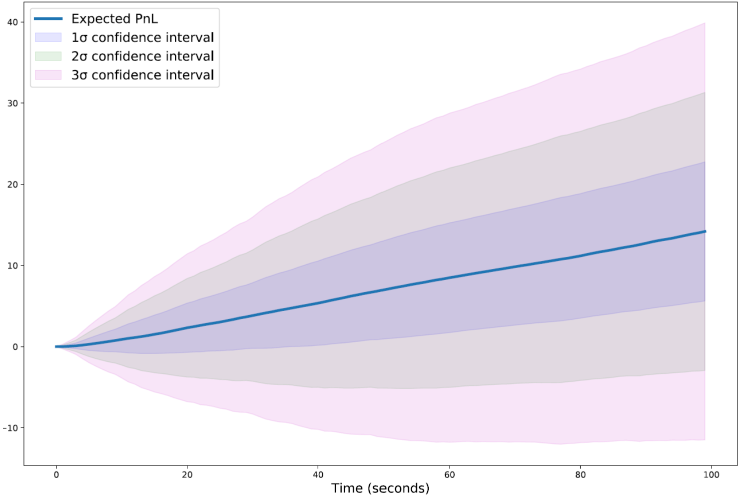
Terminal PnL Distribution



Terminal Inventory Distribution



PnL over time



Sensitivity Analysis

- Add noise to the intensity of the arrival rate $\lambda's$ of all order types.
- More precisely, three different noise sizes were considered — Gaussian noise based with mean 0 and variance 0.1, 0.2, 0.3.

Sensitivity Analysis

- Add noise to the intensity of the arrival rate $\lambda' s$ of all order types.
- More precisely, three different noise sizes were considered — Gaussian noise based with mean 0 and variance 0.1, 0.2, 0.3.

$$\lambda_k(t) = \mu_k + \sum_{l=1}^{p} \int_0^{t-} f_{k,l}(t-s)\mathrm{d}N_l(s) + \sigma B_t$$

- $\sigma = 0.1, 0.2, 0.3$

Sensitivity Analysis

TABLE II

DISTRIBUTIONAL STATISTICS — RL NOISE LEVELS

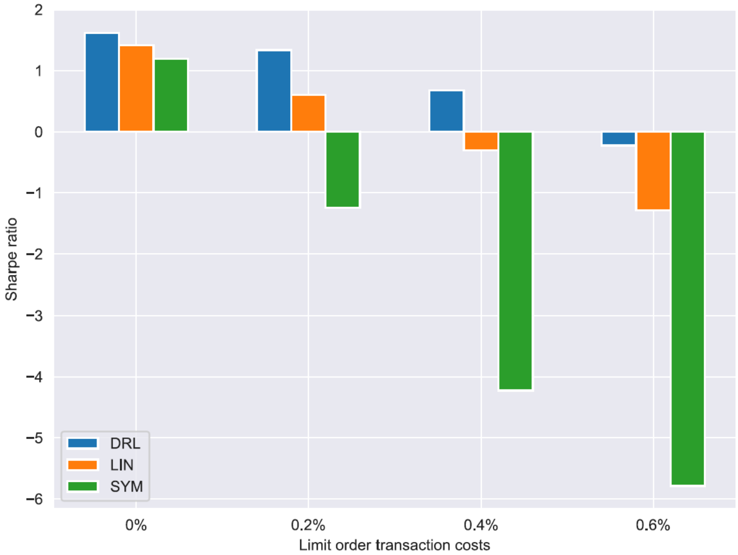
Metric	DRL-N-0.1	DRL-N-0.2	DRL-N-0.3
Mean PnL	13.2050	13.2553	15.5077
Std PnL	9.5922	13.0264	17.9523
Kurtosis PnL	4.5894	15.2773	12.0886
Skew PnL	1.6481	3.0457	2.9768
Jarque Bera PnL	665.14	5635.43	3782.90
Jarque Bera PnL p-value	0	0	0
10th percentile PnL	3.7090	2.3745	1.9472
20th percentile PnL	5.4676	4.4940	3.6387
80th percentile PnL	19.0860	19.7860	22.2980
90th percentile PnL	24.7940	26.3665	34.6055
Sharpe Ratio	1.3766	1.0175	0.8638
Mean Absolute Position	0.4443	0.5038	0.5486

Sensitivity Analysis, 2

- Vary transaction costs.

$$\mathrm{d}X_t = Q_t^a \, \mathrm{d}N_t^{\mathrm{a}} - Q_t^b \, \mathrm{d}N_t^{\mathrm{b}} - (P_t^{\mathrm{a}} + \epsilon_t) \, \mathrm{d}N_t^{\mathrm{mb}} + \left(P_t^{\mathrm{b}} - \epsilon_t\right) \mathrm{d}N_t^{\mathrm{ms}}$$

Sensitivity Analysis, 2



Zvonko Kostanjčar Bruno Gašperov.
Deep reinforcement learning for market making under a hawkes process-based limit order book model, 2022.

Baron Law and Frederi Viens.
Market making under a weakly consistent limit order book model, Jan. 2020.

Jordi Torres.
A gentle introduction to deep reinforcement learning, 2020.