

An Impulse Control Approach to Market Making in a Hawkes LOB Market

Konark Jain^{§†‡}, Nick Firoozye[†], Jonathan Kochems[‡], and Philip Treleaven[†]

Abstract. We study the optimal Market Making problem in a Limit Order Book (LOB) market simulated using a high-fidelity, mutually exciting Hawkes process. Departing from traditional Brownian-driven mid-price models, our setup captures key microstructural properties such as queue dynamics, inter-arrival clustering, and endogenous price impact. Recognizing the realistic constraint that market makers cannot update strategies at every LOB event, we formulate the control problem within an impulse control framework, where interventions occur discretely via limit, cancel, or market orders. This leads to a high-dimensional, non-local Hamilton–Jacobi–Bellman Quasi-Variational Inequality (HJB-QVI), whose solution is analytically intractable and computationally expensive due to the curse of dimensionality. To address this, we propose a novel Reinforcement Learning (RL) approximation inspired by auxiliary control formulations. Using a two-network PPO-based architecture with self-imitation learning, we demonstrate strong empirical performance with limited training, achieving Sharpe ratios above 30 in a realistic simulated LOB. In addition to that, we solve the HJB-QVI using a deep learning method inspired by Sirignano and Spiliopoulos 2018 [40] and compare the performance with the RL agent. Our findings highlight the promise of combining impulse control theory with modern deep RL to tackle optimal execution problems in jump-driven microstructural markets.

Key words. Market Making; Limit Order Book; Hawkes Process; Impulse Control; Reinforcement Learning; Hamilton–Jacobi–Bellman Equation; Quasi-Variational Inequality; Deep Learning; Financial Microstructure

AMS subject classifications. Primary: 93E20; Secondary: 91G80, 60G55, 35Q93, 68T07, 49L25

1. Introduction. Market Making in Limit Order Books (LOBs) is a high frequency trading task, where liquidity is provided with the goal of capturing the bid-ask spread. While many classical formulations rely on continuous-time dynamics and control, the true microstructure of the LOB is inherently discrete and driven by a pure jump process.

The distinction between liquidity provision and liquidity consumption represents a fundamental dichotomy in modern market microstructure theory. As noted in [24], while optimal execution strategies primarily focus on liquidity-taking behavior—the efficient liquidation of large positions or implementation of hedging strategies—market making fundamentally concerns itself with the provision of liquidity to the marketplace. This shift in perspective necessitates a fundamentally different analytical framework, one that accounts for the unique risks and opportunities inherent in standing ready to transact at posted prices.

Market makers, defined as traders who continuously provide liquidity by quoting bid and ask prices, serve as the backbone of price discovery mechanisms across diverse market structures ([24]). The heterogeneity of market making arrangements reflects the varied institutional frameworks within which these participants operate. On order-driven markets, official market makers such as Designated Market Makers (DMMs) on the NYSE operate under explicit contractual obligations to maintain fair and orderly markets, including mandatory participation in opening and closing auctions and adherence to National Best Bid and Offer (NBBO) quoting requirements. Conversely, proprietary market makers, including high-frequency trading firms, provide liquidity without formal obligations, seeking to profit from the bid-ask spread while maintaining flexibility in their market participation.

[†]Department of Computer Science, University College London, London, UK.

[‡]Quantitative Research, JP Morgan Chase, London, UK.

[§]Corresponding author (konark.jain.23@ucl.ac.uk).

Funding: This work was funded by JPMorganChase

Opinions expressed in this paper are those of the authors and do not necessarily reflect the views of JP Morgan. Submitted to the editors DATE.

The fundamental economic proposition underlying market making involves the temporal arbitrage of buying at the bid price and selling at the ask price, with the bid-ask spread representing the market maker’s expected compensation for providing immediacy services. However, this seemingly straightforward profit mechanism is complicated by the inevitable inventory risk that market makers must bear. The asynchronous nature of buy and sell transactions means that market makers typically hold non-zero inventory positions, exposing them to adverse price movements during the holding period ([24]).

Optimal Market Making (MM) in Limit Order Book (LOB) settings has been approached primarily through two methodological lenses: stochastic control and reinforcement learning. Early stochastic control models often assume Markovian mid-price dynamics with exogenously specified order flow. The seminal work of [5] established the theoretical foundation for modern market making models by formulating the problem as a continuous-time stochastic control problem. Their framework, along with the closed-form approximations developed by [20], provides elegant analytical solutions for optimal bid and ask quote placement. However, these continuous-time models face significant limitations when applied to actual order-driven markets, particularly in their treatment of price discreteness arising from tick size constraints and their inability to capture the granular dynamics of limit order book evolution. For instance, [23] model the mid-price as a continuous-time Markov chain and the bid-ask spread as a discrete Markov process, allowing the MM to place limit orders, market orders, and even aggressive internalizing strategies. These controls are optimized through a hybrid control framework involving a Hamilton-Jacobi-Bellman (HJB) equation. More recent efforts have focused on incorporating more realistic LOB dynamics. [38] consider a Hawkes-like LOB, where market orders impact a latent alpha signal and fill probabilities, making them informative. The MM, restricted to placing only limit orders, solves a stochastic control problem in this non-Markovian environment. Similarly, [30] study a quote-driven market in which market takers’ actions are modeled via Hawkes processes. They derive HJB solutions in the exponential kernel case and extend to more realistic power-law kernels via approximations. A more discrete approach is taken by [1], who model LOB dynamics using a state-dependent Poisson process and then extend the model to an exponential Hawkes process, solving the corresponding Markov Decision Process (MDP). While most of these works rely on analytically tractable models, they tend to assume known dynamics, stationarity, and full observability, which limit their applicability to real-world markets.

Reinforcement Learning (RL) has emerged as a compelling alternative, capable of handling partial observability, path-dependence, and model uncertainty. A recent survey by [21] challenges classical assumptions made in stochastic control—such as complete knowledge of the environment and fixed time horizons—and organizes RL approaches by control logic (inventory-based, signal-driven, robust) and state observability (tabular vs. deep RL). A second taxonomy classifies methods by algorithmic design, highlighting the shift from analytical and tabular methods to deep actor-critic variants. Most RL frameworks define the agent’s state using features like inventory, PnL, spread, imbalance, and short-term technical signals such as RSI. To tackle sample inefficiency, several works employ simulators or historical replay mechanisms. For instance, [22] use an 8-dimensional Hawkes-based simulator in which the MM interacts via limit and market orders in a uni-episodic loop; their Soft Actor-Critic agent outperforms a stochastic control baseline, while DQN and TD3 fail to converge. Other

approaches, like those of [41], rely on historical replay with simpler discrete actions and no explicit modeling of market impact. Recent innovations include adversarial training and auxiliary signal units to improve policy robustness and generalization. Despite RL’s flexibility, fundamental challenges remain—[24] highlights issues such as non-stationarity, the absence of feedback loops in historical data, and the inherent difficulty in designing risk-sensitive reward structures. These limitations continue to motivate hybrid models that blend the tractability of control theory with the adaptability of learning-based methods.

The discrete nature of price movements in modern electronic markets, coupled with the complex interaction dynamics within limit order books, necessitates a more sophisticated modeling approach. While continuous-time models remain well-suited to quote-driven markets such as corporate bond trading, where dealers face pricing problems analogous to those originally modeled by [25], equity markets require explicit modeling of the discrete limit order book structure. The work of [23] represents a significant advancement in this direction, providing a discrete-time framework that more accurately captures the microstructural realities of modern electronic trading venues.

The transition from continuous to discrete modeling frameworks raises fundamental questions about the nature of control in high-frequency market making environments, see for example [33]. One recent related formulation of this problem is the jump decision process (or Piecewise Deterministic Decision Process for example see [9, 12]) framework which assumes that controls can be adjusted instantaneously in response to jumps (i.e. market events). This assumption becomes increasingly unrealistic as trading speeds approach the physical limits of electronic order processing. The latency constraints inherent in real-world trading systems, combined with the discrete nature of order book updates, suggest that impulse control frameworks may provide a more realistic representation of market maker behavior.

Impulse control has been a well studied area of stochastic control with several applications, for instance see [15, 7, 35, 19, 16, 10]. Within the impulse control paradigm, market makers are viewed as making discrete intervention decisions at carefully chosen times, rather than continuously adjusting their positions. This perspective naturally accommodates the technological constraints of modern trading systems while preserving the essential economic intuition underlying market making strategies. The framework recognizes that optimal market making involves not just the selection of appropriate bid and ask prices, but also the strategic timing of order placement and cancellation decisions.

The mathematical formulation of market making under impulse control requires careful consideration of the underlying market dynamics and the constraints faced by market participants. The pure jump nature of limit order book evolution, driven by the arrival of discrete market and limit orders, naturally aligns with the impulse control framework’s emphasis on discrete intervention strategies. This alignment suggests that impulse control methods may provide both more realistic and more computationally tractable solutions to the market making problem.

Our approach builds upon these foundational insights while addressing the practical limitations of existing methodologies. By explicitly modeling the discrete nature of both market dynamics and control decisions, we develop a framework that more accurately captures the reality of modern electronic market making while maintaining analytical tractability. The resulting formulation provides a natural bridge between the theoretical elegance of continuous-

time models and the practical requirements of high-frequency trading systems.

This paper is organized as follows. Section 2 details the methodology, including the Limit Order Book model, Optimal Market Making formulation, the State–Intervention Operator, and the resulting HJB-QVI, along with its generator, solution approaches, and challenges. Section 3 applies the Deep Galerkin Method to solve the HJB-QVI and presents numerical results. Section 4 develops a Reinforcement Learning approximation of the impulse control problem, describing the state and action spaces, reward structure, training procedure using PPO and Self-Imitation Learning, and reports simulation results and sensitivity analyses. Section 6 discusses the findings and concludes while Section 7 outlines directions for future work.

2. Methodology. We discuss the methodology of the LOB model, the market making problem and its impulse control formulation in this section. We provide the mathematical details of the Hamilton-Jacobi-Bellman Quasi-Variational Inequality of the value function of this control problem. Finally, we state the impulse control problem’s solution approaches and their respective challenges.

2.1. Limit Order Book Model. We model the LOB [28] using a d -dimensional mutually-exciting Hawkes process as developed in [27, 29]. This process reproduces stylized facts of LOB dynamics such as realistic spreads, long-memory in returns, and clustered arrival times. Unlike Brownian motion-based models, the mid-price emerges endogenously from queue dynamics and event causality. We refer to [27] for more details on the LOB setup. The events that form the Hawkes process are as follows.

$$\mathcal{E} := \{LO_{\text{ask}_D}, LO_{\text{ask}_T}, CO_{\text{ask}_T}, MO_{\text{ask}}, LO_{\text{ask}_{IS}}, \\ LO_{\text{bid}_{IS}}, LO_{\text{bid}_T}, CO_{\text{bid}_T}, MO_{\text{bid}}, LO_{\text{bid}_D}\}$$

We allow for general kernel types in the Hawkes process, but restrict to exponential kernels for mathematically tractability. The model implicitly includes a concave price impact due to self-excitation.

2.2. Optimal Market Making. In this Hawkes LOB, we develop a market making agent which can interact with the LOB by sending impulses at any given time. These impulses are restricted to be one of the 12 events in \mathcal{E} for the purpose of this work. More complex impulses may include a combination of several events in one impulse however to foster discussion we restrict ourselves to the case of one order per impulse. The market making agent observes the LOB’s volumes at all the price levels fully i.e. it has level 2 access to the LOB data. The market maker’s objective is to maximise her terminal cash and the value of her terminal inventory. While doing so, a market maker prefers to keep as low an inventory at any point of time as possible. One way to achieve this is to penalise the current inventory of the market maker. For discussions around various methods of penalisation, we refer the reader to [?]. We make use of the quadratic running cost penalisation method and therefore state the objective of the market maker as below. If the market maker trades from time t to T , while observing the state \mathbf{S}_t and employing the optimal policy $u^*(t)$, we have the objective:

$$(2.1) \quad J^{(u^*)}(t, \mathbf{S}_t) = \sup_{u \in \mathcal{U}} \mathbb{E} \left[\int_t^T -\eta Y_t^2 dt + X_T + Y_T P_T^{(mid)} - \kappa Y_T^2 + \sum_{t \leq \tau_i \leq T} K(\mathbf{S}(\tau_i), \psi_i) \right]$$

where

$$K(\mathbf{S}, \psi) = \begin{cases} 0 & \text{for limit/cancel orders} \\ z p_t^{(\zeta)} & \text{for market orders} \end{cases}$$

The system state \mathbf{S}_t and policy $u(t) \in \mathcal{U}$, where \mathcal{U} denotes the set of admissible policies, at time t is defined as

$$(2.2) \quad u(t) := \{(\tau_i, \psi_i)\}_{i=1,\dots,N} \text{ where } \tau_N < t$$

$$(2.3) \quad \mathbf{S}_t := \{X_t, Y_t, p_t^{(\zeta)}, q_t^{(\zeta)}, q_t^{(\zeta,D)}, n_t^{(\zeta)}, P_t^{(mid)}, (\lambda_t^{(i)})_{(i=1,\dots,d)}\}_{\zeta \in \{\text{a, b}\}}$$

Here, X_t is the cash of the market maker, Y_t is the inventory, $n_t^{(\zeta)}$ and $q_t^{(\zeta)}$ denote the queue-priority and size, $\lambda_t^{(i)}$ are the Hawkes intensities. These state variables follow jump equations due to the Hawkes process jumps. The dynamics of the LOB state-variable are given in the Appendix A. Accordingly, the policy is a set of impulse times τ_i and impulse types $\psi_i \in \mathcal{E}$. We assume that the order sizes of the Hawkes process and that of the agent are uniformly constant. Therefore we are only interested in the optimal impulse time and type. The question of optimal order size is out of scope for this work. Finally, η and κ are the respective running and terminal inventory penalty parameters while $K(\mathbf{S}, \psi)$ is the instantaneous profit made by the market maker by sending an impulse ψ at state \mathbf{S} .

2.3. The State-Intervention Operator. The market maker's control problem is of the impulse type: at discrete intervention times τ_i , the agent selects impulses ψ_i , which correspond to submitting market orders, limit orders, or cancellations. Unlike continuous control, where adjustments are infinitesimal and ongoing, impulses induce discrete jumps in the state variables, reflecting the event-driven structure of order books.

Formally, the intervention rule is represented by the state-intervention operator:

$$\mathbf{S}(\tau_i) = \Gamma(\mathbf{S}(\tau_i^-), \psi_i),$$

where $\mathbf{S}(\tau_i^-)$ is the pre-impulse state, ψ_i is the chosen control, and Γ encodes how that control affects the state vector. For example, inserting a limit order at the top of the book ($LO_T^{(\zeta)}$) increases both the queue size and possibly modifies the maker's queue position. The intervention can be specified as follows:

$$\begin{aligned} n^{(\zeta)}(\tau_i) &= n^{(\zeta)}(\tau_i^-) \mathbf{1}(n^{(\zeta)}(\tau_i^-) \leq q^{(\zeta)}(\tau_i^-)) + q^{(\zeta)}(\tau_i^-) \mathbf{1}(n^{(\zeta)}(\tau_i^-) > q^{(\zeta)}(\tau_i^-)) \\ q^{(\zeta)}(\tau_i) &= q^{(\zeta)}(\tau_i^-) + 1 \end{aligned}$$

Similarly, cancelling a top order ($CO_T^{(\zeta)}$) removes volume, possibly shifting prices if the best quote disappears. Market orders ($MO^{(\zeta)}$) consume liquidity, changing inventory, cash, and potentially mid-price. Each of these is specified by the difference equations in the Appendix B.

2.4. Hamilton–Jacobi–Bellman Quasi-Variational Inequality. One can see from the objective's formulation in Eq. 2.1 that the running cost is $f(S_t) = -\eta Y_t^2$ and the terminal cost is $g(S_T) = X_T + Y_T P_T^{(mid)} - \kappa Y_T^2$. Let $V(t, \mathbf{S}_t)$ denote the value function at time t in state \mathbf{S}_t . The problem admits the following quasi-variational inequality (QVI) see [42] for instance:

$$(2.4) \quad \min \left\{ -\partial_t V - \mathcal{L}V, V(t, \mathbf{S}) - \sup_{\psi \in \mathcal{A}} V(t, \Gamma(\mathbf{S}, \psi)) \right\} = 0,$$

with terminal condition

$$(2.5) \quad V(T, \mathbf{S}_T) = X_T + Y_T P_T^{(mid)} - \kappa Y_T^2$$

We see that the conditions and assumptions for this HJB-QVI's wellposedness and the solutions' existence have been met (Ch. 2 [42]). Here:

- The operator \mathcal{L} is the infinitesimal generator of the controlled Hawkes-driven LOB dynamics in the absence of impulses. It accounts for stochastic jumps due to market orders, cancellations, and limit order arrivals.
- The intervention operator $\sup_{\psi \in \mathcal{A}} V(t, \Gamma(\mathbf{S}, \psi))$ encodes the value of optimally choosing an impulse at state \mathbf{S} .
- The terminal payoff consists of the cash position plus the liquidation value of inventory (using the mid-price) with κ being the terminal liquidation penalty.

This QVI formulation unifies the continuous-time Markov jump dynamics of the order book with the discrete impulse controls of the market maker. Solving it yields the optimal market-making policy: when to post or cancel limit orders, when to submit market orders, and how to manage queue positions in response to the stochastic evolution of order flow.

Let $\Phi(t, \mathbf{S}_t) = \sup_u J^{(u)}(t, \mathbf{S}_t)$ be the candidate value function. The value-intervention operator is defined as:

$$(2.6) \quad \mathcal{M}\Phi(t, \mathbf{S}_t) = \sup_{\psi} \{\Phi(t, \Gamma(\mathbf{S}_t, \psi)) + K(\mathbf{S}_t, \psi)\}$$

Introducing a binary control $d_t \in \{0, 1\}$ (impulse or not), we can rewrite the HJB-QVI as the following, see [6] for instance:

$$(2.7) \quad \sup_{d \in \{0, 1\}} \{(1 - d)(\mathcal{L}\Phi + f) + d(\mathcal{M}\Phi - \Phi)\} = 0$$

2.5. The Generator. Now for pure jump processes, $\mathcal{L}\Phi$, the generator of Φ , is given by for Poisson Process driven processes:

$$(2.8) \quad \mathcal{L}\Phi(t, s) = \Phi_t(t, s) + \sum_i \lambda^{(i)} (\Phi(t, T_i(s)) - \Phi(t, s))$$

Where $T_i(\cdot)$ is transition function of state s when an event i happens in the point process. Due to the similarity with the state-intervention operator, we omit its mathematical formulation. If a Hawkes Process drives the point processes, we have the following SDE representation of the Hawkes intensity for exponential kernels:

$$(2.9) \quad d\lambda^{(i)}(t) = \gamma_i(\mu_i - \lambda^{(i)}(t))dt + \alpha_i dN_t^{(i)}; \lambda^{(i)}(0) = \mu_i$$

$$(2.10) \quad \lambda^{(i)}(t) = \mu_i + \alpha_i \int_0^t e^{-\gamma_i(t-s)} dN_s^{(i)}$$

For multidimensional, mutually-exciting Hawkes we have M^2 such equations which when added gives us the final intensity:

$$(2.11) \quad \lambda_t^{(i)} = \mu_i + \sum_{j=1}^M \alpha_{ij} \int_0^t e^{-\gamma_{ij}(t-u)} dN_u^{(j)}$$

We note that there is no SDE representation for general kernels of the Hawkes process. As shown in several studies including [27], the more realistic choice of kernels is the power-law function however we lose the Markovian representation of the intensity process. One can approximate these power-law functions by an infinite sum of exponential kernels as shown in [31] however this leads to an infinite dimensional Markov process. In this work, we will restrict ourselves to single exponential kernels unless specified.

Abusing some notation to denote the candidate function by $\Phi(t, \lambda, s)$ instead of $\Phi(t, S_t)$, the generator becomes, see for instance [11]:

$$(2.12) \quad \begin{aligned} \mathcal{L}\Phi(t, \lambda, s) = & \Phi_t(t, \lambda, s) + \sum_i \left(\lambda^{(i)}(t) (\Phi(t, \lambda + \alpha_i, T_i(s)) - \Phi(t, s)) \right. \\ & \left. + \Phi_{\lambda_i}(t, \lambda, s) \frac{d\lambda_i}{dt}(t) \right) \end{aligned}$$

2.6. Solution Approaches and Challenges. Solving the optimal market making problem under the proposed Hawkes-driven LOB dynamics presents considerable analytical and numerical challenges. First, obtaining a closed-form solution to the resulting quasi-variational inequality (QVI) (see for instance [38]) is generally infeasible due to the high dimensionality and nonlinearity of the system. The state space, which includes variables such as the MM's inventory, cash, active order queue positions, best bid/ask prices, spread, imbalance, time since last event, and a multi-dimensional Hawkes kernel history, is of several dimensions. This renders traditional grid-based numerical techniques (see for instance [6, 18]) impractical due to the curse of dimensionality. To overcome these obstacles, deep learning-based PDE solvers have been explored for HJBs in continuous stochastic control literature, see [26] for a recent review. Another interesting approach taken in [17] was to approximate the impulse control problem as a series of optimal stopping problem and solving it with a deep learning method.

In particular, we attempted to apply the Deep Galerkin Method (DGM) introduced by [40], which approximates solutions to high-dimensional PDEs using neural networks trained by minimizing the PDE residual over a sampled domain. Despite its success in solving parabolic PDEs arising in finance and physics, DGM struggled to converge in our setting. The primary bottleneck lies in the non-local and discontinuous structure of the impulse control operator in

the QVI, which poses significant challenges for sampling-based learning methods. Extensions such as those proposed by [3] to better handle discontinuities and boundary conditions were also tested, but their effectiveness was limited by the irregular jump structure induced by the impulse control. These challenges highlight the need for more specialized neural-PDE solvers that can handle the specific structure of control problems with non-local operators and event-driven dynamics. Hybrid methods that combine analytical insights from stochastic control with data-driven approximators remain a promising yet largely unexplored avenue in this setting. Nevertheless in the below we showcase the DGM methodology and the training results we achieved.

3. Deep Galerkin Method to solve the HJB-QVI. In order to circumvent the problems with traditional approaches to solve the HJB-QVI, we employ deep learning approximations to the value function and the control function(s). Our implementation uses the Deep Galerkin Method (DGM) of [40] to solve a jump Hamilton-Jacobi-Bellman (HJB) equation arising in optimal market making with limit order books. The DGM method is a mesh-free method of fitting a neural network to the value function of the optimal control problem. The value function, the decision policy, and the control policy respectively are represented by neural networks with parameters θ, χ, ξ respectively:

$$\begin{aligned}\phi(t, \mathbf{S}_t) &= \phi_\theta(t, \mathbf{S}_t) \\ d(t, \mathbf{S}_t) &= d_\chi(t, \mathbf{S}_t) \\ u(t, \mathbf{S}_t) &= u_\xi(t, \mathbf{S}_t)\end{aligned}$$

It fits this neural network by minimizing a loss function constructed from a candidate value function using the HJB-QVI. As the HJB-QVI involves several partial derivatives of the value function, the well known method of automatic differentiation is employed to compute them in a mesh-free way again. That is, time and lambda derivatives are computed using automatic differentiation:

$$\frac{\partial \phi_\theta}{\partial t}(t, s) = \nabla_t \phi_\theta(t, s)$$

The DGM method first involves sampling time and space points from given distributions both in the interior domain as well as the boundary of the state space. The sampling strategy is to generate states from stationary distributions for example:

$$\begin{aligned}Y &\sim \mathcal{N}(0, 4) \text{ (rounded to integers)} \\ P^{mid} &\sim \mathcal{N}(200, 100) \\ \text{Spreads} &\sim \text{Geometric}(0.8) \times 0.01\end{aligned}$$

The state space consists of 23 dimensions representing market microstructure variables. As discussed earlier, the value function $\phi(t, \mathbf{S}_t)$ satisfies the jump HJB-QVI equation:

$$(3.1) \quad \sup_{d \in \{0,1\}} \left\{ (1-d)\mathcal{L}\phi(t, \mathbf{S}_t) + d \sup_{\psi \in \mathcal{A}} \mathcal{M}^\psi \phi(t, \mathbf{S}_t) \right\} = 0$$

subject to terminal condition:

$$(3.2) \quad \phi(T, \mathbf{S}_T) = X_T + Y_T P_T^{(mid)} - \kappa Y_T^2$$

Next the value function and the HJB-QVI is evaluated at these points and the loss function is calculated. The DGM loss function combines interior and boundary losses:

$$(3.3) \quad \mathcal{L}_{DGM} = \mathcal{L}_{interior} + \mathcal{L}_{boundary}$$

Interior Loss:

$$(3.4) \quad \mathcal{L}_{interior} = \mathbb{E} \left[|\{(1 - d_\chi(t, \mathbf{S}_t))\mathcal{L}\phi_\theta(t, \mathbf{S}_t) + d_\chi(t, \mathbf{S}_t)\mathcal{M}^{u_\xi}\phi_\theta(t, \mathbf{S}_t)\}|^2 \right]$$

$$(3.5) \quad \mathcal{M}^{u_\xi}\Phi(t, \mathbf{S}_t) = \Phi(t, \Gamma(\mathbf{S}_t, u_\xi(t, \mathbf{S}_t))) + K(\mathbf{S}_t, u_\xi(t, \mathbf{S}_t))$$

Boundary Loss:

$$(3.6) \quad \mathcal{L}_{boundary} = \mathbb{E} \left[|\phi_\theta(T, \mathbf{S}_T) - (X_T + Y_T P_T^{(mid)} - \kappa Y_T^2)|^2 \right]$$

The objective for the value function is to minimise these costs however as noted in [3], the objective for the control functions is to maximise these costs. Therefore it becomes a minimax optimization problem similar to an actor-critic setup. We use the ADAM backpropagation algorithm for learning the weights of the three neural networks. We used the architecture mentioned in [40] i.e. an LSTM for all three networks (ϕ, d, u). We used the relu activation function and a fixed learning rate of 10^{-3} .

3.1. Results: Training logs are shown in Figure 1 for two settings of the LOB setup - the Poisson process setting and the Hawkes process setting. In particular, we restrict ourselves to the Hawkes process setup where only Market Orders are Hawkes processes and the remaining events are Poisson processes. This is done to show a proof of concept of the Deep Galerkin Method in the Hawkes setting. We observe the convergence is oscillatory due to the opposing updates of value and policy function. The networks used were 3 layers deep and 20 neurons wide for the Poisson setting and 10 layers deep and 50 neurons wide for the Hawkes setting. These network architecture hyperparameters were established by a large scale grid search-like method of comparing losses achieved after 200 epochs of training. In the Hawkes setting we saw the d network was quickly converging to a local optima where it always decided not to act in the first few training epochs. Therefore we chose to freeze the weights of the d network for the first 50 epochs to enable learning in the ϕ and u networks. We note that in the full Hawkes LOB setting (i.e. all 12 events form a mutually exciting Hawkes Process), the model did not converge. This is primarily due to the extremely high dimensional nature of the problem since 12 mutually exciting Hawkes process require a 144 dimensional Markovian representation for the stochastic control framework.

We test the policy networks on OOS data by running them through our simulator for 5 minutes over several episodes. We calculate the annualized sharpe ratio from these simulations and the mean absolute inventory over the trajectories. The results are summarised in Table

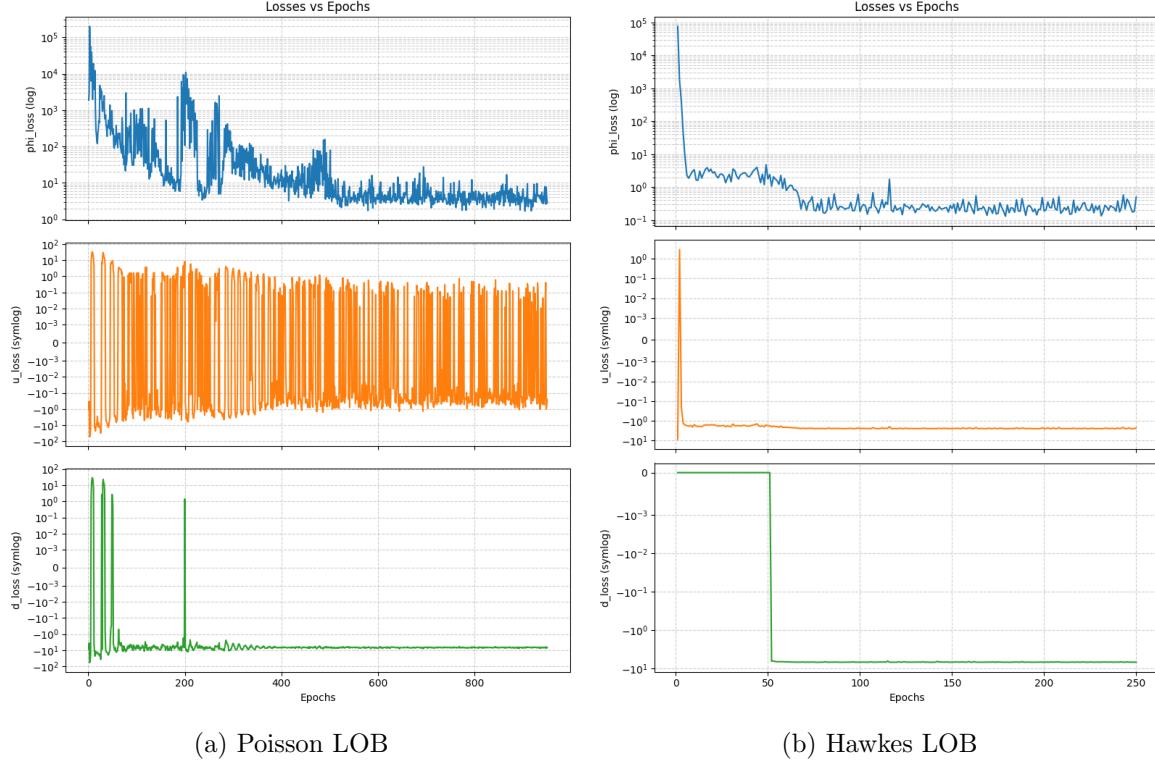


Figure 1: DGM Training Logs

Table 1: Out-of-sample Testing Results

LOB Model	Sharpe Ratio	Mean Abs. Inventory
Poisson	4.54	0.891
Hawkes - Market Orders Only	0.78	21.56
Full 12D Hawkes	Did Not Converge	Did Not Converge

1. We observe positive annualized sharpe ratios in both the Poisson and Hawkes setting however we note that the Hawkes setting has a very high mean absolute inventory. We further investigate this in Figure 2. We observe that over multiple tests (light lines) the strategy of the agent remains exactly the same. Since this seems to rely on pumping the market by sending aggressive orders initially and liquidating the inventory on a slow scale, we name this strategy ‘pump and dump’. It seems to be suboptimal as we see the profit to be near zero with a quite high variance. It is interesting that this obviously illegal strategy has been learnt by solving the HJB-QVI. Indeed it is mathematically allowed to have this strategy but it is neither market making nor legal by regulations. We note that in the setting of Hawkes with Market Orders being the only self-exciting events, with exponential kernels there is a dynamic

arbitrage as reported by [4]. This dynamic arbitrage theoretically allows for these ‘pump and dump’ strategies to be profitable on average. This could be the reason why the HJB-QVI solver converged to this strategy in this case.

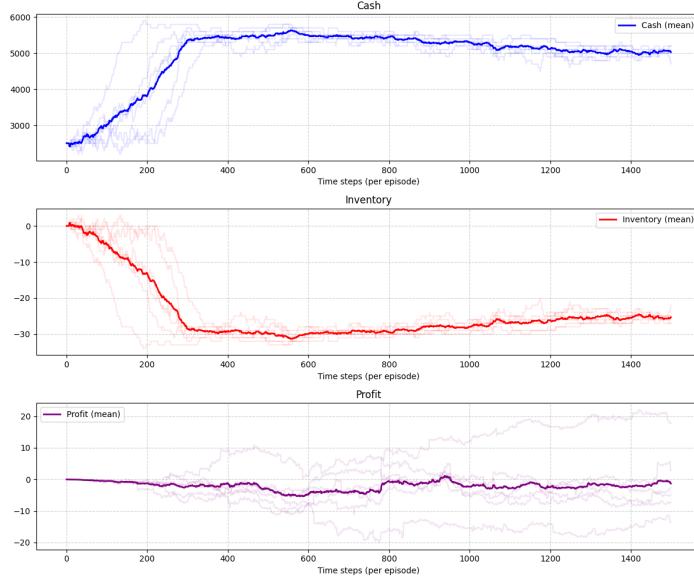


Figure 2: DGM Testing: Pump and Dump strategy of the Market Orders only Hawkes setting

It has been shown in [32] that in this Market Orders only Hawkes setting, no-arbitrage conditions imply power law kernels of the Hawkes process. This choice of power law kernels has also been widely supported in the literature, for instance see [27] for a discussion on the same. We however note that power law kernels as it is do not allow for a Markovian representation of the intensity process. We therefore were not able to formulate the intensity process’s dynamics in a SDE representation. This poses a significant challenge. [30] show that the power law kernel can be approximated as a sum of infinite number of exponential kernels. Unfortunately this does not solve our numerical problem however it does give us an approximation technique. We can indeed approximate the power law kernel to a reasonable extent by finite number of exponential kernels however this has the same curse of dimensionality as the full 12D Hawkes setting mentioned before.

These challenges motivate the use of a model free methodology instead of the model based stochastic control methodology we employed before. In the next section, we develop a reinforcement learning framework to learn the optimal strategy from the principles of policy gradient instead of trying to solve the value function’s HJB-QVI.

4. Reinforcement Learning Approximation. The analytical intractability of the HJB-QVI derived in Section 3 necessitates alternative solution methodologies. As established in Section 3.5, the generator \mathcal{L} encompasses a high-dimensional state space including queue sizes, queue priorities, Hawkes intensities, and the complete bid-ask structure across multiple price levels. The presence of impulse controls introduces non-local operators into the partial differ-

ential equation, rendering standard finite-difference schemes or backward-induction methods computationally prohibitive even under aggressive discretization. The curse of dimensionality, particularly acute in the 23-dimensional state space of our full Hawkes-driven LOB model, fundamentally limits the applicability of grid-based numerical techniques as discussed in [6].

To circumvent these computational barriers, we adopt a model-free reinforcement learning approach that learns optimal policies through simulated interaction with the Hawkes-driven limit order book environment, thereby avoiding the need to solve the QVI directly. This methodological pivot aligns with recent developments in the algorithmic trading literature reviewed in Section 1, where reinforcement learning has emerged as a compelling alternative.

4.1. Decomposition of the Impulse Control Problem. Drawing inspiration from the auxiliary control formulations explored in [36], we approximate the impulse control problem by decomposing it into two interacting reinforcement learning agents. This decomposition directly mirrors the structure of the QVI presented in Section 3.4, which inherently distinguishes between the continuation region (no impulse) and the intervention region (apply an impulse).

Specifically, we introduce:

- **Decision network d_χ (timing):** A policy that determines *when* to intervene in the limit order book. This network approximates the binary control $d_t \in \{0, 1\}$ introduced in the reformulated HJB-QVI of Section 3.4, thereby capturing the stopping aspect of the impulse control problem. The decision network effectively learns to identify states where the value of intervention exceeds the value of continuation.
- **Action network u_ξ (impulse selection):** A policy that determines *what* order action to execute, conditional on the decision to intervene. This network approximates the impulse selection operator $\sup_{\psi \in \mathcal{A}} V(t, \Gamma(\mathbf{S}, \psi))$ from the QVI formulation, where ψ_i corresponds to submitting market orders, limit orders, or cancellations as enumerated in the event set \mathcal{E} defined in Section 3.1.

This architectural separation preserves the natural structure of the quasi variational inequality while enabling gradient-based learning through policy optimization. Unlike the Deep Galerkin Method discussed in Section 3.6, which attempts to directly parameterize the value function $V(t, \mathbf{S}_t)$ and solve the PDE residual, the reinforcement learning formulation learns the optimal policy implicitly through interaction with the environment, thereby bypassing the non-local operator evaluation challenges that hindered DGM convergence in high-dimensional settings.

4.2. State Space Construction. The state representation for the reinforcement learning agent must capture both the instantaneous market conditions relevant for execution risk and the historical information necessary for predicting future order flow dynamics. Recognizing that the Hawkes process intensities $\lambda_t^{(i)}$ provide a sufficient statistic for the infinitesimal arrival rates conditional on the filtration generated by past events (as established in Section 3.6), we construct the augmented state space as:

$$(4.1) \quad \mathbf{S}_t := \left\{ X_t, Y_t, s_t, \frac{n_t^{(\zeta)}}{q_t^{(\zeta)}}, \zeta \in \{a, b\}, \lambda_t^{(i)}, \mathcal{H}_{t-\tau_H:t} \right\}$$

The components of this state vector warrant detailed justification:

- $X_t \in \mathbb{R}$: The market maker's cash position, which evolves according to the jump equations specified in Appendix A.
- $Y_t \in \mathbb{Z}$: The inventory position, which appears in both the running cost $f(S_t) = -\eta Y_t^2$ and the terminal liquidation value.
- $s_t := p_t^{(a)} - p_t^{(b)} \in \mathbb{R}_+$: The bid-ask spread, which determines the maximum profit capturable per round-trip transaction and influences both fill probabilities and the mid-price evolution $P_t^{(mid)}$ that enters the terminal condition.
- $\frac{n_t^{(\zeta)}}{q_t^{(\zeta)}} \in [0, 1]$ for $\zeta \in \{a, b\}$: The relative queue position of the market maker's limit orders on each side of the book. As detailed in Section 3.1, the queue dynamics $(n_t^{(\zeta)}, q_t^{(\zeta)}, q_t^{(\zeta, D)})$ govern the fill probabilities and therefore critically impact the expected profit from posted limit orders. The normalization by total queue size $q_t^{(\zeta)}$ ensures scale invariance across different liquidity regimes.
- $\lambda_t^{(i)} \in \mathbb{R}_+$ for $i = 1, \dots, d$: The Hawkes process intensities corresponding to each event type in \mathcal{E} .
- $\mathcal{H}_{t-\tau_H:t}$: A truncated window of the recent event history spanning the interval $[t - \tau_H, t]$. While the Hawkes intensities provide a sufficient statistic for the infinitesimal conditional intensity, finite memory windows can capture additional predictive information about market microstructure regimes not fully summarized by the exponential kernel parameterization. This augmentation addresses the limitation noted in Section 3.6 regarding the restriction to exponential kernels, as more realistic power-law kernels do not admit finite-dimensional Markovian representations.

4.3. Objective Function and Reward Structure. To enable gradient-based policy learning, we discretize the continuous-time objective function presented in Equation 2.1 over a uniform grid $\{t_0, t_1, \dots, t_N\}$ with $t_0 = t$, $t_N = T$, and timestep $\Delta t = t_{i+1} - t_i$. The discrete-time approximation of the value function becomes:

$$(4.2) \quad J^{(u)}(t, \mathbf{S}_t) = \mathbb{E} \left[\sum_{i=0}^{N-1} \left(\int_{t_i}^{t_{i+1}} -\eta Y_s^2 ds + \Delta X_{t_i} + \Delta(Y_{t_i} P_{t_i}^{(mid)}) \right. \right. \\ \left. \left. + \mathbb{1}_{\{\tau_j \in (t_i, t_{i+1})\}} K(\mathbf{S}(\tau_j), \psi_j) \right) \right]$$

where $\Delta X_{t_i} := X_{t_{i+1}} - X_{t_i}$ represents the change in cash due to order fills during the interval $[t_i, t_{i+1}]$, and $\Delta(Y_{t_i} P_{t_i}^{(mid)})$ captures the marked-to-market change in inventory value. The indicator function $\mathbb{1}_{\{\tau_j \in (t_i, t_{i+1})\}}$ equals one if the agent chose to intervene at some impulse time τ_j within the interval, with the instantaneous cost/profit $K(\mathbf{S}(\tau_j), \psi_j)$ defined in Section 3.2.

The reward signal at each timestep thus aggregates:

1. The instantaneous inventory penalty $-\eta \int_{t_i}^{t_{i+1}} Y_s^2 ds \approx -\eta Y_{t_i}^2 \Delta t$, which discourages the accumulation of directional positions and aligns with the risk-aversion objective discussed by [13].

2. The incremental profit/loss from inventory changes and order executions, reflecting both the bid-ask spread capture mechanism emphasized in Section 1 and the adverse selection risks inherent in providing liquidity.
3. The intervention cost $K(\mathbf{S}(\tau_j), \psi_j)$, which for limit and cancel orders equals zero, while for market orders equals the immediate liquidity consumption cost. This asymmetry naturally discourages excessive aggressive trading while permitting strategic use of market orders for inventory management.

4.4. Training Methodology.

4.4.1. Proximal Policy Optimization. We implement the training procedure using Proximal Policy Optimization (PPO) [39], selected for its empirical robustness to noisy gradient estimates and its ability to handle the continuous-time approximation inherent in our episodic simulation framework. PPO optimizes policies through a clipped surrogate objective that constrains policy updates to remain within a trust region, thereby mitigating the variance and instability that plague vanilla policy gradient methods in high-dimensional action spaces.

The PPO objective for the decision network d_χ and action network u_ξ takes the form:

$$(4.3) \quad \mathcal{L}^{PPO}(\chi, \xi) = \mathbb{E}_{\tau \sim \pi_{\chi, \xi}} \left[\min \left(r_t(\chi, \xi) \hat{A}_t, \text{clip}(r_t(\chi, \xi), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$$

where $r_t(\chi, \xi) := \frac{\pi_{\chi, \xi}(a_t | s_t)}{\pi_{\chi_{old}, \xi_{old}}(a_t | s_t)}$ denotes the probability ratio between the current and previous policies, \hat{A}_t is an estimate of the advantage function, and ϵ (typically set to 0.2) controls the size of the trust region. The clipping operation ensures that the policy does not change too drastically in a single update, which is particularly important given the instability observed in Section 3.6 when training the DGM-based value and policy networks with opposing gradient updates.

4.4.2. Simulation Environment. The training environment is instantiated using the high-fidelity Hawkes-driven limit order book simulator developed in [27], which reproduces the stylized facts of LOB dynamics discussed in Section 3.1, including realistic spreads, long-memory in returns, and clustered arrival times. Each training episode spans a trading horizon of $T = 300$ seconds for computational feasibility of training over several episodes.

The agent's decision frequency is discretized at intervals of $\Delta\tau = 0.1$ seconds, at which points both the decision network d_χ evaluates whether to intervene and, conditional on intervention, the action network u_ξ selects the specific order type to execute. This discretization acknowledges the realistic constraint, emphasized in Section 1, that market makers cannot update strategies at every LOB event due to latency constraints and the discrete nature of order book updates. The chosen frequency represents a conservative estimate of achievable reaction times in modern electronic trading systems, falling well within the impulse control framework's applicability regime while remaining computationally tractable for training.

4.4.3. Action Space Restriction. To reduce the complexity of the learning problem and focus on the fundamental market making mechanisms, we restrict the action space \mathcal{A} to a subset of the full event set \mathcal{E} defined in Section 3.1:

$$(4.4) \quad \mathcal{A}_{\text{restricted}} := \{LO_T^{(a)}, LO_T^{(b)}, CO_T^{(a)}, CO_T^{(b)}\}$$

This restriction includes:

- $LO_T^{(\zeta)}$ for $\zeta \in \{a, b\}$: Placement of limit orders at the top of the book on either the ask or bid side.
- $CO_T^{(\zeta)}$ for $\zeta \in \{a, b\}$: Cancellation of the agent’s existing limit orders at the top of the book.

Market orders ($MO^{(\zeta)}$) are explicitly excluded from $\mathcal{A}_{\text{restricted}}$ at this stage to focus policy learning on queue management and symmetric liquidity provision rather than aggressive liquidity taking. This design choice contrasts with the behavior observed in Section 3.6, where the DGM-trained agent converged to a “pump and dump” strategy heavily reliant on market order submission.

The restriction to top-of-book orders reflects the empirical reality that, for large-tick assets, the vast majority of trading volume occurs at the best bid and offer. Orders placed deeper in the book face significantly lower fill probabilities while providing limited additional option value, particularly given the inventory constraints typical of high-frequency market makers.

4.4.4. Transaction Costs. To ensure that learned policies remain profitable under realistic trading conditions, we impose a transaction fee structure on all inventory liquidations. Specifically, when computing the reward, we apply a transaction cost of 1 basis point (0.01%) on the absolute value of terminal inventory:

$$(4.5) \quad \hat{r}(S_t) = r(S_t) - 0.0001 \cdot |Y_T| \cdot P_T^{(\text{mid})}$$

This fee structure serves multiple purposes. First, it prevents the emergence of unrealistic strategies that rely on costless round-trip transactions to generate artificial profits, addressing a common criticism of simulation-based reinforcement learning in finance noted by [24]. Second, it incentivizes the agent to minimize unnecessary inventory turnover, thereby encouraging stable quoting behavior rather than rapid order cycling. Third, it approximates the exchange fees and market impact costs that real market makers face, as discussed in the context of Designated Market Makers and proprietary trading firms in Section 1.

The magnitude of 1 basis point reflects a conservative estimate of combined exchange fees, clearing costs, and micro-scale market impact for a medium-liquidity equity. While actual fee structures vary significantly across venues and participant types—with some designated market makers receiving rebates for providing liquidity—our choice ensures that any profitable strategy identified by the RL agent would remain viable under realistic cost assumptions.

4.5. Self-Imitation Learning. Standard policy gradient methods, including PPO, suffer from high variance and poor sample efficiency when applied to financial trading problems. This inefficiency stems from the sparsity of high-reward trajectories in the policy distribution,

particularly during early training phases when the agent has not yet discovered profitable trading patterns. The problem is exacerbated in our setting by the complexity of the Hawkes-driven dynamics and the high dimensionality of the state space, which create a vast exploration space with sparse reward signals.

To address these challenges, we augment the PPO training procedure with self-imitation learning (SIL) following [37]. The key insight underlying SIL is that the agent should explicitly imitate its own past trajectories that achieved returns exceeding its current value function estimate, thereby accelerating the propagation of successful behaviors throughout the policy.

4.5.1. Mechanism. The SIL augmentation modifies the PPO loss function by adding a cross-entropy term that encourages the current policy to replicate actions from a replay buffer of high-performing trajectories:

$$(4.6) \quad \mathcal{L}^{SIL}(\chi, \xi) = -\mathbb{E}_{(s_t, a_t, R_t) \sim \mathcal{B}_{good}} \left[\mathbf{1}_{\{R_t > V_\theta(s_t)\}} \log \pi_{\chi, \xi}(a_t | s_t) \right]$$

where \mathcal{B}_{good} denotes the replay buffer containing past experiences, R_t is the empirical return-to-go from state s_t , and $V_\theta(s_t)$ is the current value function estimate. The indicator function $\mathbf{1}_{\{R_t > V_\theta(s_t)\}}$ ensures that only trajectories exceeding current expectations contribute to the imitation loss.

The combined training objective becomes:

$$(4.7) \quad \mathcal{L}^{total}(\chi, \xi, \theta) = \mathcal{L}^{PPO}(\chi, \xi) + \beta_{SIL} \mathcal{L}^{SIL}(\chi, \xi) - \beta_{entropy} \mathcal{H}(\pi_{\chi, \xi})$$

where β_{SIL} controls the strength of the self-imitation signal, $\mathcal{H}(\pi_{\chi, \xi})$ is the policy entropy that encourages exploration, and $\beta_{entropy}$ balances exploration with exploitation. The hyperparameters ($\beta_{SIL}, \beta_{entropy}$) are tuned to ensure that self-imitation does not prematurely collapse the policy distribution before sufficient exploration has occurred.

4.5.2. Benefits in the Market Making Context. The integration of self-imitation learning provides several critical advantages in our impulse control setting:

1. **Avoidance of catastrophic forgetting:** By explicitly maintaining and imitating past successes, SIL ensures that profitable patterns, once discovered, remain accessible to the policy.
2. **Accelerated convergence:** The sparsity of highly profitable trajectories means that vanilla policy gradients provide weak learning signals, particularly during early training when the agent's exploration is largely undirected. SIL effectively amplifies the learning signal from rare successful experiences by repeatedly reinforcing the actions that led to those outcomes, thereby accelerating the convergence to profitable strategies.
3. **Variance reduction:** By focusing policy updates on trajectories with high returns rather than the full distribution of explored behaviors, SIL reduces the variance of gradient estimates.

The combination of PPO's trust-region optimization with SIL's experience replay creates a training framework that is simultaneously stable, sample-efficient, and capable of discovering complex temporal strategies in high-dimensional state spaces—properties essential for

learning optimal impulse control policies in the realistic, jump-driven market microstructure environment established in Section 3.1.

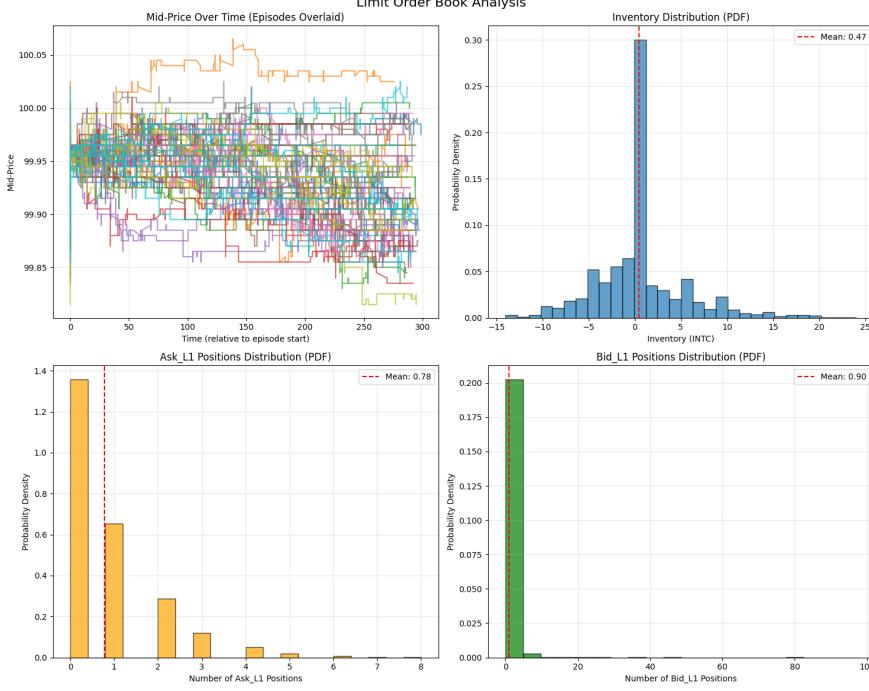


Figure 3: Policy Statistics for $\eta = 10$

4.6. Results. With an initial cash balance of \$2000, inventory penalty $\eta = 10$, and trading horizon $T = 300$ seconds, the trained agent achieves an annualized Sharpe ratio of 31.54 within only 60 training episodes. Despite the high dimensionality of the state space and the non-locality of the jump-PDE, the RL formulation converges rapidly and consistently to a policy that resembles genuine market making: the agent symmetrically provides liquidity at the top of the book and dynamically balances inventory. To benchmark the RL approximation, we compare it against the Deep Galerkin Method (DGM) implementation of the QVI from Section 3. As discussed earlier, the DGM training logs (Figure 1) display oscillatory convergence, reflecting the opposing updates of the value and policy networks. In contrast to the pump and dump strategy learned by the DGM method, RL-based approximation avoids such degenerate equilibria:

- The PPO agent with self-imitation learning converges within ~ 60 episodes, while the DGM-based approach requires orders of magnitude more iterations without escaping local minima.
- The RL agent learns symmetric liquidity provision strategies rather than exploiting transient price impact.
- Sharpe ratios under the RL approach (31.54 annualized) are both significantly higher and far more stable than the noisy near-zero returns observed under the DGM-trained

agent.

Overall, the comparison suggests that while the QVI formulation admits mathematically feasible but economically unrealistic solutions, the reinforcement learning approximation with an explicit inventory penalty guides the agent towards economically meaningful and stable market making strategies.

4.7. Sensitivity and Ablation Study. We perform a hyperparameter sensitivity and state ablation study to understand some reasons behind the good performance of the RL agent. Indeed such results of a high sharpe strategy require rigorous assumption and realism checks to be useful to drive production insights.

Table 2: Sensitivity Analysis of Model Performance Across Parameters and Kernel Types

Parameter	Value	Kernel Type	OOS Sharpe
Standard		Exponential	31.54
		Power-Law	28.81
Self-Imitation	Off		Pump & Dump
	On	Exponential	31.54
Probabilistic Agent	Section 4.7.1	Exponential	7.73
		Power-Law	20.12
Inventory Penalty η	0.1		Pump & Dump
	1.0		Pump & Dump
	10	Exponential	31.54
	100		21.32
Transaction Costs	1bps		31.54
	2bps		3.02
	4bps	Exponential	-5.67
	8bps		-19.48
	1bps		28.81
	2bps		2.40
	4bps	Power-Law	-25.22
	8bps		-107.08

To this end, we systematically vary key model components and environment parameters to assess the robustness and interpretability of the learned policy. The sensitivity analysis focuses on hyperparameters influencing the temporal structure of the Hawkes process, the regularization of inventory risk, and the impact of transaction costs, under both Exponential and Power-Law kernel specifications. Complementarily, the ablation study evaluates the informational contribution of each state variable by selectively removing components of the agent’s observation space. Together, these experiments aim to identify which design elements—both architectural and environmental—are essential for stable learning, realistic

execution behaviour, and sustained profitability out-of-sample.

4.7.1. Probabilistic Agent Baseline. The Probabilistic Agent extends a standard limit order book (LOB) trading agent by incorporating probabilistic reasoning about the timing and direction of the next market order (MO) event, as inferred from the estimated Hawkes intensities.

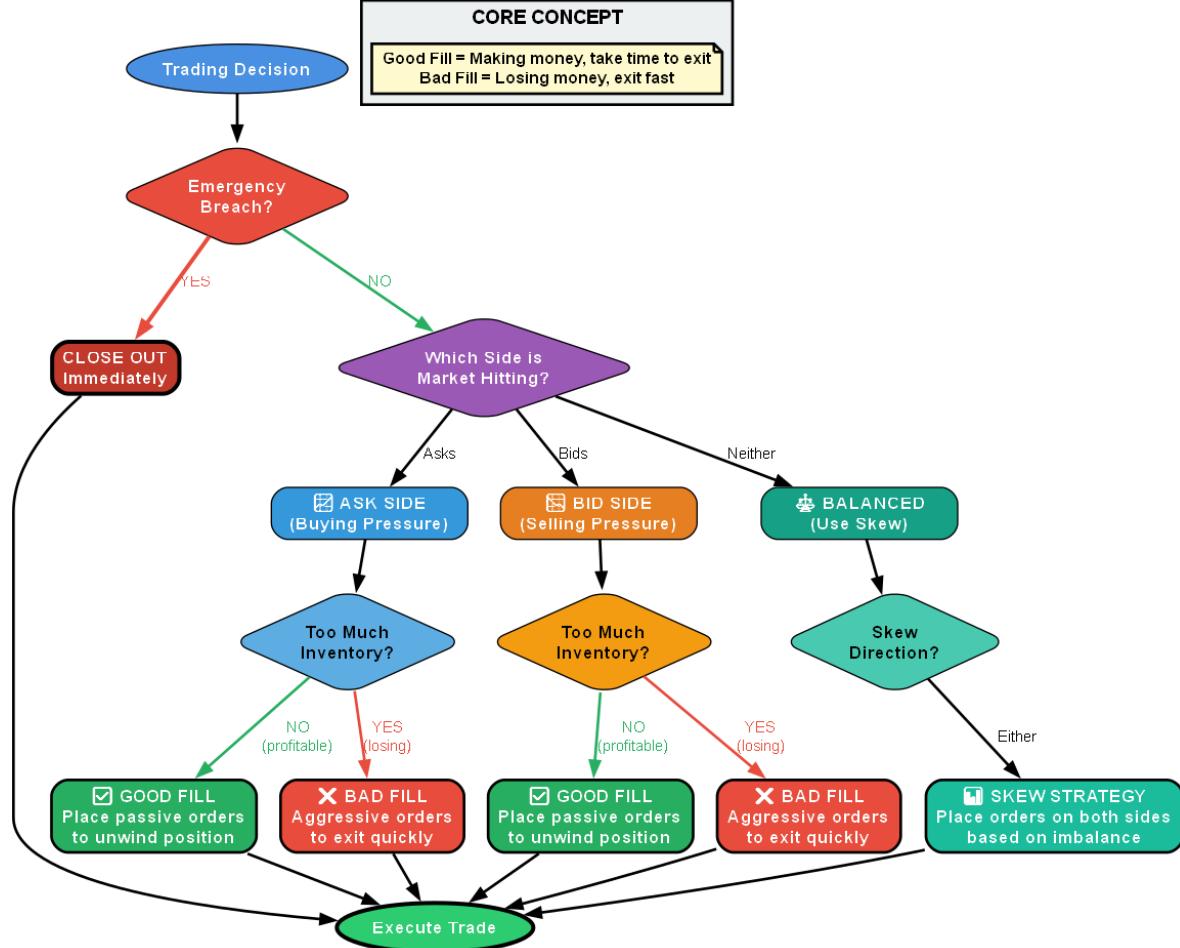


Figure 4: Baseline - Probabilistic Agent

At each decision time, the agent normalizes the Hawkes intensities to form a probability distribution over event types, thus estimating which side of the book is most likely to be hit next. The decision logic, as summarised in Fig. 4, then adjusts its quoting or market-taking behaviour accordingly. For instance, if the next most probable event is a market buy (MOBid), the agent anticipates upward price pressure and may reduce short inventory, place liquidity on the bid side, or cancel vulnerable asks. Conversely, if a market sell (MOAsk) is likely, the agent increases bid-side exposure or unwinds long positions.

Inventory management is incorporated through threshold-based controls on inventory imbalance. When the absolute inventory exceeds a fixed threshold, the agent enforces corrective market orders to re-centre its position. The action selection process also accounts for posted quotes' skew—the relative imbalance between bid and ask quotes of the market maker—which influences whether the agent improves quotes within the spread or cancels existing orders at deeper levels.

4.7.2. Discussion. Table 2 indicate that the RL agent's performance is broadly robust but exhibits sensitivity to kernel structure, transaction costs, and inventory penalization. The Exponential kernel consistently outperforms the Power-Law, suggesting that short-memory excitation provides a more accurate representation of transient order flow dependencies, while the Power-Law kernel offers marginally greater stability under partial state information. Stable convergence further requires the inclusion of Self-Imitation, whose removal induces pathological “Pump & Dump” dynamics. Comparing to the Probabilistic Agent, the RL agent consistently outperforms this baseline.

Moderate inventory penalization ($\eta = 10$) achieves the best balance between stability and responsiveness, whereas excessive penalization suppresses profitability. The strategy remains viable up to transaction costs of 2bps but collapses beyond that threshold, underscoring its dependence on low-friction environments.

The ablation study in Table 3 highlights the essential role of intensity and relative position features, whose exclusion results in unprofitable or unstable policies. Overall, the most effective configuration combines Exponential Hawkes dynamics, moderate regularization, and comprehensive state representations that capture both queue positioning and order flow intensity.

Table 3: State Ablation Study and Kernel Types

Parameter Removed from State	Kernel Type	OOS Sharpe
None (Standard)	Exponential	31.54
	Power-Law	28.81
History $\mathcal{H}_{t-\tau_H:t}$	Exponential	Pump & Dump
	Power-Law	30.85
Intensity $\lambda_t^{(.)}$	Exponential	-20.23
	Power-Law	-51.21
Spread s_t	Exponential	Pump & Dump
	Power-Law	17.66
Relative Position $\frac{n_t^{(\zeta)}}{q_t^{(\zeta)}}_{\zeta \in \{a,b\}}$	Exponential	Pump & Dump
	Power-Law	Pump & Dump (unprofitable)

5. Conclusion. This work formulates optimal market making in a high-fidelity limit order book environment as an impulse control problem, departing from traditional continuous-time Brownian frameworks to explicitly capture microstructural realities including queue dynamics, clustered arrivals, and endogenous price impact through a mutually-exciting Hawkes process. The resulting Hamilton-Jacobi-Bellman Quasi-Variational Inequality poses substantial computational challenges due to high dimensionality and non-local operators, rendering classical finite-difference methods impractical. We explored two solution approaches: a Deep Galerkin Method inspired by [40] and a model-free reinforcement learning approximation based on Proximal Policy Optimization with self-imitation learning. While the DGM framework successfully converged in simplified settings—achieving positive Sharpe ratios of 4.54 in the Poisson case and 0.78 when only market orders exhibited Hawkes dynamics—it encountered fundamental limitations. In the market-orders-only Hawkes setting, the learned strategy converged to an economically unrealistic “pump and dump” behavior exploiting the dynamic arbitrage conditions identified by [4] under exponential kernels. Moreover, the DGM approach failed to converge entirely in the full 12-dimensional mutually-exciting Hawkes environment, highlighting the curse of dimensionality inherent to neural PDE solvers in non-local impulse control settings. In contrast, the reinforcement learning approximation, which decomposes the impulse control problem into timing and action subproblems via a two-network PPO architecture, demonstrated substantially superior performance. Within only 60 training episodes, the RL agent achieved an annualized Sharpe ratio of 31.54 while learning symmetric liquidity provision strategies consistent with genuine market making behavior. The integration of self-imitation learning proved critical in accelerating convergence and reducing variance by focusing policy updates on rare but profitable trajectories. The comparison between methods reveals a fundamental trade-off: while the HJB-QVI formulation admits mathematically feasible solutions, it lacks intrinsic mechanisms to exclude economically unrealistic equilibria such as manipulative strategies. The RL framework, augmented with explicit inventory penalties and transaction costs, naturally guides learning toward stable and interpretable policies. These findings underscore the promise of combining impulse control theory with modern deep reinforcement learning to address optimal execution problems in jump-driven microstructural markets, while also highlighting the need for specialized neural-PDE solvers capable of handling non-local operators and event-driven dynamics in high-dimensional settings.

The sensitivity study shows that the agent is robust to moderate inventory penalties but performance deteriorates sharply under high transaction costs or extreme penalization. The ablation study highlights that intensity and relative position features are critical for maintaining profitability and stability. Notably, the Exponential kernel often leads to profitable “Pump & Dump” patterns, whereas this is rarely observed under the Power-Law kernel, providing some evidence that the DGM’s convergence to such behaviour may be driven by the Exponential kernel specification. Finally, we use the Probabilistic Agent as an interpretable baseline to benchmark and contextualize the performance of the black-box RL strategy.

6. Future Work. While the current study demonstrates the effectiveness of deep reinforcement learning for high-dimensional Hawkes-driven market making, several avenues remain to advance both theoretical understanding and computational efficiency.

Stochastic maximum principle (SMP) [14] approaches could provide alternative analytical

characterizations of optimal impulse controls in jump-driven LOBs, potentially yielding closed-form or semi-analytical feedback policies. Forward-backward stochastic differential equation (FBSDE) techniques may enable scalable approximations of the HJB-QVI by decoupling forward state evolution from backward value propagation [26]. Stochastic partial differential equation (SPDE) [34] solvers could generalize neural-PDE methods to fully handle the non-local operators arising from mutually-exciting Hawkes dynamics. Delay differential equation [8] frameworks could capture the intrinsic lagged dependencies in the LOB and order flow while providing tractable approximations of temporal microstructure effects. Finally, Volterra control [2] methods appear promising to address the seemingly non-Markovian dynamics induced by Power-Law kernels, enabling control strategies that account for long-memory effects while maintaining computational tractability.

Disclaimer. Opinions and estimates constitute our judgement as of the date of this Material, are for informational purposes only and are subject to change without notice. This Material is not the product of J.P. Morgan’s Research Department and therefore, has not been prepared in accordance with legal requirements to promote the independence of research, including but not limited to, the prohibition on the dealing ahead of the dissemination of investment research. This Material is not intended as research, a recommendation, advice, offer or solicitation for the purchase or sale of any financial product or service, or to be used in any way for evaluating the merits of participating in any transaction. It is not a research report and is not intended as such. Past performance is not indicative of future results. Please consult your own advisors regarding legal, tax, accounting or any other aspects including suitability implications for your particular circumstances. J.P. Morgan disclaims any responsibility or liability whatsoever for the quality, accuracy or completeness of the information herein, and for any reliance on, or use of this material in any way. Important disclosures at: www.jpmorgan.com/disclosures

Acknowledgements. Konark Jain would like to acknowledge JP Morgan Chase & Co. for his PhD scholarship. We are thankful for the discussions and feedback received at the SIAM Financial Mathematics 2025 conference in Miami, USA. Finally, we are grateful to the anonymous reviewers for their constructive feedback.

REFERENCES

- [1] F. ABERGEL AND A. JEDIDI, *A Mathematical Approach to Order Book Modeling*, Mar. 2013, <http://arxiv.org/abs/1010.5136> (accessed 2024-02-06). Issue: arXiv:1010.5136 arXiv:1010.5136 [math, q-fin].
- [2] N. AGRAM AND B. ØKSENDAL, *Malliavin calculus and optimal control of stochastic volterra equations*, Journal of Optimization Theory and Applications, 167 (2015), pp. 1070–1094.
- [3] A. AL-ARADI, A. CORREIA, G. JARDIM, D. DE FREITAS NAIFF, AND Y. SAPORITO, *Extensions of the deep Galerkin method*, Applied Mathematics and Computation, 430 (2022), p. 127287, <https://doi.org/10.1016/j.amc.2022.127287>, <https://www.sciencedirect.com/science/article/pii/S0096300322003617> (accessed 2025-02-19).
- [4] A. ALFONSI AND P. BLANC, *Extension and calibration of a Hawkes-based optimal execution model*, June 2015, <https://doi.org/10.48550/arXiv.1506.08740>, <http://arxiv.org/abs/1506.08740> (accessed 2025-10-15). arXiv:1506.08740 [q-fin].

- [5] M. AVELLANEDA AND S. STOIKOV, *High-frequency trading in a limit order book*, Quantitative Finance, 8 (2008), pp. 217–224, <https://doi.org/10.1080/14697680701381228>, <https://doi.org/10.1080/14697680701381228> (accessed 2024-10-08). Publisher: Routledge eprint: <https://doi.org/10.1080/14697680701381228>.
- [6] P. AZIMZADEH, *Impulse Control in Finance: Numerical Methods and Viscosity Solutions*, Feb. 2018, <https://doi.org/10.48550/arXiv.1712.01647>, <http://arxiv.org/abs/1712.01647> (accessed 2025-02-19). arXiv:1712.01647 [math].
- [7] R. AÏD, M. BASEI, G. CALLEGARO, L. CAMPI, AND T. VARGIOLU, *Nonzero-Sum Stochastic Differential Games with Impulse Controls: A Verification Theorem with Applications*, Mathematics of Operations Research, 45 (2020), pp. 205–232, <https://doi.org/10.1287/moor.2019.0989>, <https://pubsonline.informs.org/doi/abs/10.1287/moor.2019.0989> (accessed 2025-02-10). Publisher: INFORMS.
- [8] B. BALACHANDRAN, T. KALMÁR-NAGY, AND D. E. GILSINN, *Delay differential equations*, Springer, 2009.
- [9] E. BANDINI AND C. KELLER, *Non-local Hamilton-Jacobi-Bellman equations for the stochastic optimal control of path-dependent piecewise deterministic processes*, Aug. 2024, <https://doi.org/10.48550/arXiv.2408.02147>, <http://arxiv.org/abs/2408.02147> (accessed 2025-04-23). arXiv:2408.02147 [math] version: 1.
- [10] E. BAYRAKTAR, T. EMMERLING, AND J.-L. MENALDI, *On the Impulse Control of Jump Diffusions*, SIAM Journal on Control and Optimization, 51 (2013), pp. 2612–2637, <https://doi.org/10.1137/120863836>, <https://epubs.siam.org/doi/10.1137/120863836> (accessed 2025-02-19). Publisher: Society for Industrial and Applied Mathematics.
- [11] A. BENOUESSAN AND B. CHEVALIER-ROIGNANT, *Stochastic control for diffusions with self-exciting jumps: An overview*, Mathematical Control and Related Fields, 14 (2024), pp. 1452–1476, <https://doi.org/10.3934/mcrf.2024038>, <https://www.aims sciences.org/en/article/doi/10.3934/mcrf.2024038> (accessed 2025-02-03). Publisher: Mathematical Control and Related Fields.
- [12] N. BÄUERLE AND U. RIEDER, *Markov Decision Processes with Applications to Finance*, Universitext, Springer, Berlin, Heidelberg, 2011, <https://doi.org/10.1007/978-3-642-18324-9>, <https://link.springer.com/10.1007/978-3-642-18324-9> (accessed 2024-08-09).
- [13] Á. CARTEA, S. JAUMUNGAL, AND J. PENALVA, *Algorithmic and high-frequency trading*, Cambridge University Press, 2015.
- [14] M. CHAHIM, R. F. HARTL, AND P. M. KORT, *A tutorial on the deterministic Impulse Control Maximum Principle: Necessary and sufficient optimality conditions*, European Journal of Operational Research, 219 (2012), pp. 18–26, <https://doi.org/10.1016/j.ejor.2011.12.035>, <https://www.sciencedirect.com/science/article/pii/S0377221711011295>.
- [15] F. CHEN, N. MARTIN, P.-Y. CHEN, X. WANG, Z. REN, AND F. BUET-GOLFOUSE, *Deciding Bank Interest Rates – A Major-Minor Impulse Control Mean-Field Game Perspective*, Jan. 2025, <https://doi.org/10.48550/arXiv.2411.14481>, <http://arxiv.org/abs/2411.14481> (accessed 2025-02-10). arXiv:2411.14481 [math].
- [16] Y.-S. A. CHEN AND X. GUO, *Impulse Control of Multidimensional Jump Diffusions in Finite Time Horizon*, SIAM Journal on Control and Optimization, 51 (2013), pp. 2638–2663, <https://doi.org/10.1137/110854205>, <https://epubs.siam.org/doi/10.1137/110854205> (accessed 2025-02-19). Publisher: Society for Industrial and Applied Mathematics.
- [17] E. CHEVALIER, Y. HAFSI, AND V. L. VATH, *Optimal Execution under Incomplete Information*, Nov. 2024, <https://doi.org/10.48550/arXiv.2411.04616>, <http://arxiv.org/abs/2411.04616> (accessed 2025-07-18). arXiv:2411.04616 [q-fin].
- [18] A. CLEYNEN AND B. D. SAPORTA, *Numerical method to solve impulse control problems for partially observed piecewise deterministic Markov processes*, July 2023, <https://doi.org/10.48550/arXiv.2112.09408>, <http://arxiv.org/abs/2112.09408> (accessed 2025-05-23). arXiv:2112.09408 [math].
- [19] M. H. A. DAVIS, X. GUO, AND G. WU, *Impulse Control of Multidimensional Jump Diffusions*, SIAM Journal on Control and Optimization, 48 (2010), pp. 5276–5293, <https://doi.org/10.1137/090780419>, <https://epubs.siam.org/doi/10.1137/090780419> (accessed 2025-02-19). Publisher: Society for Industrial and Applied Mathematics.
- [20] J. FERNANDEZ-TAPIA, O. GUÉANT, AND J.-M. LASRY, *Optimal Real-Time Bidding Strategies*, June 2016, <https://doi.org/10.48550/arXiv.1511.08409>, <http://arxiv.org/abs/1511.08409> (accessed 2025-04-21).

- arXiv:1511.08409 [math].
- [21] B. GAŠPEROV, S. BEGUŠIĆ, P. POSEDEL ŠIMOVIĆ, AND Z. KOSTANJČAR, *Reinforcement Learning Approaches to Optimal Market Making*, Mathematics, 9 (2021), p. 2689, <https://doi.org/10.3390/math9212689>, <https://www.mdpi.com/2227-7390/9/21/2689> (accessed 2024-09-23). Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.
- [22] B. GAŠPEROV AND Z. KOSTANJČAR, *Deep Reinforcement Learning for Market Making Under a Hawkes Process-Based Limit Order Book Model*, IEEE Control Systems Letters, 6 (2022), pp. 2485–2490, <https://doi.org/10.1109/LCSYS.2022.3166446>, https://ieeexplore.ieee.org/abstract/document/9754690?casa_token=ml8dsDpva4UAAAAA:qhSif821WLeYWJ6ti6Ggqs4IQIX6ggjvXxpbhK9mGEj-t3XZUsdnXFHt0plhgMy0V2oNV_zung (accessed 2024-09-23). Conference Name: IEEE Control Systems Letters.
- [23] F. GUILBAUD AND H. PHAM, *Optimal high-frequency trading with limit and market orders*, Quantitative Finance, 13 (2013), pp. 79–94, <https://doi.org/10.1080/14697688.2012.708779>, <http://www.tandfonline.com/doi/abs/10.1080/14697688.2012.708779> (accessed 2024-02-05). Number: 1.
- [24] O. GUÉANT, *The Financial Mathematics of Market Liquidity: From Optimal Execution to Market Making*, 2016.
- [25] T. HO AND H. R. STOLL, *Optimal dealer pricing under transactions and return uncertainty*, Journal of Financial Economics, 9 (1981), pp. 47–73, [https://doi.org/https://doi.org/10.1016/0304-405X\(81\)90020-9](https://doi.org/https://doi.org/10.1016/0304-405X(81)90020-9), <https://www.sciencedirect.com/science/article/pii/0304405X81900209>.
- [26] R. HU AND M. LAURIERE, *Recent developments in machine learning methods for stochastic control and games*, arXiv preprint arXiv:2303.10257, (2023).
- [27] K. JAIN, N. FIROOZYE, J. KOCHEMS, AND P. TRELEAVEN, *Limit Order Book dynamics and order size modelling using Compound Hawkes Process*, Finance Research Letters, 69 (2024), p. 106157, <https://www.sciencedirect.com/science/article/pii/S1544612324011863> (accessed 2025-02-20). Publisher: Elsevier.
- [28] K. JAIN, N. FIROOZYE, J. KOCHEMS, AND P. TRELEAVEN, *Limit order book simulations: A review*, SSRN Electronic Journal, (2024), <https://doi.org/10.2139/ssrn.4745587>.
- [29] K. JAIN, J.-F. MUZY, J. KOCHEMS, AND E. BACRY, *No Tick-Size Too Small: A General Method for Modelling Small Tick Limit Order Books*, Nov. 2024, <https://doi.org/10.48550/arXiv.2410.08744>, <http://arxiv.org/abs/2410.08744> (accessed 2025-06-26). arXiv:2410.08744 [q-fin].
- [30] P. JUSSELIN, *Optimal Market Making with Persistent Order Flow*, SIAM Journal on Financial Mathematics, 12 (2021), pp. 1150–1200, <https://doi.org/10.1137/20M1376054>, <https://epubs.siam.org/doi/10.1137/20M1376054> (accessed 2024-10-24).
- [31] P. JUSSELIN, *Optimal market making with persistent order flow*, SIAM Journal on Financial Mathematics, 12 (2021), pp. 1150–1200, <https://doi.org/10.1137/20M1376054>, <https://doi.org/10.1137/20M1376054>, <https://arxiv.org/abs/https://doi.org/10.1137/20M1376054>.
- [32] P. JUSSELIN AND M. ROSENBAUM, *No-arbitrage implies power-law market impact and rough volatility*, Mathematical Finance, 30 (2020), pp. 1309–1336, <https://doi.org/https://doi.org/10.1111/mafi.12254>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/mafi.12254>, <https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1111/mafi.12254>.
- [33] B. LAW AND F. VIENS, *Market making under a weakly consistent limit order book model*, High Frequency, 2 (2019), pp. 215–238, <https://doi.org/10.1002/hf2.10050>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/hf2.10050> (accessed 2024-08-28). eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hf2.10050>.
- [34] G. J. LORD, C. E. POWELL, AND T. SHARDLOW, *An introduction to computational stochastic PDEs*, vol. 50, Cambridge University Press, 2014.
- [35] S. LV AND J. XIONG, *Hybrid optimal impulse control*, Automatica, 140 (2022), p. 110233, <https://doi.org/10.1016/j.automatica.2022.110233>, <https://www.sciencedirect.com/science/article/pii/S0005109822000784> (accessed 2025-02-17).
- [36] D. MGUNI, A. SOOTLA, J. ZIOMEK, O. SLUMBERS, Z. DAI, K. SHAO, AND J. WANG, *Timing is Everything: Learning to Act Selectively with Costly Actions and Budgetary Constraints*, June 2023, <https://doi.org/10.48550/arXiv.2205.15953>, <http://arxiv.org/abs/2205.15953> (accessed 2025-02-07). arXiv:2205.15953 [cs].
- [37] J. OH, Y. GUO, S. SINGH, AND H. LEE, *Self-Imitation Learning*, in Proceedings of the 35th International

- Conference on Machine Learning, PMLR, July 2018, pp. 3878–3887, <https://proceedings.mlr.press/v80/oh18b.html> (accessed 2025-06-09). ISSN: 2640-3498.
- [38] J. RICCI, *Applied Stochastic Control in High Frequency and Algorithmic Trading*, SSRN Electronic Journal, (2014), <https://doi.org/10.2139/ssrn.2504061>, <http://www.ssrn.com/abstract=2504061> (accessed 2024-02-05).
- [39] J. SCHULMAN, F. WOLSKI, P. DHARIWAL, A. RADFORD, AND O. KLIMOV, *Proximal Policy Optimization Algorithms*, Aug. 2017, <https://doi.org/10.48550/arXiv.1707.06347>, <http://arxiv.org/abs/1707.06347> (accessed 2025-05-23). arXiv:1707.06347 [cs].
- [40] J. SIRIGNANO AND K. SPILIOPOULOS, *DGM: A deep learning algorithm for solving partial differential equations*, Journal of Computational Physics, 375 (2018), pp. 1339–1364, <https://doi.org/10.1016/j.jcp.2018.08.029>, <https://www.sciencedirect.com/science/article/pii/S0021999118305527> (accessed 2025-02-19).
- [41] T. SPOONER, J. FEARNLEY, R. SAVANI, AND A. KOUKORINIS, *Market Making via Reinforcement Learning*, Apr. 2018, <https://doi.org/10.48550/arXiv.1804.04216>, <http://arxiv.org/abs/1804.04216> (accessed 2024-08-28). arXiv:1804.04216 [cs, q-fin].
- [42] B. K. ØKSENDAL AND A. SULEM, *Applied stochastic control of jump diffusions*, Universitext, Springer, Berlin : New York, 2005.

Appendix A. Dynamics of State Variables:

$$(A.1) \quad \text{Policy} \equiv u(t) := \{(\tau_i, \psi_i)\}_{i=1,\dots,N} \text{ where } \tau_N < t$$

$$(A.2) \quad \text{State} \equiv \mathcal{S}_t := \{X_t, Y_t, p_t^{(\zeta)}, q_t^{(\zeta)}, q_t^{(\zeta,D)}, n_t^{(\zeta)}, P_t^{(mid)}\}_{\zeta \in \{a, b\}}$$

$$(A.3) \quad \text{Cash} \equiv X_t \text{ s.t. } dX_t = \sum_{\zeta} -z \mathbb{1}_{(\zeta)}(t) (P_t^{(mid)} + z p_t^{(\zeta)}) dN_t^{(MO_{\zeta})}$$

$$(A.4) \quad \text{Inventory} \equiv Y_t \text{ s.t. } dY_t = \sum_{\zeta} -z \mathbb{1}_{(\zeta)}(t) dN_t^{(MO_{\zeta})}$$

$$(A.5) \quad \begin{aligned} \text{Best price at } \zeta \equiv p_t^{(\zeta)} \text{ s.t.} \quad dp_t^{(\zeta)} &= z(-dN_t^{IS_{\zeta}} \\ &+ \mathbb{1}(q_t^{(\zeta)} = 1)(dN_t^{(MO_{\zeta})} + dN_t^{(CO_{\zeta})})) \end{aligned}$$

$$(A.6) \quad \begin{aligned} \text{Best quote at } \zeta \equiv q_t^{(\zeta)} \text{ s.t.} \quad dq_t^{(\zeta)} &= (1 - q_t^{(\zeta)}) dN_t^{IS_{\zeta}} + dN_t^{(LO_{\zeta})} \\ &- (\mathbb{1}(q_t^{(\zeta)} > 1) + \mathbb{1}(q_t^{(\zeta)} = 1)(q_t^{(\zeta,D)} - q_t^{(\zeta)})) \\ &\times (dN_t^{(CO_{\zeta})} + dN_t^{(MO_{\zeta})}) \end{aligned}$$

$$(A.7) \quad \begin{aligned} \text{2nd quote at } \zeta \equiv q_t^{(\zeta,D)} \text{ s.t.} \quad dq_t^{(\zeta,D)} &= (q_t^{(\zeta,D)} - q_t^{(\zeta)}) dN_t^{IS_{\zeta}} \\ &+ dN_t^{(LO_{\zeta}^D)} + dN_t^{(CO_{\zeta}^D)} \end{aligned}$$

$$(A.8) \quad \text{Mid-Price} \equiv P_t^{(mid)} \text{ s.t.} \quad dP_t^{(mid)} = \frac{dp_t^{(a)} + dp_t^{(b)}}{2}$$

$$\begin{aligned}
\text{Queue Priority at } \zeta \equiv n_t^{(\zeta)} \text{ s.t.} \quad dn_t^{(\zeta)} = & dN_t^{IS_\zeta} - dN_t^{(MO_\zeta)} - \frac{n_t^{(\zeta)}}{q_t^{(\zeta)}} dN_t^{(CO_\zeta)} \\
& - \mathbb{1}(n_t^{(\zeta)} > q_t^{(\zeta)}) \frac{n_t^{(\zeta)} - q_t^{(\zeta)}}{q_t^{(\zeta,D)}} dN_t^{(CO_\zeta^D)} \\
& + \mathbb{1}(n_t^{(\zeta)} = 0) \tilde{n}_t^{(\zeta)} dN_t^{(MO_\zeta)}
\end{aligned} \tag{A.9}$$

Appendix B. State-Intervention Operator.

1. $LO_T^{(\zeta)}$

$$\begin{aligned}
n^{(\zeta)}(\tau_i) = & n^{(\zeta)}(\tau_i^-) \mathbb{1}(n^{(\zeta)}(\tau_i^-) \leq q^{(\zeta)}(\tau_i^-)) \\
& + q^{(\zeta)}(\tau_i^-) \mathbb{1}(n^{(\zeta)}(\tau_i^-) > q^{(\zeta)}(\tau_i^-))
\end{aligned} \tag{B.1}$$

$$q^{(\zeta)}(\tau_i) = q^{(\zeta)}(\tau_i^-) + 1 \tag{B.2}$$

2. $LO_D^{(\zeta)}$

$$\begin{aligned}
n^{(\zeta)}(\tau_i) = & n^{(\zeta)}(\tau_i^-) \mathbb{1}(n^{(\zeta)}(\tau_i^-) \leq q^{(\zeta)}(\tau_i^-) + q^{(\zeta,D)}(\tau_i^-)) \\
& + (q^{(\zeta)}(\tau_i^-) + q^{(\zeta,D)}(\tau_i^-)) \\
& \times \mathbb{1}(n^{(\zeta)}(\tau_i^-) > q^{(\zeta)}(\tau_i^-) + q^{(\zeta,D)}(\tau_i^-))
\end{aligned} \tag{B.3}$$

$$q^{(\zeta,D)}(\tau_i) = q^{(\zeta,D)}(\tau_i^-) + 1 \tag{B.4}$$

3. $LO_{IS}^{(\zeta)}$

$$n^{(\zeta)}(\tau_i) = 0 \tag{B.5}$$

$$q^{(\zeta)}(\tau_i) = 1 \tag{B.6}$$

$$q^{(\zeta,D)}(\tau_i) = q^{(\zeta)}(\tau_i^-) \tag{B.7}$$

$$p^{(\zeta)}(\tau_i) = p^{(\zeta)}(\tau_i^-) - z^{(\zeta)} \tag{B.8}$$

$$P^{mid}(\tau_i) = P^{mid}(\tau_i^-) - \frac{z^{(\zeta)}}{2} \tag{B.9}$$

4. $CO_T^{(\zeta)}$: only possible when there are some orders in the LOB that are the agent's.

$$n^{(\zeta)}(\tau_i) = \tilde{n}^{(\zeta)}(\tau_i^-) \tag{B.10}$$

$$\begin{aligned}
q^{(\zeta)}(\tau_i) = & (q^{(\zeta)}(\tau_i^-) - 1) \mathbb{1}(q^{(\zeta)}(\tau_i^-) > 1) \\
& + q^{(\zeta,D)}(\tau_i^-) \mathbb{1}(q^{(\zeta)}(\tau_i^-) = 1)
\end{aligned} \tag{B.11}$$

$$p^{(\zeta)}(\tau_i) = p^{(\zeta)}(\tau_i^-) + z^{(\zeta)} \mathbb{1}(q^{(\zeta)}(\tau_i^-) = 1) \tag{B.12}$$

$$P^{mid}(\tau_i) = P^{mid}(\tau_i^-) + \frac{z^{(\zeta)}}{2} \mathbb{1}(q^{(\zeta)}(\tau_i^-) = 1) \tag{B.13}$$

5. $MO^{(\zeta)}$: disabled if the top order on the resp. side is the agent's.

$$(B.14) \quad n^{(\zeta)}(\tau_i) = n^{(\zeta)}(\tau_i^-) - 1$$

$$q^{(\zeta)}(\tau_i) = (q^{(\zeta)}(\tau_i^-) - 1)\mathbb{1}(q^{(\zeta)}(\tau_i^-) > 1)$$

$$(B.15) \quad + q^{(\zeta,D)}(\tau_i^-)\mathbb{1}(q^{(\zeta)}(\tau_i^-) = 1)$$

$$(B.16) \quad p^{(\zeta)}(\tau_i) = p^{(\zeta)}(\tau_i^-) + z^{(\zeta)}\mathbb{1}(q^{(\zeta)}(\tau_i^-) = 1)$$

$$(B.17) \quad P^{mid}(\tau_i) = P^{mid}(\tau_i^-) + \frac{z^{(\zeta)}}{2}\mathbb{1}(q^{(\zeta)}(\tau_i^-) = 1)$$

$$(B.18) \quad X(\tau_i) = X(\tau_i^-) + z^{(\zeta)}p^{(\zeta)}(\tau_i^-)$$

$$(B.19) \quad Y(\tau_i) = Y(\tau_i^-) - z^{(\zeta)}$$