Reinforcement Learning in the Capital Markets

Edoardo Vittori

King's College London

10 March 2023

AGENDA

Introduction to Banks

- Introduction
- Capital Markets
- Wealth Management
- Order Execution

Algorithms in the Financial Markets

- Introduction
- Reinforcement Learning
- Use cases

Profit Centres of Banks

Introduction – Main services offered by banks and their technological focus

	Retail Bank	CIB	Private Banking/Wealth Management
Services	• Receive deposits • Offer loans	• Investment banking: M&A, ECM, DCM • Capital markets: sales & trading • Structured finance	• Mutual funds • Hedge funds • Private equity • Private banking
Revenue	• Difference between loan interest and deposit interest	• Advisory fees • Capital gains + margins • Interest rate	• % fee on AUM + performance fee
Technological focus	• Chatbots • Targeted ads for products • Metaverse?	• Analysing financial statements • Compiling slides • Automating traders? • Client segmentation	• Stock picking • Portfolio optimization • Analysing financial statements

Focus next (CIB)

Capital Markets

CIB | Capital Markets

	Market Making	Prop Trading	Corporate Derivatives Business
Scope	• Offering liquidity to the markets by continuously pricing assets. • It is important to continuously hedge	• Trading with the bank's capital. VaR limits. Intraday investments. • Buy low… sell high!	• Origination of derivatives for corporates. • Collaboration between sales, structuring, market making, XVAs and Financial Engineering
Technological focus	• Auto pricing • Auto hedging	• Returns prediction • Earnings prediction • Trading signals • Analytics	• Auto hedging • Analysing financial statements and transactions to forecast needs

Market Making: Offering liquidity to the markets

CIB | Capital Markets

Regulated market example

Last	Last Vol	Total Vol	Close	Daily Low	Daily High
4045.00	2	367267	4097.50	4033.50	4101.50

Implied

Bid		Offer	
Volume	Price	Price	Volume
136	4044.50	4045.00	62
327	4044.00	4045.50	293
348	4043.50	4046.00	427
620	4043.00	4046.50	426
358	4042.50	4047.00	463
330	4042.00	4047.50	348
325	4041.50	4048.00	327
318	4041.00	4048.50	294
305	4040.50	4049.00	281
512	4040.00	4049.50	288

Dealer market example - OTC

MARKIT ITRX EUR SNR FIN 06/26 | 92) Order Book | 91) RFS | 97) Settings

11:56:39 | 95) Buy | 96) Sell | BTFE | Filter By | All

PCS	Firm Name	CCP	Bid Spd	Ask Spd	BSz(MM)	ASz(MM)
CSDE	CREDIT SUISSE INTL	ICEE	54.6900	55.0100	50	50
CCGC	Citi CCGC	ICEE	54.7650	55.0350	50	50
GSMX	GS MINI	ICEE	54.7350	55.0350	15	15
JCTT	JP MORGAN	ICEE	54.7600	55.0400	100	100
BXCZ	Barclays Minis	ICEE	54.8400	55.0400	75	75
MSTI	MORGAN STANLEY MINI	ICEE	54.8000	55.0400	50	50
ABNP	BNP Paribas	ICEE	54.8000	55.0500	51	51
SGMI	SocGen Mini	ICEE	54.7380	55.0880	50	50
CSEO	CS iTraxx Mini	ICEE	54.610	55.090	100	100
BARX	Barclays	ICEE	54.7650	55.1150	250	250
CGCX	Citi CGCX	ICEE	54.6800	55.1200	100	100
EBNP	BNP Paribas	ICEE	54.7250	55.1250	101	101
SCDS	SocGen	ICEE	54.6890	55.1380	125	125
DBVD	DB Index-(DBDV)	ICEE	54.8500	55.1500	100	100
GSET	GOLDMAN SACHS	ICEE	54.6100	55.2100	75	75
CCGB	Citi CCGB	ICEE	54.5600	55.2400	200	200
JPOS	JP Morgan	ICEE	54.5600	55.2400	200	200
MSTT	MORGAN STANLEY MAXI	ICEE	54.5500	55.2900	100	100
CSXE	Credit Suisse EU	ICEE	54.406	55.294	200	200

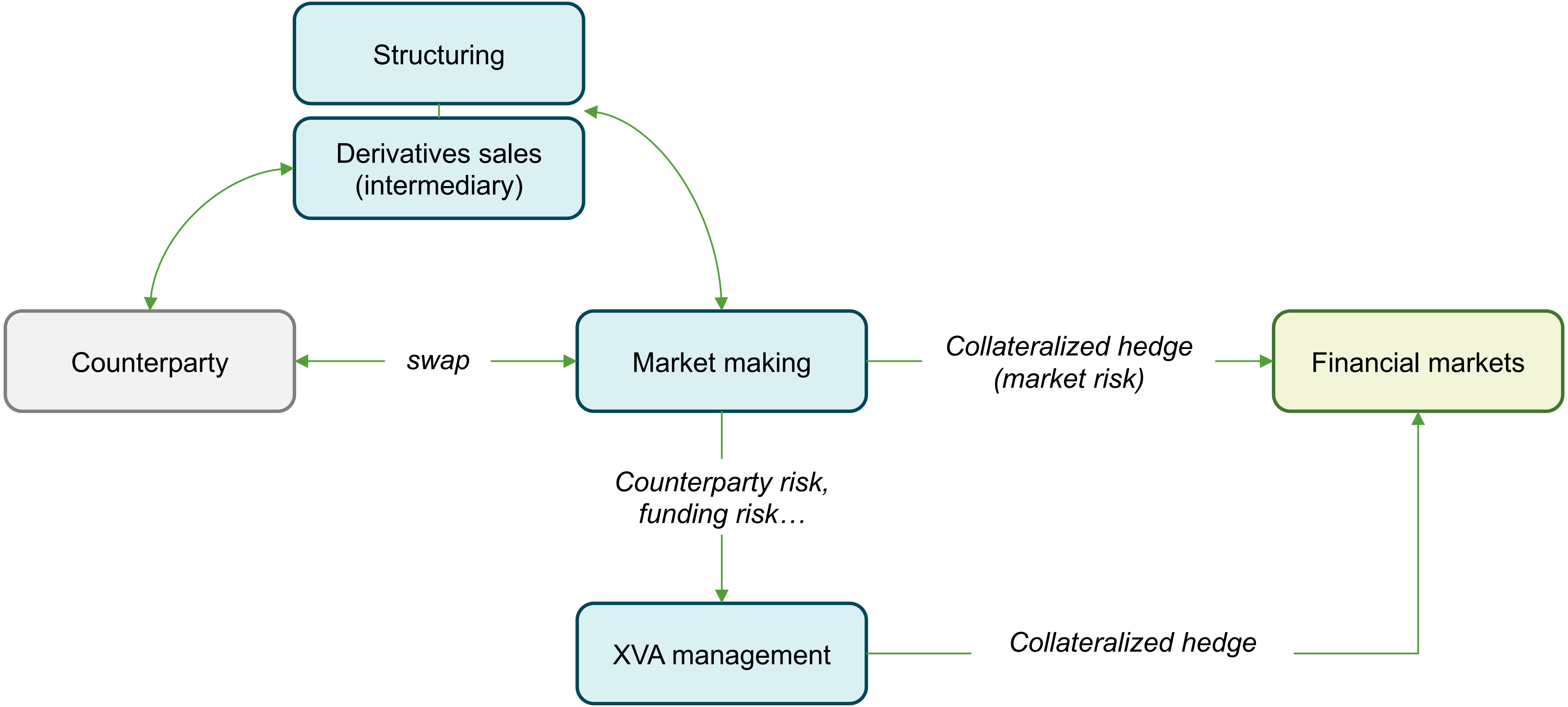
RFQ Example

Client buys protection 200mln
Price: ____

Send

Corporate Derivatives: Swap components

CIB | Capital Markets



XVA’s: Valuation adjustments (1/2)

CIB | Capital Markets

Valuation Adjustment	Description
CVA	Counterparty credit risk. An extra charge given the risk of the counterparty
DVA	Own counterparty risk. A discount on the price in exchange for my liability.
FVA	Funding cost (or benefit) if the corporate derivative is ITM, then the hedge is OTM and I need to pay collateral which must be funded
MVA	Cost of financing initial margins
KVA	Capital resources required to match regulatory requirements from Basel III and SACCR.
CollVA, AVA	…

Profit Centres of Banks

Introduction – Main services offered by banks and their technological focus

	Retail Bank	CIB	Private Banking/Wealth Management
Services	• Receive deposits • Offer loans	• Investment banking: M&A, ECM, DCM • Capital markets: sales & trading • Structured finance	• Mutual funds • Hedge funds • Private equity • Private banking
Revenue	• Difference between loan interest and deposit interest	• Advisory fees • Capital gains + margins • Interest rate	• % fee on AUM + performance fee
Technological focus	• Chatbots • Ads for products • Metaverse?	• Analysing financial statements • Compiling slides • Client segmentation • Automating traders?	• Portfolio optimization • Stock picking • Analysing financial statements

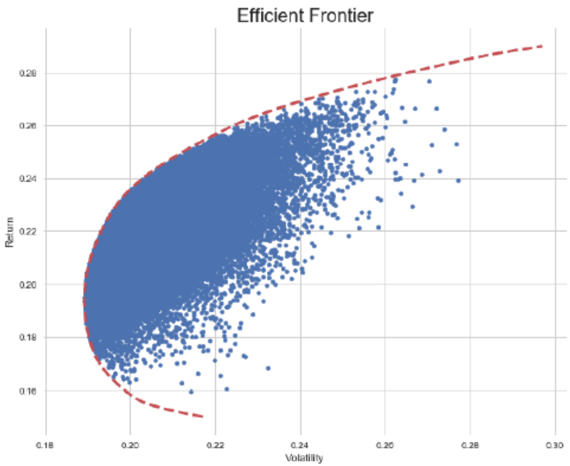
Focus next

Portfolio Optimization

Wealth Management

Definition

- Given an investment universe of M assets, the objective is to decide what proportion of the total available budget to invest in each of the M assets



Background

- **Modern Portfolio Optimization**
  [Markowitz, 1952]
  - Calculate variance and correlations
  - Single period
- **Intertemporal CAPM**
  [Merton, 1969]
  - Make assumptions on asset dynamics
  - Multi period
- **Online Portfolio Optimization**
  [Cover and Ordentlich, 1996]
  - Adversarial market
  - Multi period

Optimal Execution

Order Execution

Description

- In prop trading, the trader decides his strategy and also executes the trades
- In asset management, the portfolio manager decides the portfolio allocation, and the execution is done by an execution desk
- When the execution desk receives an order of size X, the objective is to execute in a specified amount of time, by minimizing the difference between the arrival price and the execution price

Limit order book example

Last	Last Vol	Total Vol	Close	Daily Low	Daily High
4045.00	2	367267	4097.50	4033.50	4101.50

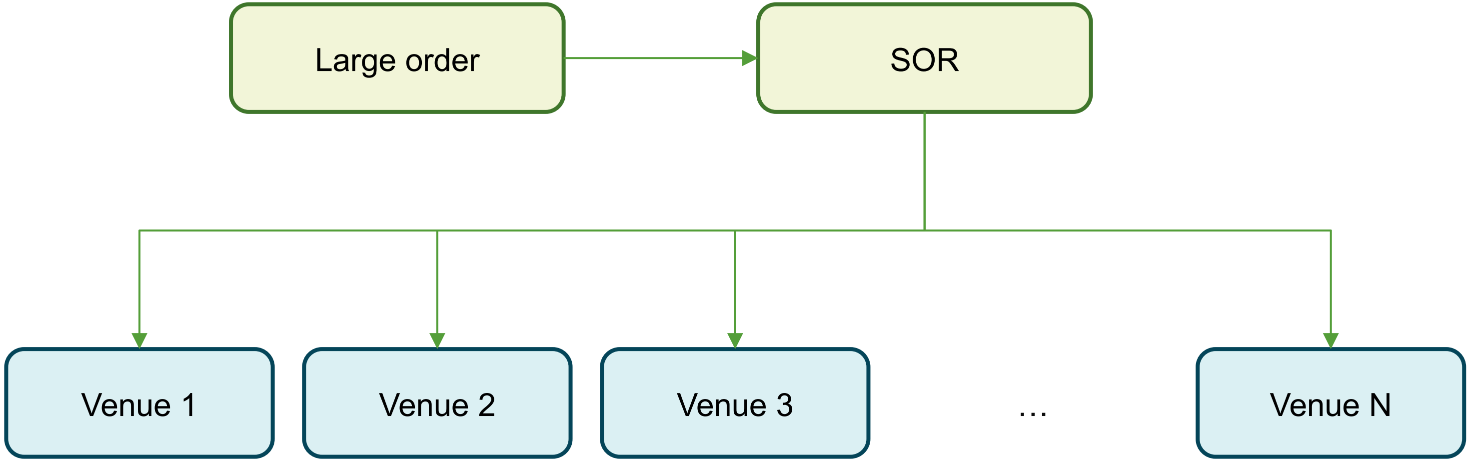
Implied

Bid		Offer	
Volume	Price	Price	Volume
136	4044.50	4045.00	62
327	4044.00	4045.50	293
348	4043.50	4046.00	427
620	4043.00	4046.50	426
358	4042.50	4047.00	463
330	4042.00	4047.50	348
325	4041.50	4048.00	327
318	4041.00	4048.50	294
305	4040.50	4049.00	281
512	4040.00	4049.50	288

Smart Order Routing

Order Execution

- Smart Order Routing (SOR): optimally splitting an order over multiple venues.



AGENDA

Introduction to Banks
- Introduction
- Capital Markets
- Wealth Management
- Order Execution

Algorithms in the Financial Markets
- Introduction
- Reinforcement Learning
- Use cases

Algorithms in the Financial Markets

1 Algorithmic Trading

2 Reinforcement Learning

3 Quantitative Trading

4 Online Portfolio Optimization

5 Optimal Execution

6 Smart Routing with CMABs

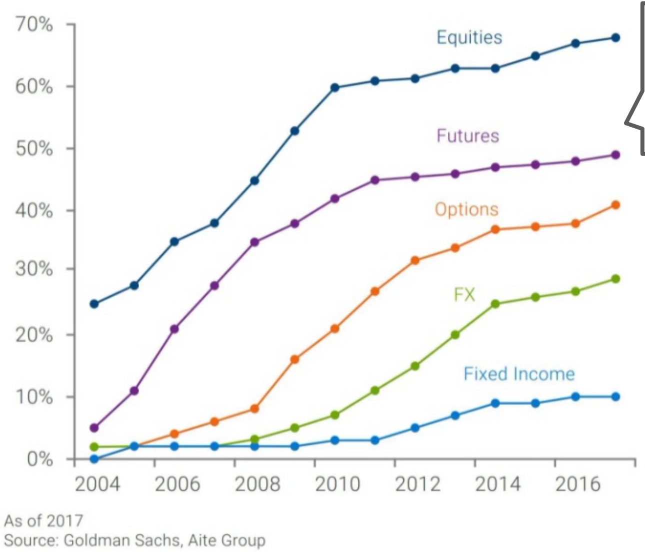
7 Market Making with MFGs

8 Hedging with Risk Averse RL

Algorithmic Trading

Market and types of trading algorithms

Share of algorithmic trading market by asset class



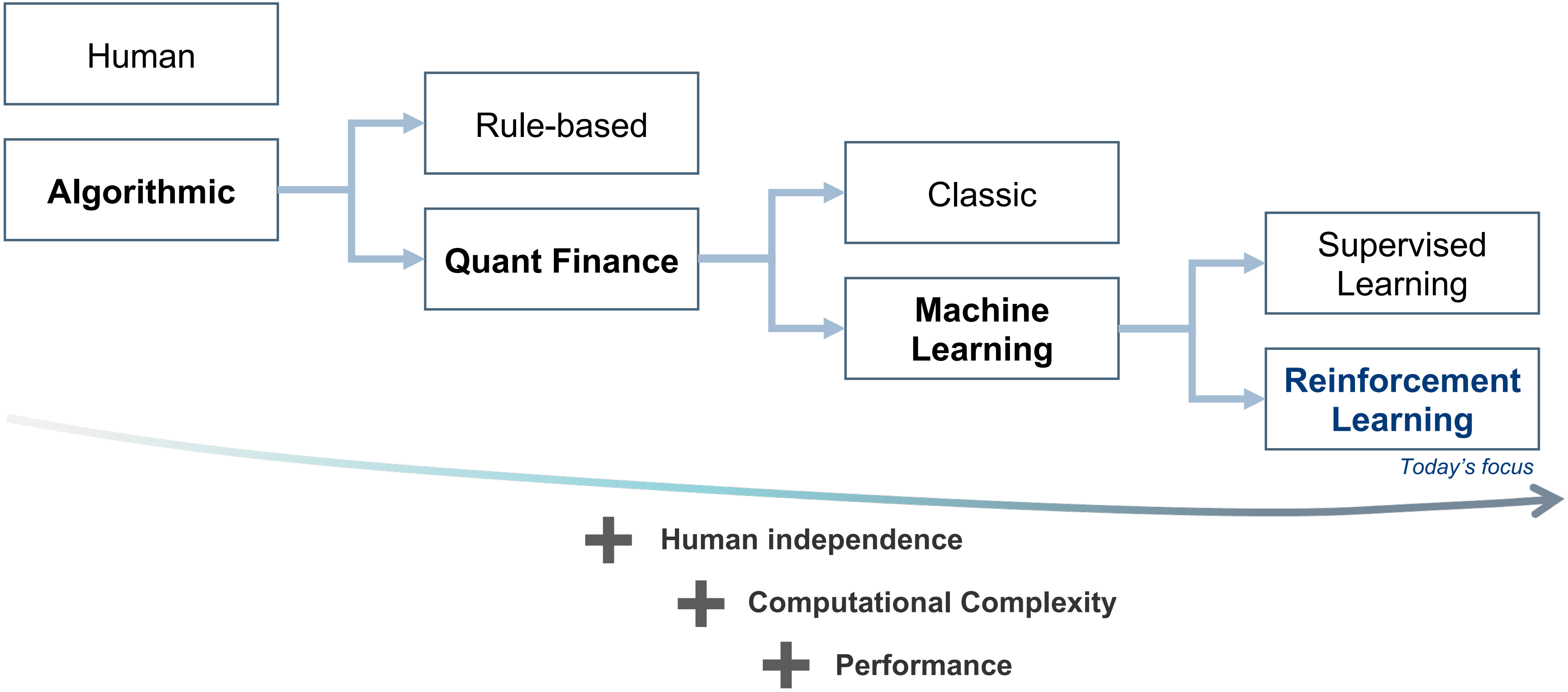
The algorithmic trading market grows with a CAGR of **~11%** ('21-'26)

Main types of algorithms

- Optimal execution and smart routing
- Market making
- Hedging
- Trading
- Portfolio optimization

Algorithmic Trading Technologies

Classification by technology type



Reinforcement Learning for Trading

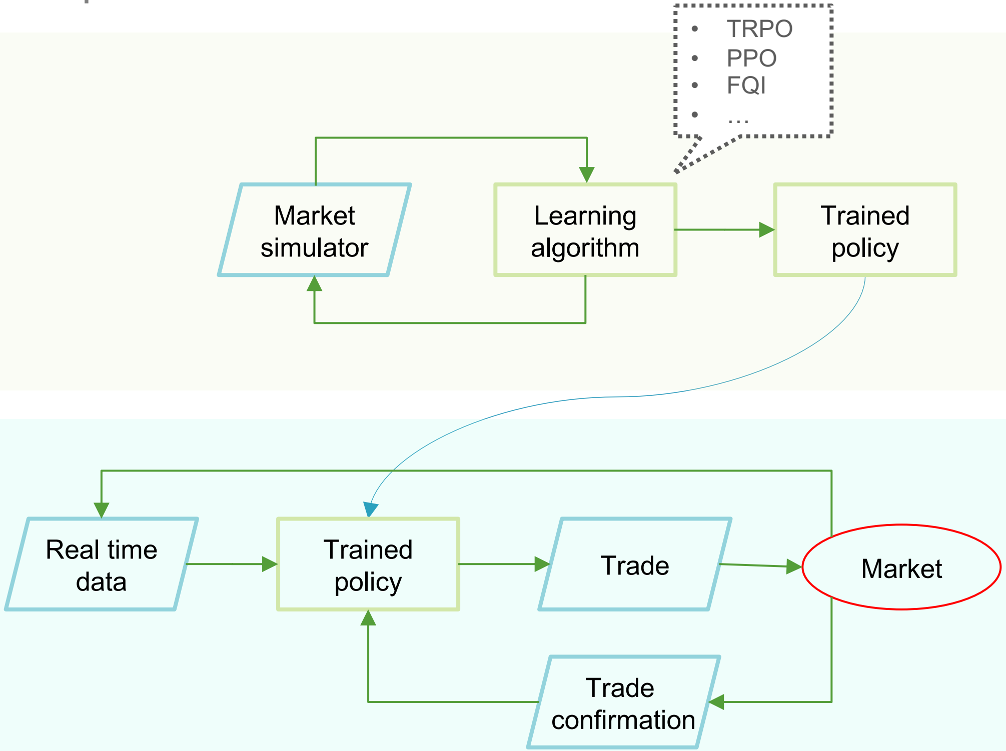
Training, testing and use in production

Phase 1

- Training
- Hyperparameter tuning
- Backtesting

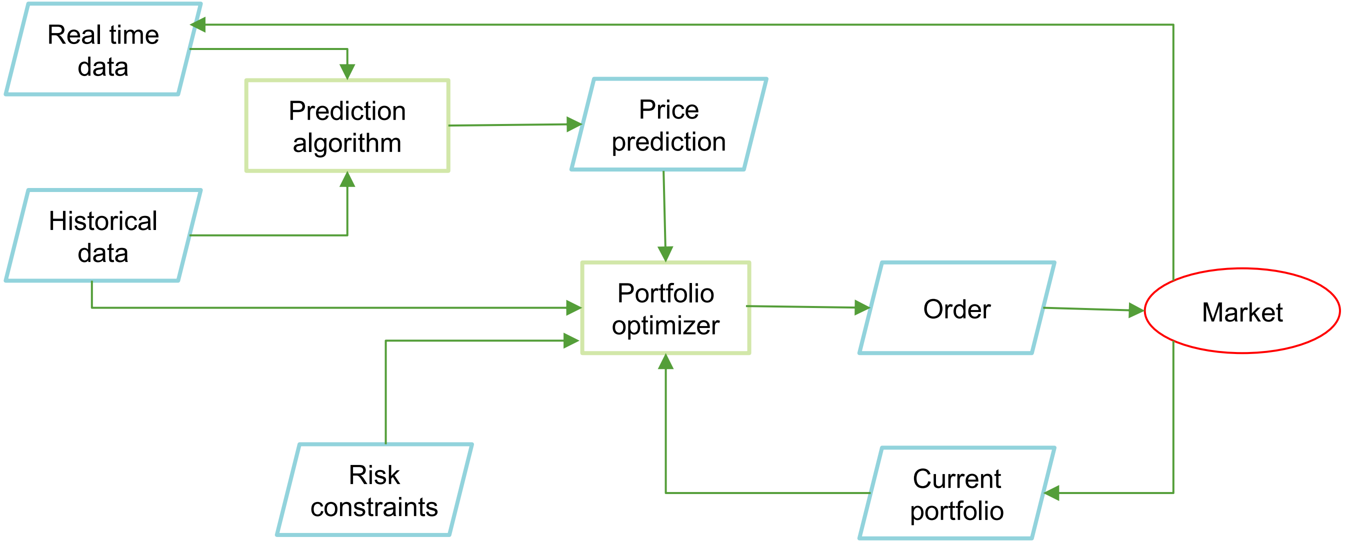
Phase 2

- Production



Supervised learning for Quantitative Trading

Trading system architecture using a supervised learning approach



Key Points

- Necessary to create a labelled dataset
- Supervised algorithm output is a prediction
- It is necessary to have a portfolio optimiser

Algorithms in the Financial Markets

1 Algorithmic Trading

2 Reinforcement Learning

3 Quantitative Trading

4 Online Portfolio Optimization

5 Optimal Execution

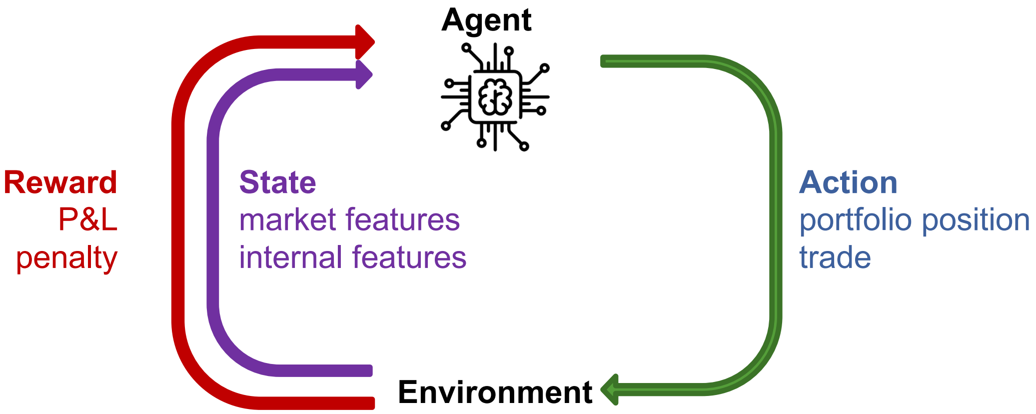
6 Smart Routing with CMABs

7 Market Making with MFGs

8 Hedging with Risk Averse RL

Reinforcement Learning Basics

Markov Decision Process: process which describes interaction between agent and environment



- The objective is finding the policy π which maximizes the discounted sum of the rewards
- $J = \max_{\pi} \mathbb{E}_t[\sum \gamma^t R_t]$

Q-function and Policy

RL algorithms enable the learning of the policy π

The objective is to find the π that maximises J : $J = \max_{\pi} \mathbb{E}_{\pi}[\sum \gamma^t R_t]$

Q-learning

- Q-function

  $$Q_{\pi} = \mathbb{E}_{\pi}[\sum \gamma^t R_t | s_0, a_0]$$

- Bellman Equation

  $$Q_{\pi} = r(s,a) + \gamma \mathbb{E}_{s',a'}[Q_{\pi}(s',a')]$$

- Q-learning algorithm

  $$Q_t(s,a) = r(s,a) + \gamma \max_{a'} Q_t(s',a')$$

- Q-learning is a tabular algorithm which can be generalized using function approximators such as Xgboost.

Policy Search

- Policy gradient theorem

  $$\nabla_{\theta} J_{\pi_{\theta}} = \mathbb{E}[\nabla \log \pi_{\theta}(a|s) Q_{\pi_{\theta}}(s,a)]$$

- Policy update

  $$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta} J_{\pi_{\theta}}$$

- The policy is a parametric and differentiable function, usually a neural network

Multi Armed Bandits (MAB)

Partial feedback algorithms – stochastic environments

Characteristics

- Field of research close to RL
- Objective is to learn sequential decision processes
- Online algorithms
- MAB algorithms choose at each timestep which arm to pull
- Regret guarantees: finding the best arm in sub-linear time
- Regret: $$R_T = \sum_{t=1}^{T} [f_t(a_t, y_t) - f_t(a^*, y_t)]$$

*a** is the best arm

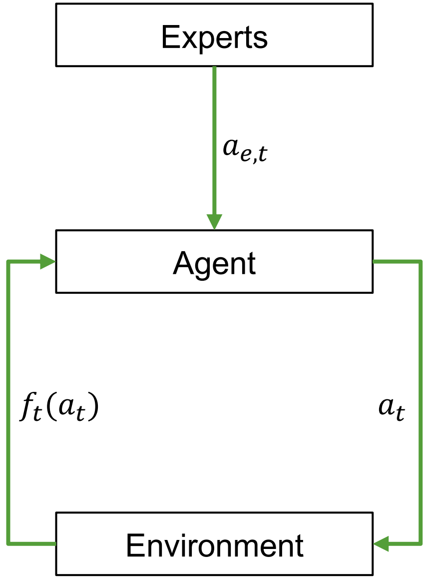
Expert Learning

Full feedback algorithms – adversarial environments

Characteristics

- Field of research close to RL
- Objective is to learn sequential decision processes
- Online algorithms
- Expert learning algorithms choose at each timestep which experts to follow
- Regret guarantees: finding the best expert in sub-linear time
- Regret $R_T = \sum_{t=1}^{T} f_t(a_t, y_t) - \inf_{e \in \mathcal{E}} \sum_{t=1}^{T} f_t(a_{e,t}, y_t).$

Expert interaction scheme



Algorithms in the Financial Markets

1 Algorithmic Trading

2 Reinforcement Learning

3 Quantitative Trading

4 Online Portfolio Optimization

5 Optimal Execution

6 Smart Routing with CMABs

7 Market Making with MFGs

8 Hedging with Risk Averse RL

Reinforcement Learning for Quantitative Trading

Problem description and MDP definition

Quantitative Trading

Definition

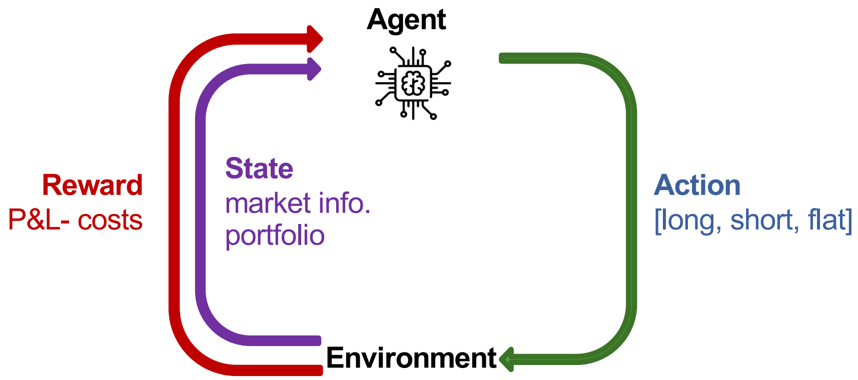
- At each timestep, decide whether to go long, short or flat to maximize gains

MDP

- **State:** price window, bid-ask spread, current portfolio, date/time
- **Action:** long, short, flat
- **Reward:** P&L – transaction costs

Characteristics

- Alpha seeking
- Low market correlation



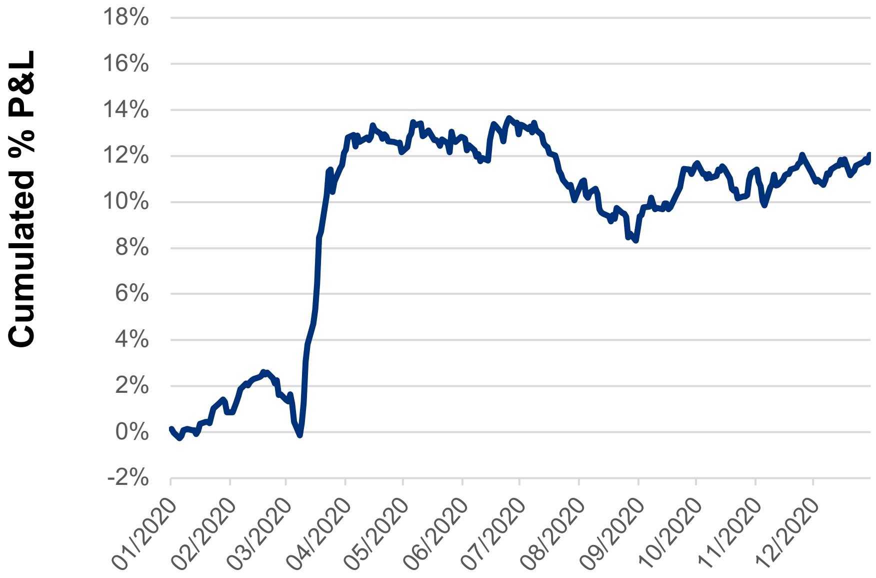
Reinforcement Learning for FX Trading (1/2)

Experimental results - performance

Experiment

- Intraday trading on EURUSD FX
- Training with FQI on historical data 2017-2018
- Validation on historical data 2019
- Backtesting on historical data out-of-sample 2020

P&L of backtest EURUSD FX trading on 2020



Learning FX Trading Strategies with FQI and Persistent Actions, ICAIF 2021

Reinforcement Learning for FX Trading (2/2)

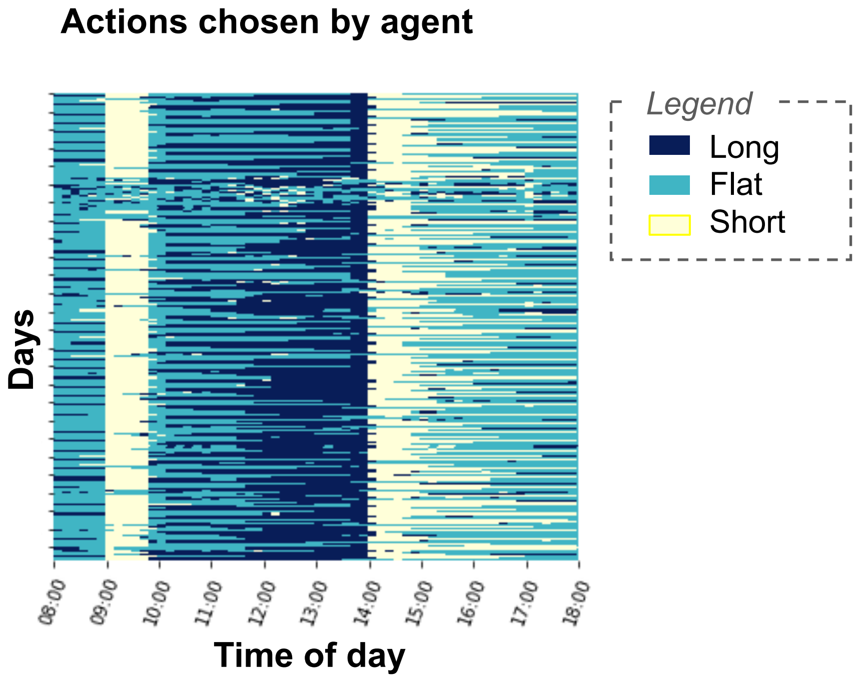
Experimental results - policy

Experiment

- Intraday trading on EURUSD FX
- Training with FQI on historical data 2017-2018
- Validation on historical data 2019
- Backtesting on historical data out-of-sample 2020

Can we improve?

- Market non-stationarity



Learning FX Trading Strategies with FQI and Persistent Actions, ICAIF 2021

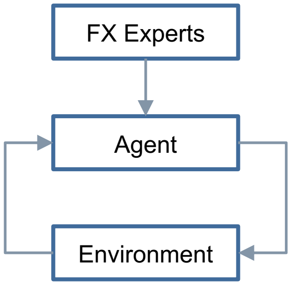
Reinforcement ed Expert Learning per FX Trading

Expert Learning on FX trading

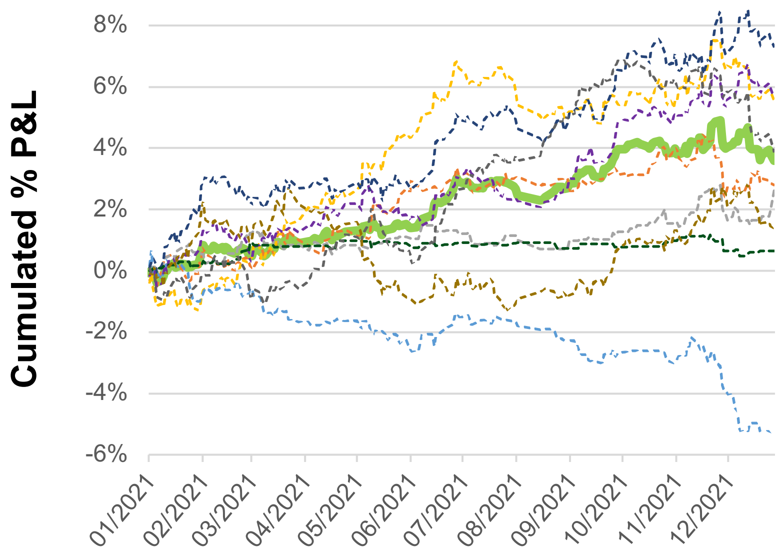
Description

- = trading strategies
- = expert learning strategies

Expert interaction scheme



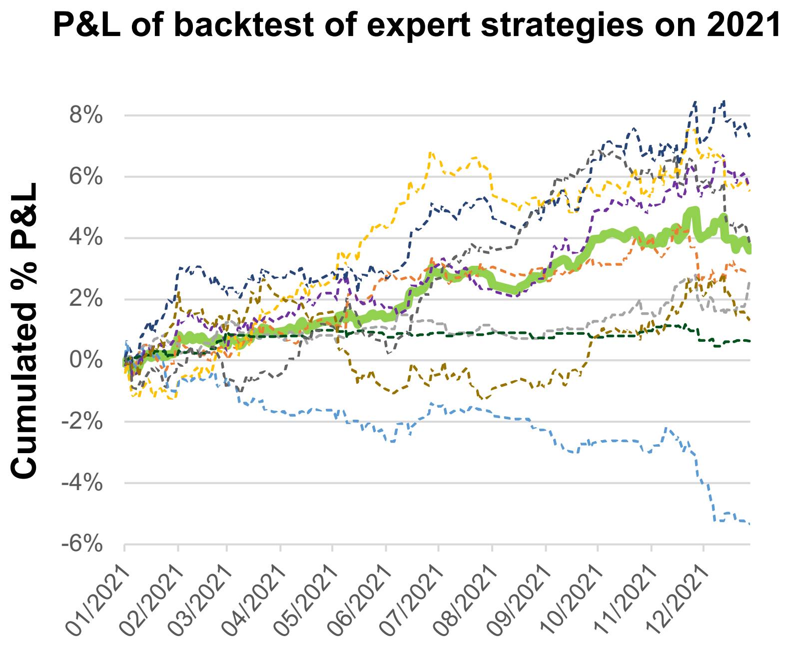
P&L of backtest on 2021

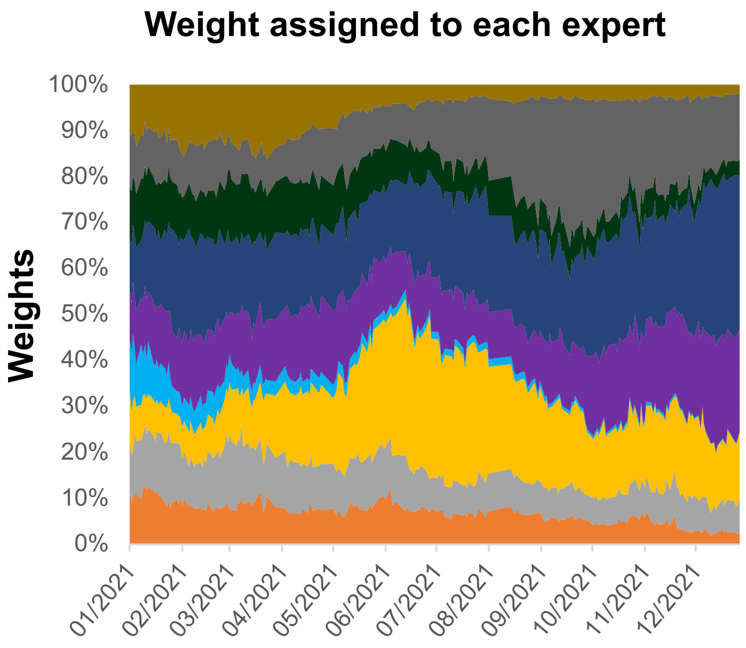


Addressing Non-Stationarity in FX Trading with Online Model Selection of Offline RL Experts, ICAIF 2022

Reinforcement and Expert Learning for FX Trading

Example using Expert Learning on FX trading





Algorithms in the Financial Markets

1 Algorithmic Trading

2 Reinforcement Learning

3 Quantitative Trading

4 Online Portfolio Optimization

5 Optimal Execution

6 Smart Routing with CMABs

7 Market Making with MFGs

8 Hedging with Risk Averse RL

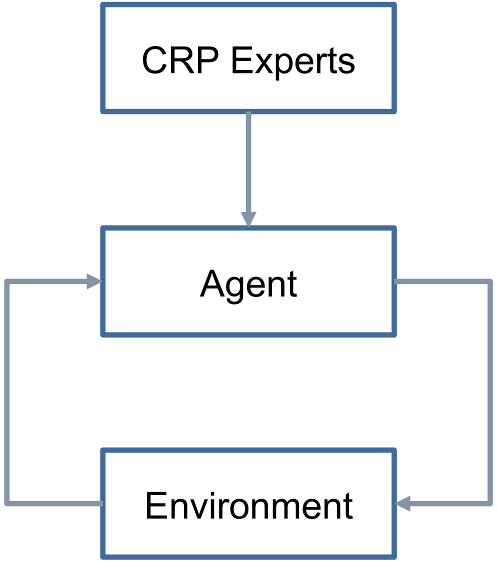
Online Portfolio Optimization

From Expert Learning to Online Portfolio Optimization (OPO)

Definitions and notation

- $\boldsymbol{a}_t \in \Delta_{M-1}$ is the portfolio allocation, with M assets
- The experts are Constant Rebalancing Portfolios (CRPs)
- $\boldsymbol{a}^* = \operatorname{argmin}_{\boldsymbol{a}\in\Delta_{M-1}} \sum_t f_t(\boldsymbol{a}, \boldsymbol{y}_t)$ is the best CRP
- $f_t(\boldsymbol{a}, \boldsymbol{y}_t) = -\log < \boldsymbol{a}, \boldsymbol{y}_t >$ is the loss
- $y_t = \left(\frac{p_{t,1}}{p_{t-1,1}}, \dots, \frac{p_{t,M}}{p_{t-1,M}}\right)$ are the price relatives
- $W_T(a_1, \dots, a_T) = \Pi_t^T < a_{t,} y_t >$ is the wealth
- Regret $R_T = \sum_{t=1}^{T} f_t(\mathbf{a}_t, \mathbf{y}_t) - \min_{a\in\Delta_{M-1}} \sum_{t=1}^{T} f_t(\mathbf{a}, \mathbf{y}_t)$

OPO interaction scheme



Universal Portfolios (UP)

The first algorithm in the OPO field

Algorithm 3 Universal Portfolios [Cover and Ordentlich, 1996]

1: Input M assets, set $\mathbf{a}_1 \leftarrow \frac{1}{M}\mathbf{1}$, initialize $\mathbf{W}_1$
2: **for** $t \in \{1, \ldots, T\}$ **do**
3: Select $\mathbf{a}_{t+1} \leftarrow \frac{\int_{\mathbf{b}\in\Delta_{M-1}} \mathbf{b} W_t(\mathbf{b}) \mathrm{d}\mu(\mathbf{b})}{\int_{\mathbf{b}\in\Delta_{M-1}} W_t(\mathbf{b}) \mathrm{d}\mu(\mathbf{b})}$
4: Observe $\mathbf{y}_{t+1}$ from the market
5: Get wealth increase $\langle \mathbf{y}_{t+1}, \mathbf{a}_{t+1} \rangle$
6: **end for**

- Regret $O(M \log T)$
- Computational Complexity $\Theta(T^M)$

Online Gradient Descent (OGD)

Moving towards the minimum of the log loss function

Algorithm 4 Online Gradient Descent [Zinkevich, 2003]

Require: learning rate sequence $\{\eta_1, \ldots, \eta_T\}$

1: Input M assets, set $\mathbf{a}_1 \leftarrow \frac{1}{M}\mathbf{1}$

2: **for** $t \in \{1, \ldots, T\}$ **do**

3: Select $\mathbf{a}_{t+1} \leftarrow \Pi_{\Delta_{M-1}}\left(\mathbf{a}_t + \eta_t \frac{\mathbf{y}_t}{\langle \mathbf{y}_t, \mathbf{a}_t \rangle}\right)$

4: Observe $\mathbf{y}_{t+1}$ from the market

5: Get wealth increase $\langle \mathbf{y}_{t+1}, \mathbf{a}_{t+1} \rangle$

6: **end for**

- Regret $O(\sqrt{T})$
- Computational Complexity $\Theta(M)$

Online Gradient Descent with Momentum (OGDM)

Keeping transaction costs under control

Algorithm 6 OGDM in OPO with Transaction Costs

Require: learning rate sequence $\{\eta_1, \ldots, \eta_T\}$, momentum parameter sequence $\{\lambda_1, \ldots, \lambda_T\}$

1: Set $\mathbf{a}_1 \leftarrow \frac{1}{M}\mathbf{1}$
2: **for** $t \in \{1, \ldots, T\}$ **do**
3: Select $\mathbf{a}_{t+1} \leftarrow \Pi_{\Delta_{M-1}}\left(\mathbf{a}_t + \eta_t \frac{\mathbf{y}_t}{\langle \mathbf{y}_t, \mathbf{a}_t \rangle} - \frac{\lambda_t}{2}(\mathbf{a}_t - \mathbf{a}_{t-1})\right)$
4: Observe $\mathbf{y}_{t+1}$ from the market
5: Get wealth $\log(\langle \mathbf{y}_{t+1}, \mathbf{a}_{t+1} \rangle) - \gamma ||\mathbf{a}_{t+1} - \mathbf{a}_t||_1$
6: **end for**

- Total Regret $O(\sqrt{T})$
- Computational Complexity $\Theta(M)$

$$R_T^C = \underbrace{\sum_{t=1}^{T} f_t(\mathbf{a}_t, \mathbf{y}_t) - \min_{a \in \Delta_{M-1}} \sum_{t=1}^{T} f_t(\mathbf{a}, \mathbf{y}_t)}_{R_T:\ \text{standard regret}} + \gamma \underbrace{\sum_{t=1}^{T} ||\mathbf{a}_t - \mathbf{a}_{t-1}||_1}_{C_T:\ \text{transaction costs}}$$

Dealing with Transaction Costs in Portfolio Optimization: Online Gradient Descent with Momentum, ICAIF 2020

Online Newton Step (ONS)

Second order algorithm

Algorithm 5 Online Newton Step [Agarwal et al., 2006]

Require: β, δ

1: Input M assets, set $\mathbf{a}_1 \leftarrow \frac{1}{M}\mathbf{1}_M$

2: **for** $t \in \{1, \ldots, T\}$ **do**

3: Select $\mathbf{a}_{t+1} \leftarrow \Pi^{\mathbf{A}_t}_{\Delta_{M-1}}(\mathbf{a}_t - \frac{1}{\beta}\mathbf{A}_t^{-1}\mathbf{b}_t)$, where:

$\mathbf{b}_t = \sum_{\tau=1}^{t} \nabla[\log_\tau(\mathbf{a}_\tau \cdot \mathbf{y}_\tau)])$

$\mathbf{A}_t = \sum_{\tau=1}^{t} \nabla^2[\log(\mathbf{a}_\tau \cdot \mathbf{y}_\tau)] + \mathbf{1}_M$

$\Pi^{\mathbf{A}_t}_{\Delta_{M-1}}$ is the projection in the norm induced by $\mathbf{A}_t$

4: Observe $\mathbf{y}_{t+1}$ from the market

5: Get wealth increase $\langle \mathbf{y}_{t+1}, \mathbf{a}_{t+1} \rangle$

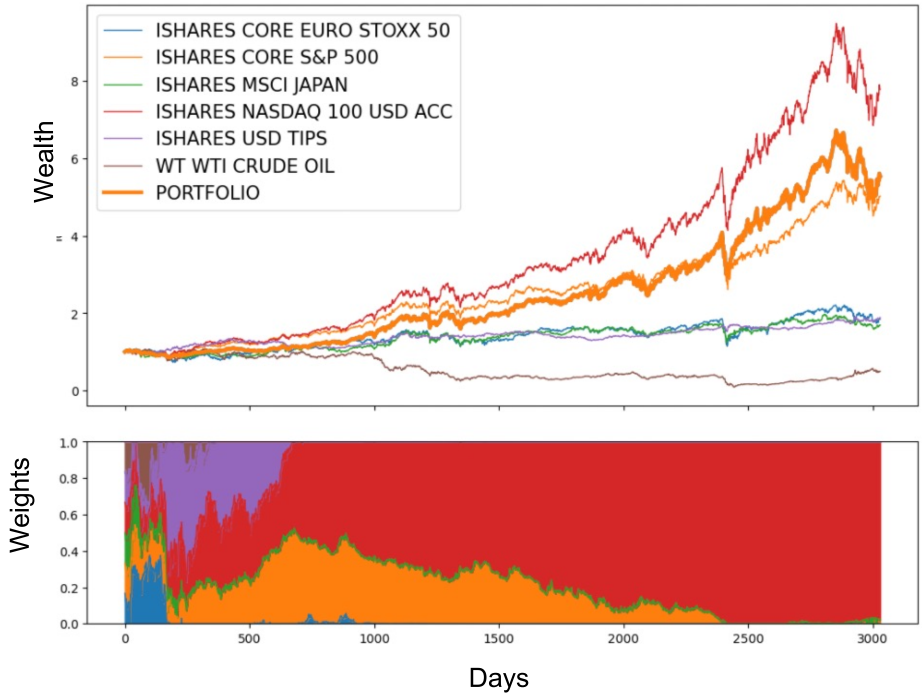
6: **end for**

- Regret $O(M \log T)$
- Computational Complexity $\Theta(M^2)$

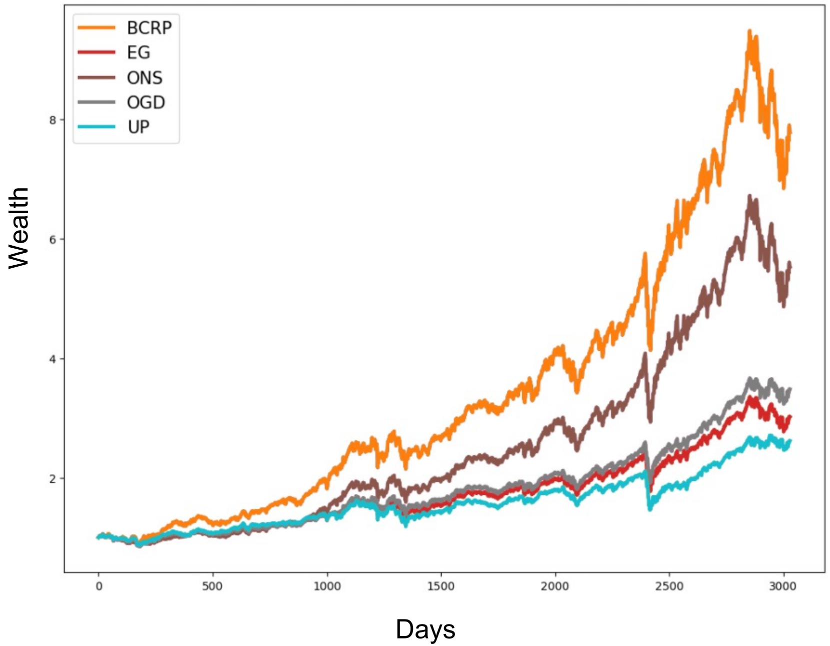
Algorithm Comparison

OPO experimental examples

ONS performance and weights



Wealth of expert strategies



If we consider market impact?

- Up to now we considered transaction costs but no market impact.
- What happens if we have market impact?

?

Algorithms in the Financial Markets

1 Algorithmic Trading

2 Reinforcement Learning

3 Quantitative Trading

4 Online Portfolio Optimization

5 Optimal Execution

6 Smart Routing with CMABs

7 Market Making with MFGs

8 Hedging with Risk Averse RL

Limit Order Book

Definition and limit order book example

Characteristics

- Limit order book is the record of all limit orders which have not been executed
- Limit order is an order which specifies both price and volume of a trade
- Market order is an order to execute immediately at the best price possible

Example of Limit Order Book

Last	Last Vol	Total Vol	Close	Daily Low	Daily High
4045.00	2	367267	4097.50	4033.50	4101.50

Implied

Bid		Offer	
Volume	Price	Price	Volume
136	4044.50	4045.00	62
327	4044.00	4045.50	293
348	4043.50	4046.00	427
620	4043.00	4046.50	426
358	4042.50	4047.00	463
330	4042.00	4047.50	348
325	4041.50	4048.00	327
318	4041.00	4048.50	294
305	4040.50	4049.00	281
512	4040.00	4049.50	288

Reinforcement Learning for Optimal Execution

Problem definition and MDP description

Optimal Execution

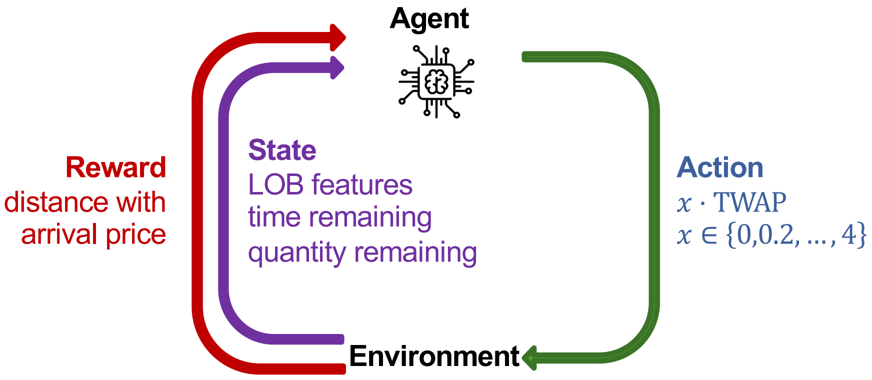
Definition

- Execute X shares in N timesteps
- Decide at each timestep the trade to execute so to minimize difference between arrival and execution price

MDP

- **State:** LOB features, remaining timesteps, remaining quantity
- **Action:** x·TWAP with $x \in \{0, 0.2, \ldots, 4\}$
- **Reward:** distance with arrival price

$$r_t = \left(1 - \frac{P_{fill} - P_{arr}}{P_{fill}}\right) \lambda \frac{n_t}{X}$$



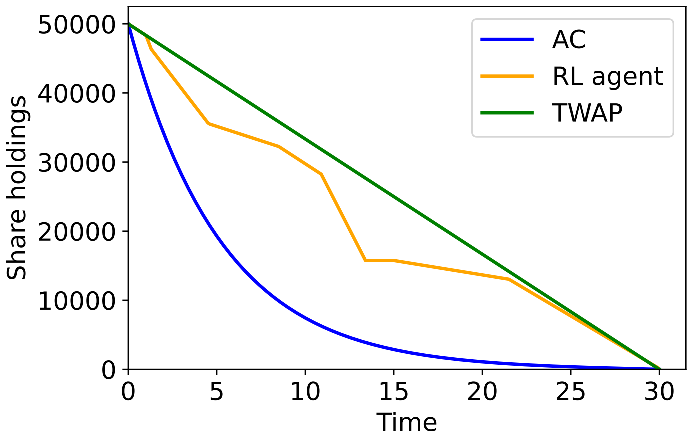
Experimental Results

Return comparison between RL agent and benchmark on a market simulated with ABIDES

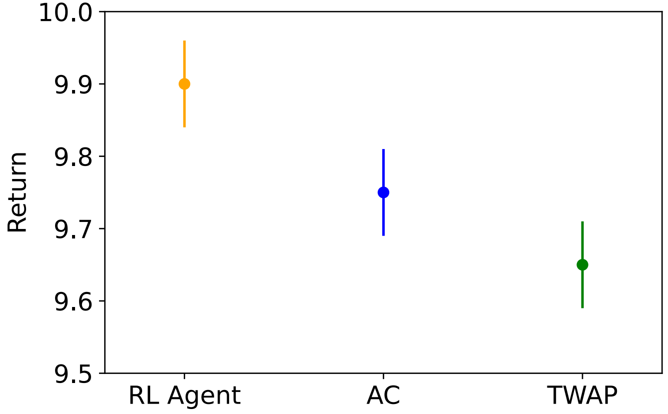
Characteristics

- Simulating with ABIDES the optimal execution exercise
- 30 minutes to execute 50k shares
- $r_t = \left(1 - \frac{P_{fill} - P_{arr}}{P_{fill}}\right) \lambda \frac{n_t}{X}$

Execution trajectories



Average RL agent returns vs benchmark



Algorithms in the Financial Markets

1 Algorithmic Trading

2 Reinforcement Learning

3 Quantitative Trading

4 Online Portfolio Optimization

5 Optimal Execution

6 Smart Routing with CMABs

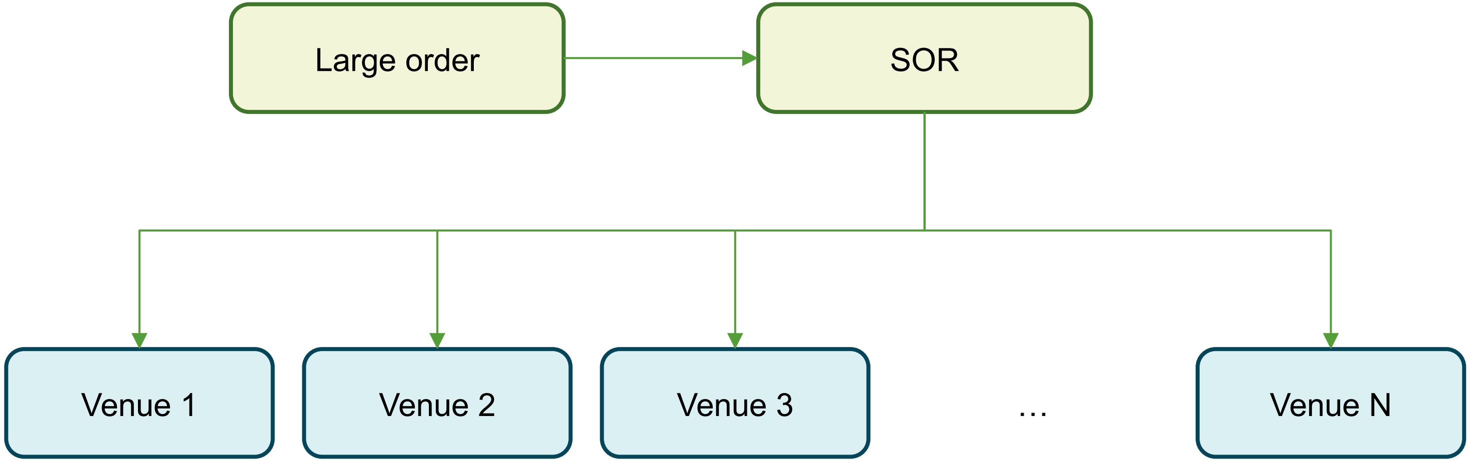
7 Market Making with MFGs

8 Hedging with Risk Averse RL

Smart Order Routing

Order Execution

- Smart Order Routing (SOR): optimally splitting an order over multiple venues.



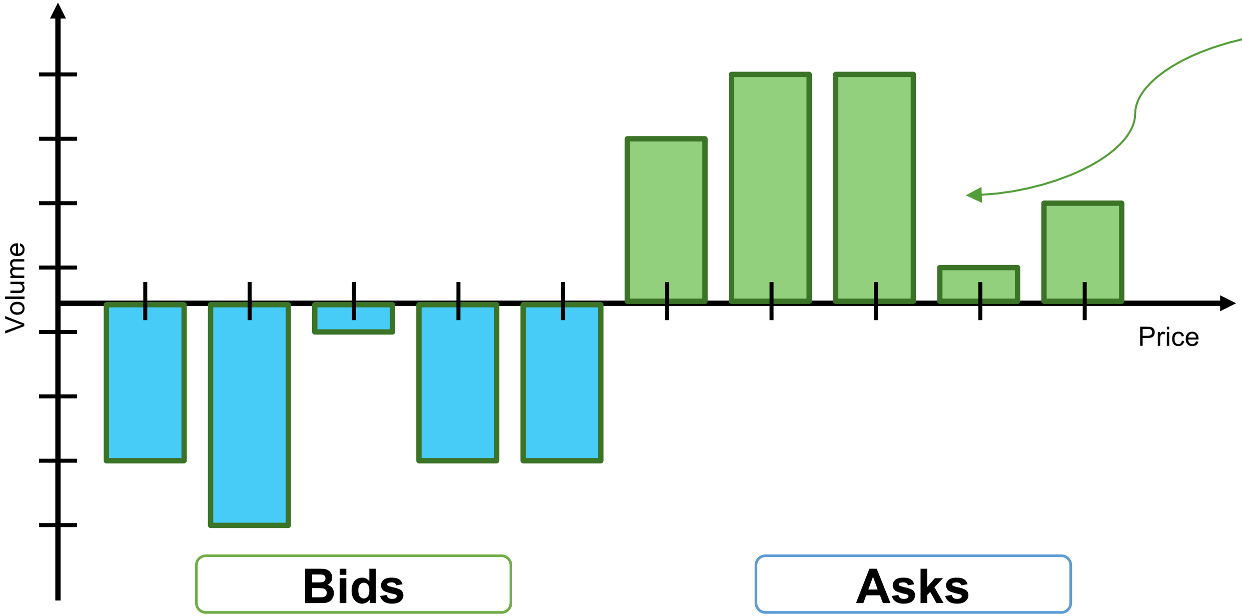
Regulated Exchanges - Limit Order Book

LOB visualization

Last	Last Vol	Total Vol	Close	Daily Low	Daily High
4045.00	2	367267	4097.50	4033.50	4101.50

Implied

Bid		Offer	
Volume	Price	Price	Volume
136	4044.50	4045.00	62
327	4044.00	4045.50	293
348	4043.50	4046.00	427
620	4043.00	4046.50	426
358	4042.50	4047.00	463
330	4042.00	4047.50	348
325	4041.50	4048.00	327
318	4041.00	4048.50	294
305	4040.50	4049.00	281
512	4040.00	4049.50	288



Dark Pools

The latent limit order book is invisibile to the market participants

Dark Pool Smart Order Routing - DPSOR

Defining SOR as a sequential decision problem

Task

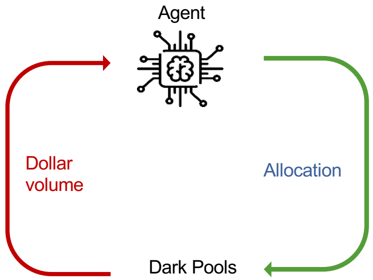
- Create and maintain an **estimate** of **hidden liquidity** of multiple dark pools
- Make optimal joint **routing and pricing decisions**
- Optimize the **dollar volume**

Assumptions

- **Multiple dark pools** for a single asset
- Stationary liquidity
- **Limit orders** are admitted

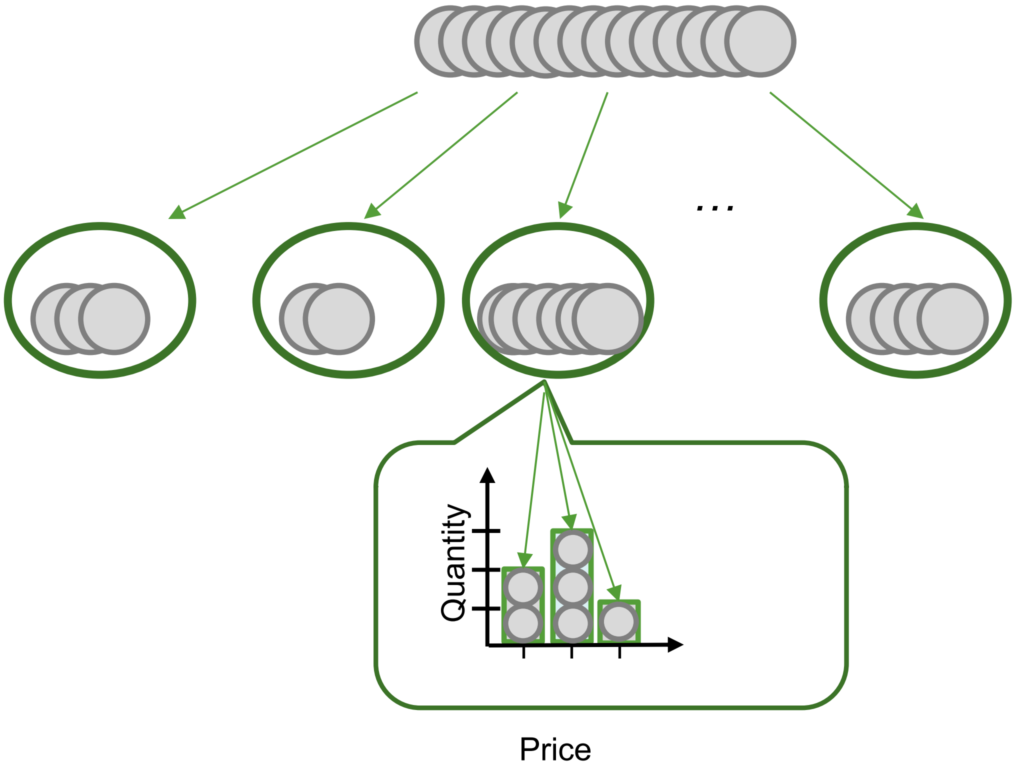
Formulation

- **Sequential decision problem** where at each round *t*, an agent, given a volume V of shares to execute, must maximize the dollar volume by allocating the shares across K dark pools, specifying the price



Joint routing and pricing allocation

Defining both the dark pool and the limit price



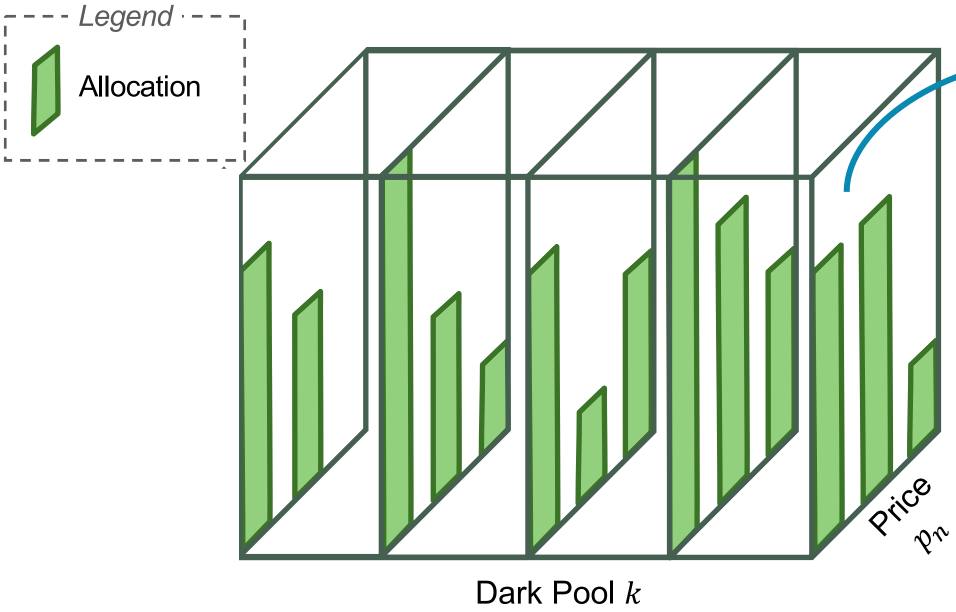
V units to sell

Allocate to K dark pools

Specify amount to allocate at a specific price

Problem formalization and notation

Defining constraints and censored feedback



A^t_{kn}: amount allocated at round t to dark pool k at price p_n

- We have the constraint that

$$\sum_{k=1}^{K}\sum_{n=1}^{N} A^t_{kn} = V$$

- Our objective is the allocation that maximizes dollar volume

$$R_t(\mathfrak{U}) = \sum_{k=1}^{K}\sum_{n=1}^{N} r^t_{kn}\, p_n.$$

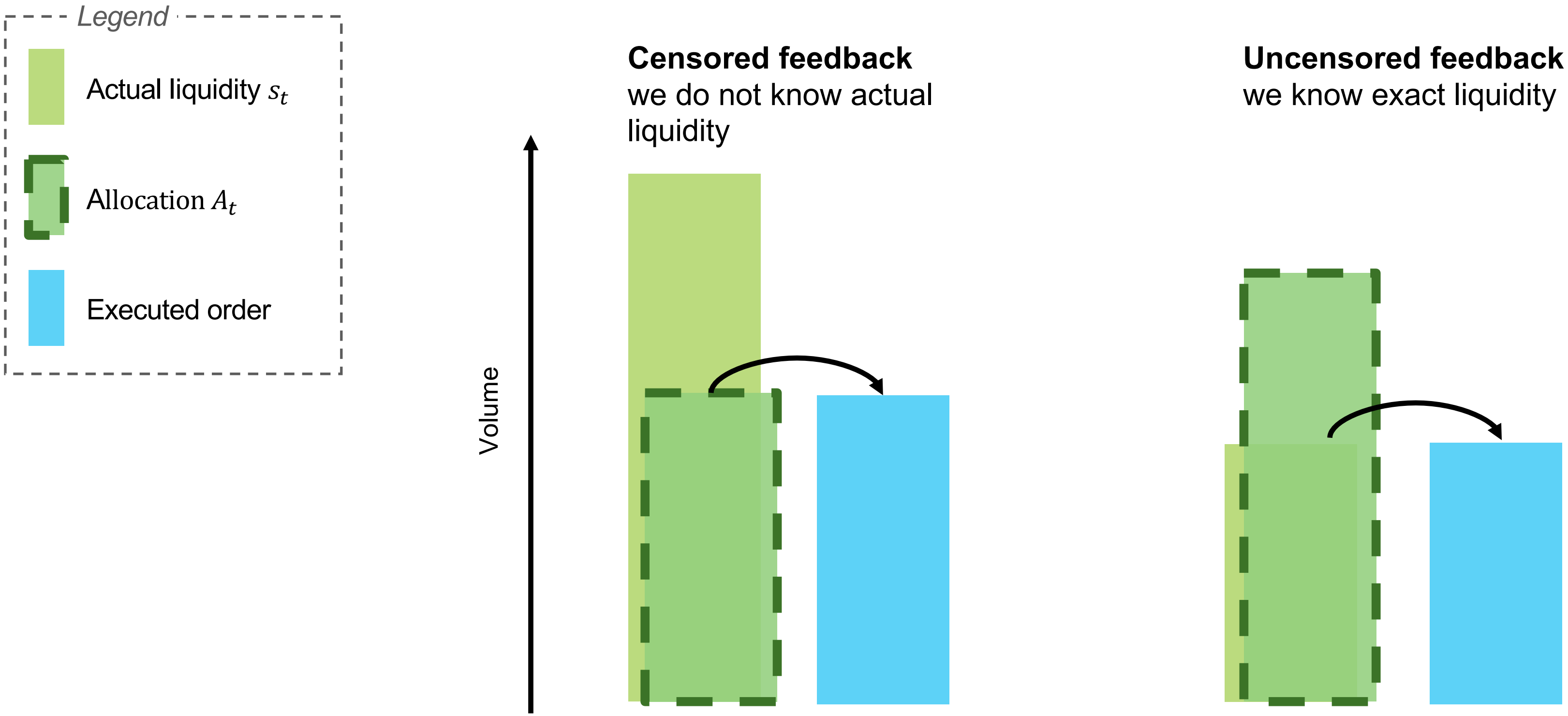
Censored feedback

$$r^t_{kn} = \min\{A^t_{kn}, s^t_{kn}\}$$

s^t_{kn} is the actual liquidity present at time t in dark pool k at price p_n

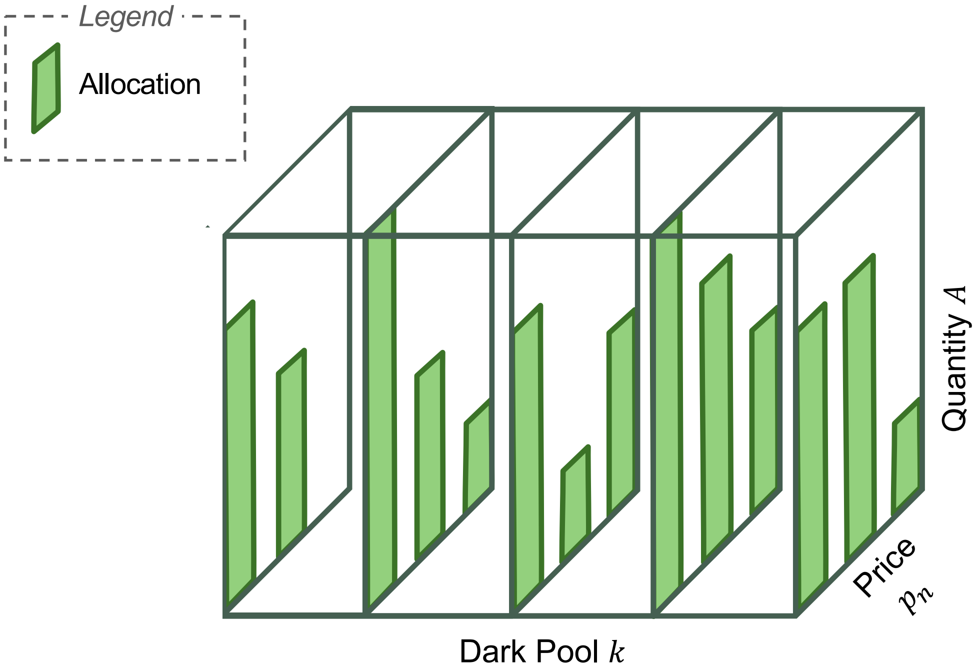
Censored feedback

Send small orders will keep the actual liquidity hidden



Combinatorial MAB [Chen et al., 2013]

Solving the DPSOR problem by framing it as a CMAB



- We are in a CMAB setting, where the superarms are all the combinations of A_{kn}^t which satisfy the following constraint:

$$\sum_{k=1}^{K}\sum_{n=1}^{N} A_{kn}^t = V$$

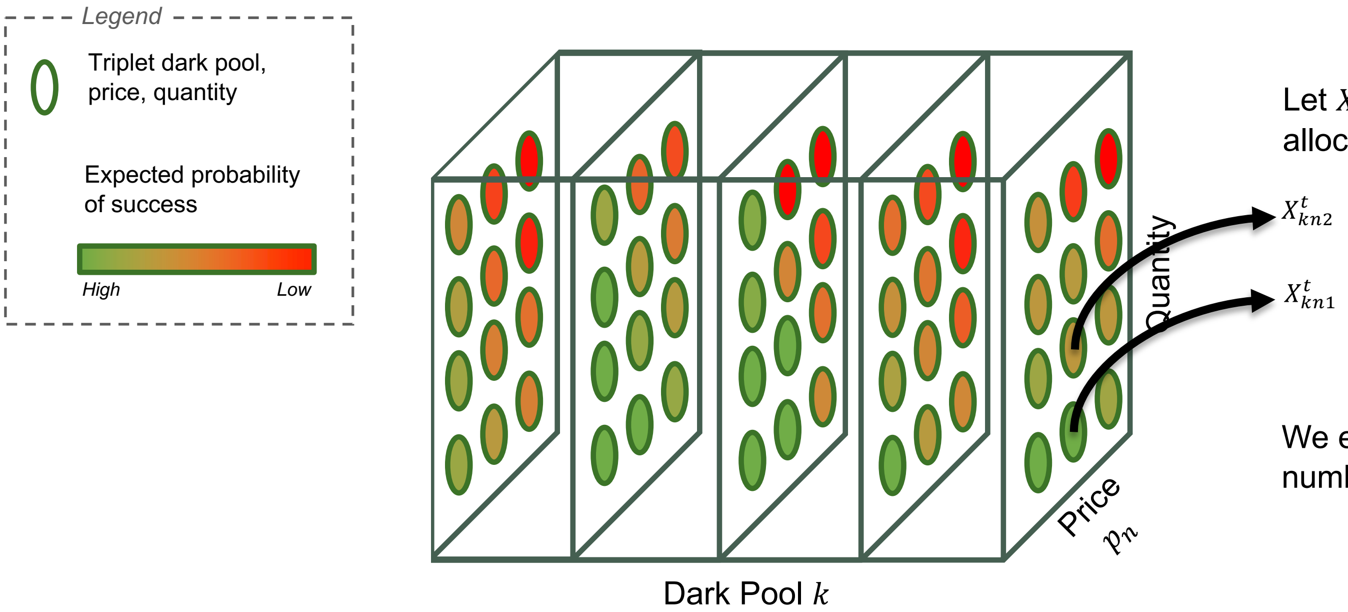
- We want to minimize pseudo-regret w.r.t. the expected dollar value of the optimal superarm r^*

$$Reg_t(\mathfrak{U}) := t\, r^* - \sum_{h=1}^{t}\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}[r_{kn}^h]\mathbb{1}\{A_{nk}^h > 0\}\, p_n$$

$$r_{kn}^t = \min\{A_{kn}^t, s_{kn}^t\}$$

Estimating liquidity

Count the number of successes and failures of each triplet

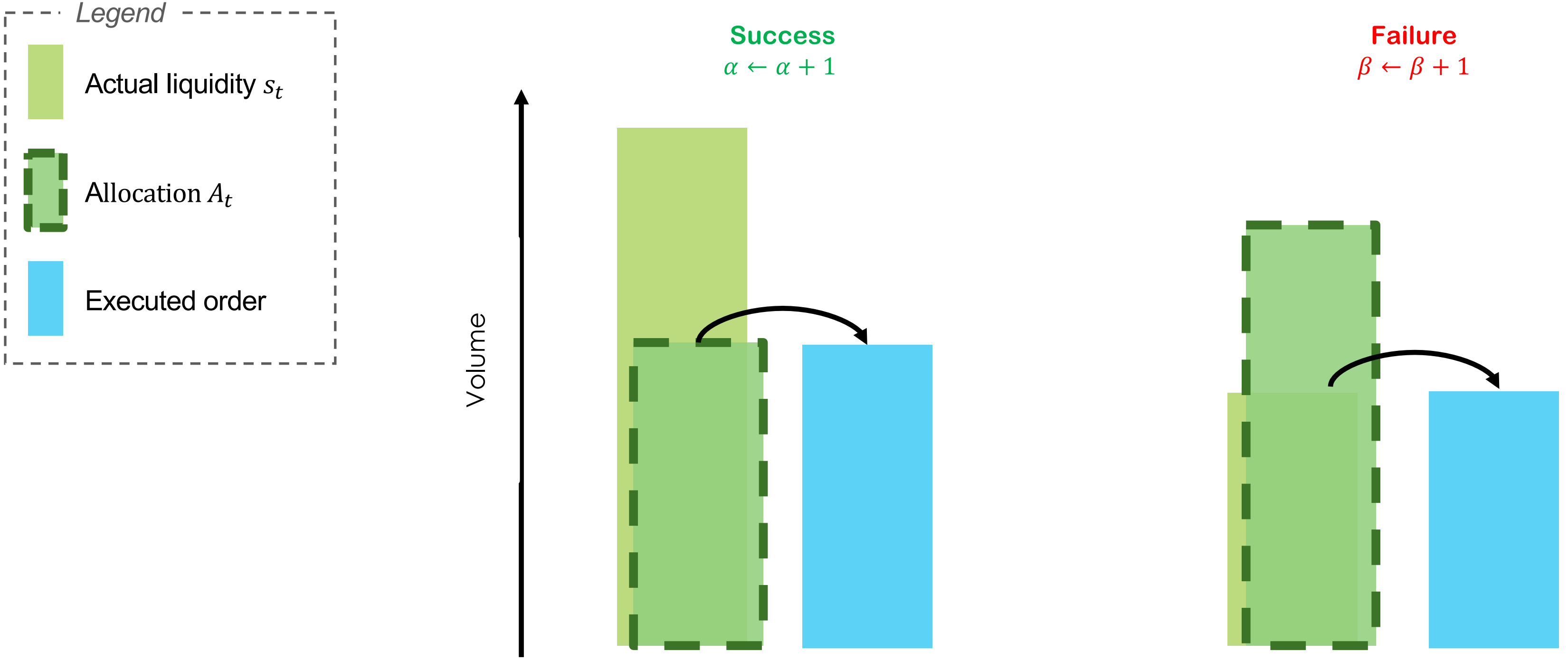


Let X^t_{knv} the probability that a specific allocation is successful

We estimate X^t_{knv} by counting the number of successes and failures

Counting successes α and failures β

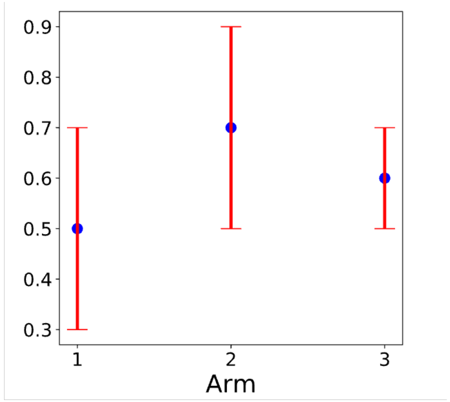
Using successes and failures to estimate liquidity



DP-CMAB Algorithm – θ Selection

Using successes and failures to estimate liquidity

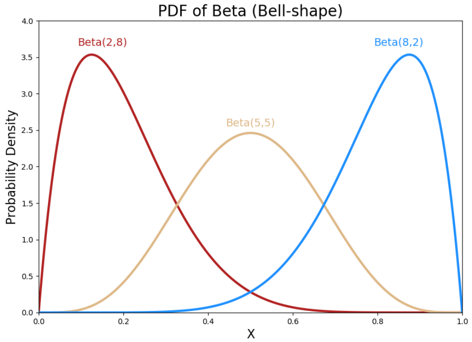
DP-CUCB



Mean and uncertainty

$$\theta^t_{knv} = v\left(\underbrace{\frac{\alpha^t_{knv} - 1}{\alpha^t_{knv} + \beta^t_{knv} - 2} + \sqrt{\frac{2\log(t)}{\alpha^t_{knv} + \beta^t_{knv} - 2}}}_{X^t_{knv}}\right)$$

DP-TS



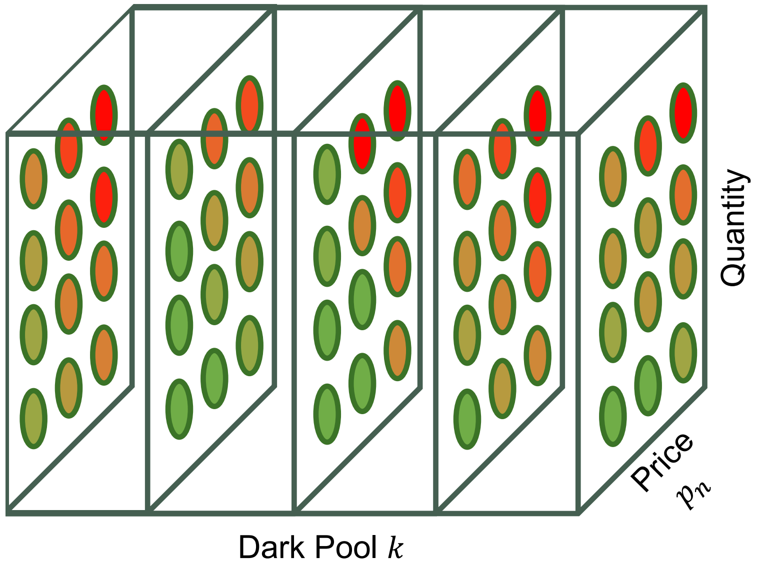
Sample from the Beta distribution

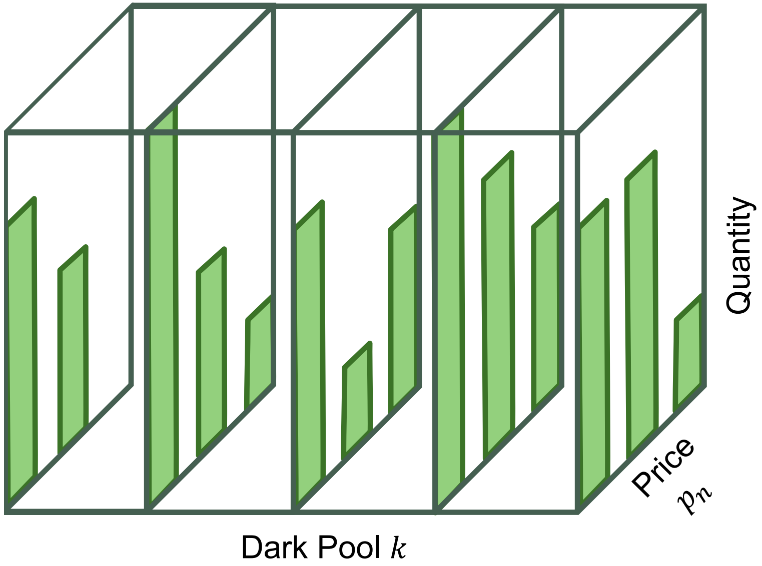
$$\theta^t_{knv} \sim v\, \underbrace{Beta\left(\alpha^t_{knv}, \beta^t_{knv}\right)}_{X^t_{knv}}$$

Translating liquidity to allocation

Using an optimization oracle and dynamic programming to decide the allocation matrix

$$\boldsymbol{\theta}_t = \boldsymbol{v}\boldsymbol{X}_t \qquad \mathrm{Opt}(\boldsymbol{\theta}_t) \rightarrow \boldsymbol{A}_t \qquad \boldsymbol{A}_t$$





DP CMAB High Level Pseudo Code

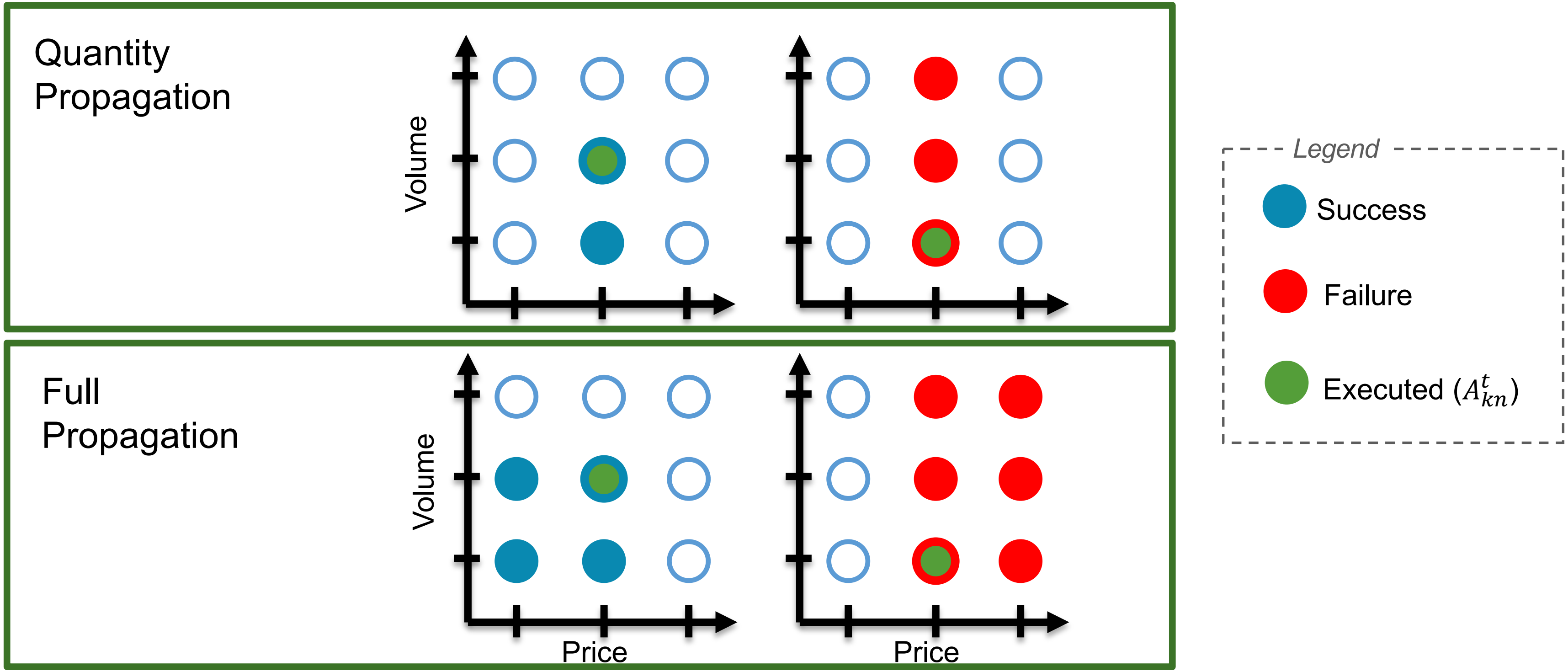
At each round t:

- Calculate the liquidity estimate $\boldsymbol{\theta_t}$ using $\boldsymbol{\alpha_t}, \boldsymbol{\beta_t}$ and the appropriate update CUCB or TS
- Calculate the action matrix $\boldsymbol{A_t} \leftarrow \text{Opt}(\boldsymbol{\theta_t})$
- Play allocation $\boldsymbol{A_t}$
- Receive feedbacks $\boldsymbol{r_t}$ from played arms
- Calculate the parameters $\boldsymbol{\alpha_{t+1}}$ and $\boldsymbol{\beta_{t+1}}$

Dark-Pool Smart Order Routing: a Combinatorial Multi-armed Bandit Approach, ICAIF 2022

Can we do better?

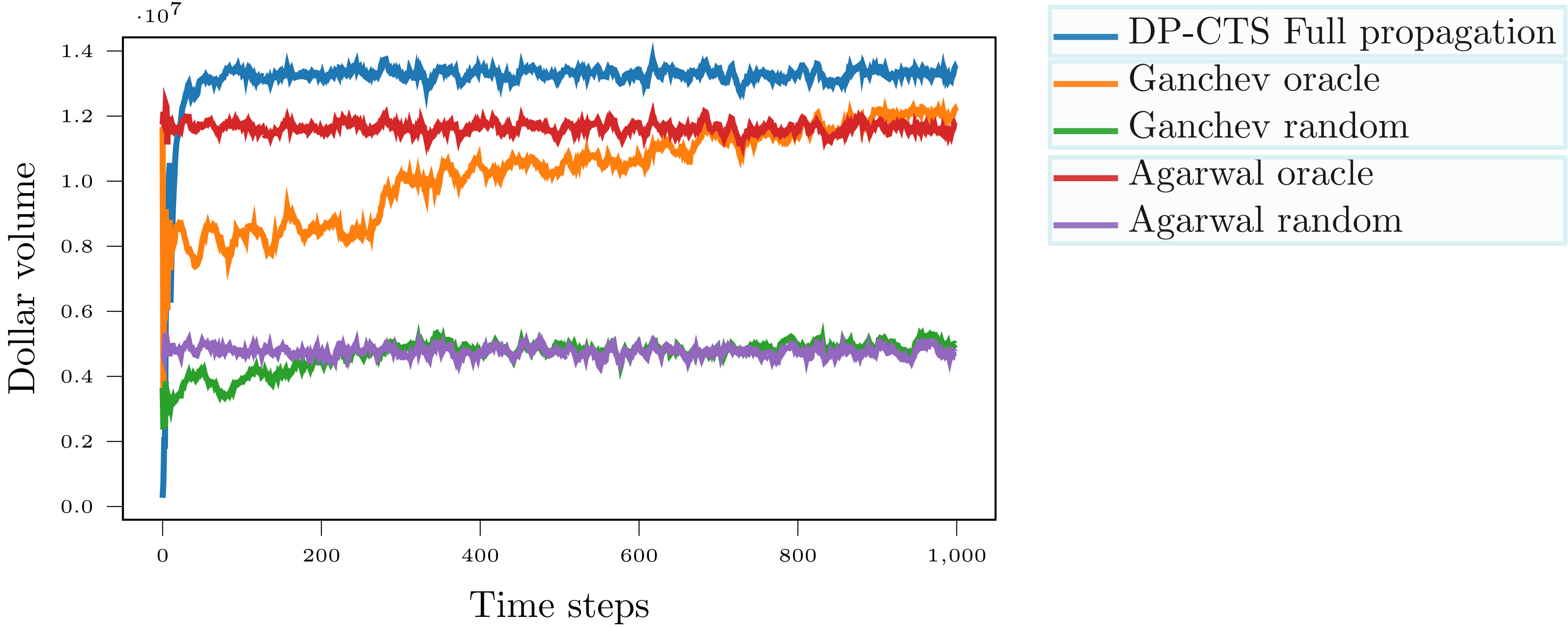
Using domain knowledge to improve learning



Dark-Pool Smart Order Routing: a Combinatorial Multi-armed Bandit Approach, ICAIF 2022

Experimental results – Dollar volume

We want to maximise dollar volume (volume times price)



Dark-Pool Smart Order Routing: a Combinatorial Multi-armed Bandit Approach, ICAIF 2022

Algorithms in the Financial Markets

Dealer Markets

Market structure

Characteristics

- Dealer markets
- Request for quote
- High frequency job

Dealers quotes for a bond

PCS	Firm Name	Bid Px / Ask Px	Bid Yld / Ask Yld	BSz... x AS...	Time ↓
	Total Axe Size			205 x	
CBBT	FIT COMPOSITE	91.844 / 91.868	1.833 / 1.830	x	11:59
BVAL	BVAL (Score: 10)	91.624 / 91.640	1.858 / 1.856	x	09:00
	Last Trade	91.856	--	7.7	11:34
NOMX	NOMURA INTL PLC LDN	91.848 / 91.882	1.832 / 1.828	50 x 10	11:59
MZHO	MIZUHO INTL	91.8400 / 91.8928	1.832 / 1.827	5 x 10	11:59
IMIG	INTESA SANPAOLO IMIG	91.795 / 91.895	1.838 / 1.827	10 x 10	11:59
MSEG	MORGAN STANLEY LOND	91.847 / 91.922	1.832 / 1.823	3 x 10	11:59
BSGB	SANTANDER Ex	91.848 / 91.918	1.831 / 1.824	25 x 5	11:59
HVGO	UniCredit Bank AG	91.800 / 91.919	1.837 / 1.824	5 x 5	11:59
DZBK	DZ BANK	91.796 / 91.916	1.838 / 1.824	5 x 5	11:59
HELA	HELABA AUTO EX	91.781 / 91.930	1.840 / 1.823	5 x 5	11:59
DEKA	DEKABANK	91.806 / 91.906	1.837 / 1.825	2.5 x 2.5	11:59
BPEG	BNP PARIBAS EURO G...	91.863 / 91.937	1.830 / 1.822	2 x 2	11:59

Reinforcement Learning for Market Making

Problem definition and MDP description

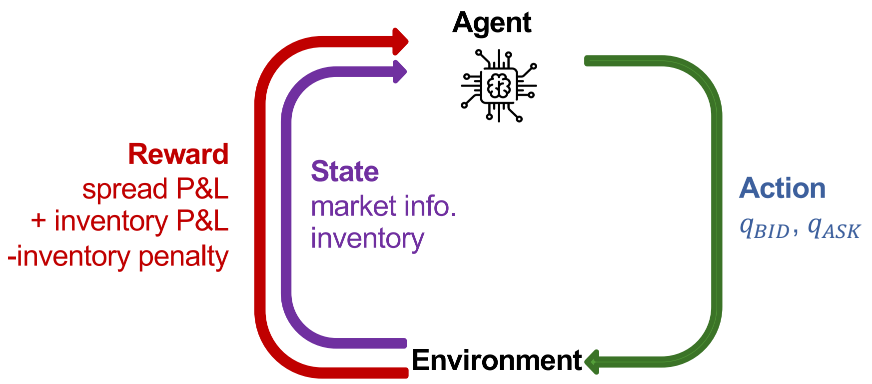
Market Making

Definition

- Continuously quote bid and ask prices in order to maximize P&L with minimizing inventory

MDP

- **State:** market information (prices, volumes etc.), current inventory
- **Action:** bid price, ask price
- **Reward:** spread P&L + inventory P&L – inventory penalty

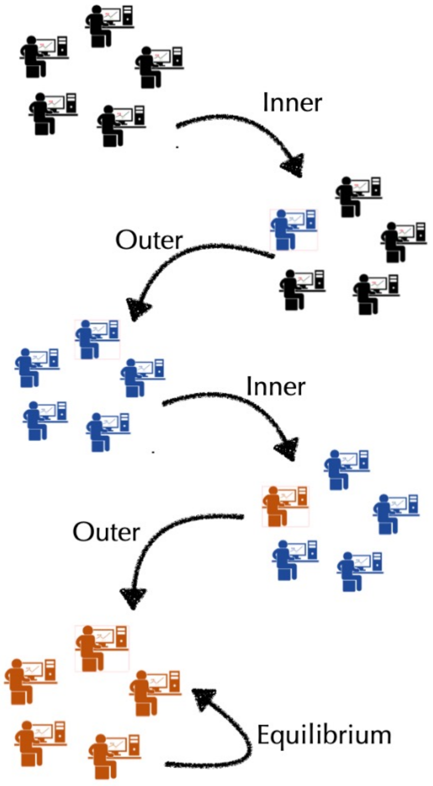


Learning in Mean-Field Games

Learning a competitive strategy

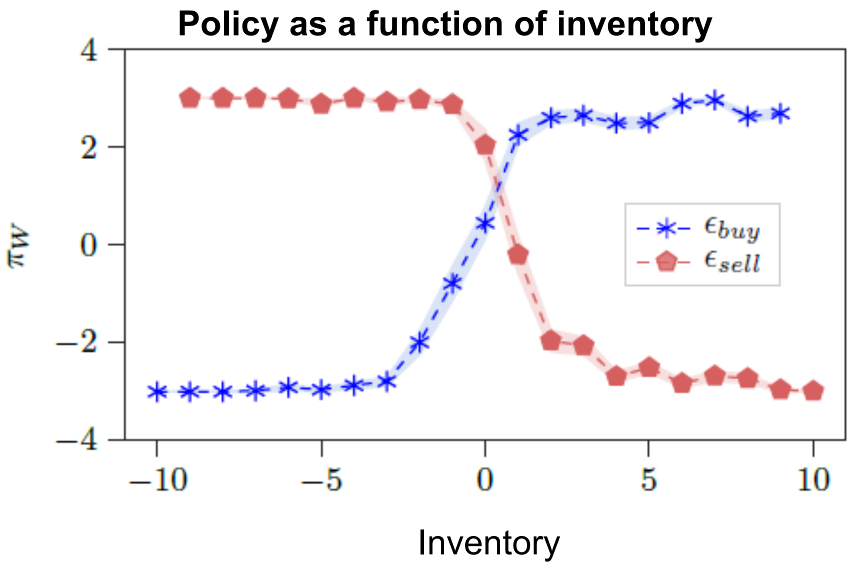
Definitions and notation

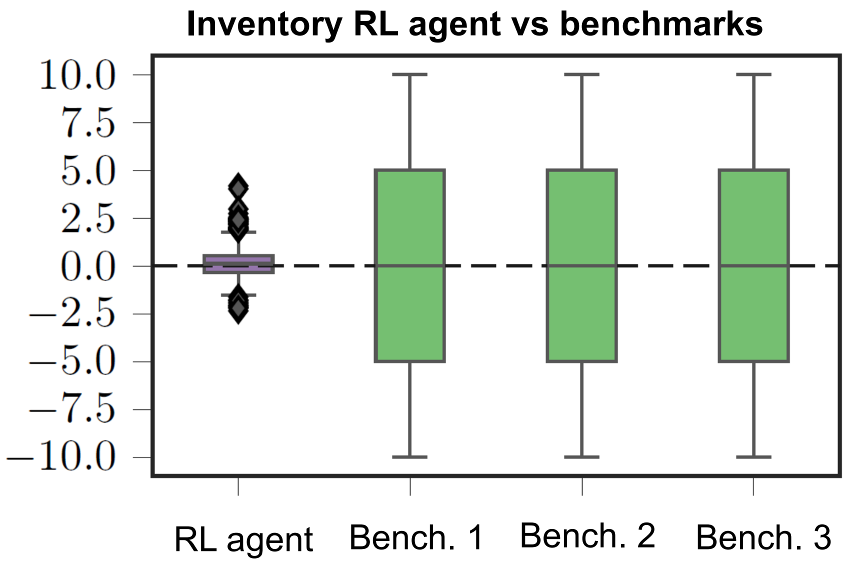
- Assume homogeneity/anonymity
- Mean-Field $\mathcal{L}$ represents players' distrubtion
- π is the policy
- Nash Equilibrium



Experimental Results

Policy and inventory in a simulated environment





Dealer Markets: A Reinforcement Learning Mean Field Game Approach, SSRN 2022

Algorithms in the Financial Markets

1 Algorithmic Trading

2 Reinforcement Learning

3 Quantitative Trading

4 Online Portfolio Optimization

5 Optimal Execution

6 Smart Routing with CMABs

7 Market Making with MFGs

8 Hedging with Risk Averse RL

Reinforcement Learning for Option Hedging

Problem definition and MDP description

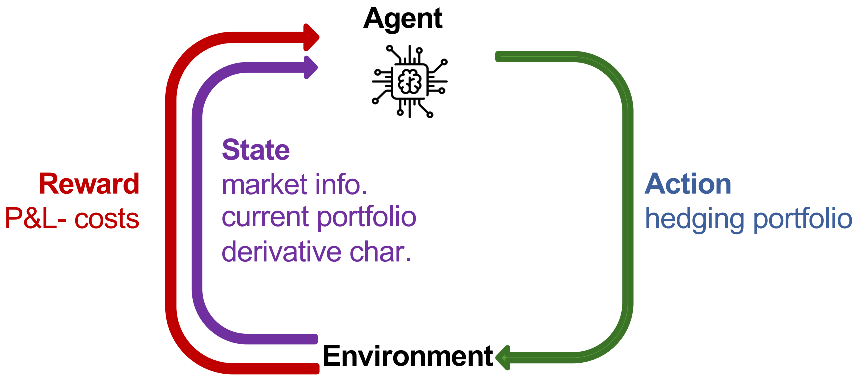
Option Hedging

Definition

- Choose, for each timestep, the hedging portfolio so to minimize the price variations caused by the option
- A risk averse objective is necessary

MDP

- **State:** market prices, hedging portfolio, option details
- **Action:** hedging portfolio
- **Reward:** $P\&L_c$ – $P\&L_h$ – transaction costs



Risk aversion in RL

Different approaches to risk aversion

Reward volatility

$$\nu_\pi^2 = (1-\gamma) \underset{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}}{\mathbb{E}} \left[\sum_{t=0}^{\infty} \gamma^t \left(\mathcal{R}(s_t, a_t) - J_\pi \right)^2 \right]$$

Return variance

$$\sigma_\pi^2 := \underset{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}}{\mathbb{E}} \left[\left(\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) - \frac{J_\pi}{1-\gamma} \right)^2 \right]$$

Risk-Averse Trust Region Optimization for Reward-Volatility Reduction, IJCAI 2020

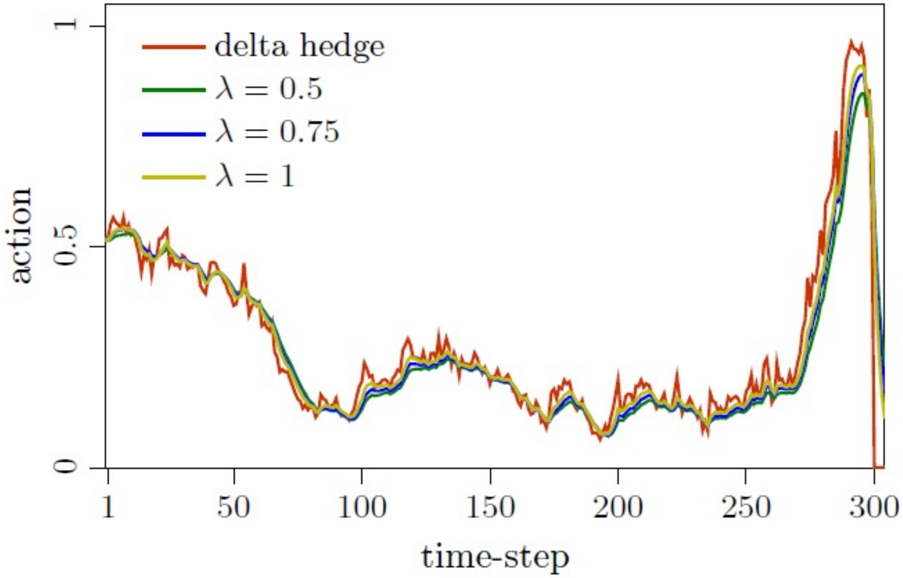
Experimental Results (1/2)

Hedging a call option, single scenario

Characteristics

- Objective: $J - \lambda\nu^2$
- Simulated market
- Hedge a vanilla call option with a TTM of 60 days
- We are considering transaction costs

Plot of policy



Option Hedging with Risk Averse Reinforcement Learning, ICAIF 2020

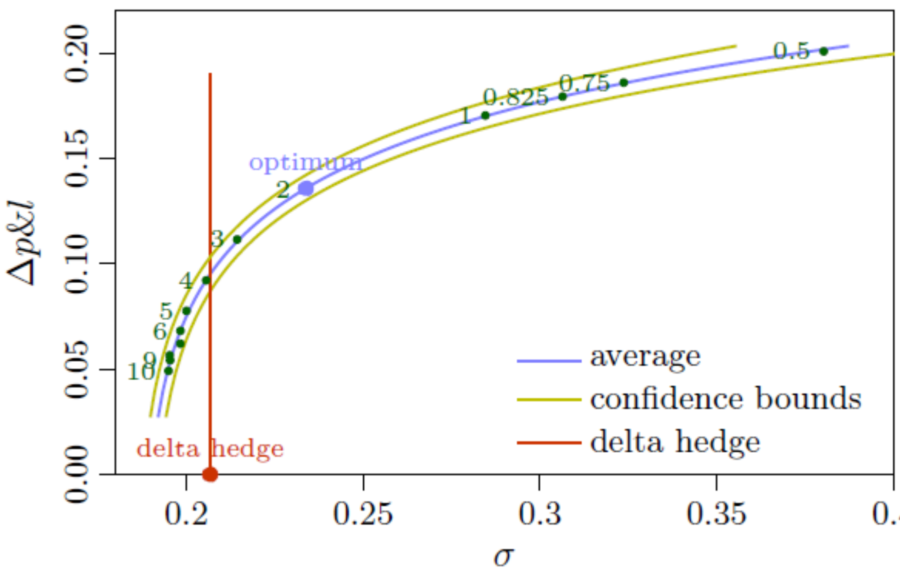
Experimental Results (2/2)

Hedging a call option, average results

Characteristics

- Simulated market
- Hedge a vanilla call option with a TTM of 60 days
- $\Delta p\&l$ is the difference between the return of the strategy and that of the delta hedge
- σ is the $p\&l$ volatility

Pareto frontier



Option Hedging with Risk Averse Reinforcement Learning, ICAIF 2020

Reinforcement Learning in the Capital Markets

Edoardo Vittori
edoardo.vittori@intesasanpaolo.com