AARHUS
UNIVERSITET

Machine learning approaches for high-frequency financial data analysis

Alexandros Iosifidis
Aarhus University, Department of Engineering, ECE
alexandros.iosifidis@eng.au.dk

High-frequency financial data analysis

From an Eng/CS and Machine Learning point of view

High-frequency financial data analysis

From an Eng/CS and Machine Learning point of view:

Interesting research topic for three reasons

High-frequency financial data analysis

Interesting from an Eng/CS and Machine Learning point of view for 3 reasons:

1. Lots of data! à perfect for methodologies we have been working on for many decades

High-frequency financial data analysis

Interesting from an Eng/CS and Machine Learning point of view for 3 reasons:

1. Lots of data! à perfect for methodologies we have been working on for many decades
2. Lots of data! à potential to reveal/observe/model patterns and trends

High-frequency financial data analysis

Interesting from an Eng/CS and Machine Learning point of view for 3 reasons:

1. Lots of data! à perfect for methodologies we have been working on for many decades
2. Lots of data! à potential to reveal/observe/model patterns and trends
3. Lots of data! à big challenges in data management, processing and analysis

High-frequency financial data analysis

Interesting from an Eng/CS and Machine Learning point of view for 3 reasons:

1. Lots of data! à perfect for methodologies we have been working on for many decades
2. Lots of data! à potential to reveal/observe/model patterns and trends
3. Lots of data! à big challenges in data management, processing and analysis

Sexy topic: it is expected to receive much attention in ML community

Definitions and Concepts

Financial Exchanges

- Exchanges are organized marketplaces where securities (like stocks), commodities and derivatives (among other financial instruments) are traded
- Investors need to agree upon a price before exchanging an asset, and this needs to be done centrally so the transactions can be monitored
- Exchanges use what is called a Limit Order Book to achieve this coordination

Definitions and Concepts

Limit Order Book (LOB)

- The Limit Order Book is a transparent trading system that matches customer orders (bids and asks) using a “price first time second” priority basis
- When an order is submitted it contains 3 attributes:
  - Whether it is an ask (sell) or bid (buy) order
  - The price limit p(t)
  - The size of the order v(t)
- The order book has two sides, the bid side, containing buy orders, and the ask side, containing sell orders

Definitions and Concepts

Message List (or Message Book)

Timestamp	Id	Price	Quantity	Event	Side
1275386347944	6505727	126200	400	Cancellation	Ask
1275386347981	6505741	126500	300	Submission	Ask
1275386347981	6505741	126500	300	Cancellation	Ask
1275386348070	6511439	126100	17	Execution	Bid
1275386348070	6511439	126100	17	Submission	Bid
1275386348101	6511469	126600	300	Cancellation	Ask

time

Definitions and Concepts

Limit Order Book (LOB)

- When an order is submitted to the Limit Order Book, its price and volume is checked against existing orders of the opposite side
- Whenever a bid order price exceeds an ask order price $p_b^{(i)}(t) > p_a^{(j)}(t)$, they “annihilate”, executing the orders and exchanging the traded assets between the investors.
- If there is not enough volume to fully fill the size of an order at the requested price limit, the leftover size is added to the Order Book

Definitions and Concepts

Limit Order Book

depth of L levels

time

			Level 1				Level 2				...
			Ask		Bid		Ask		Bid		
Timestamp	Mid-price	Spread	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity	
1275386347944	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348101	126050	100	126100	291	126000	2800	126200	300	125900	1120	...

Ask: $p_a^{(l)}(t) \leq p_a^{(l+1)}(t)$

Bid: $p_b^{(l)}(t) \geq p_b^{(l+1)}(t)$

Definitions and Concepts

Limit Order Book – Mid-price

			Level 1				Level 2				...
			Ask		Bid		Ask		Bid		
Timestamp	Mid-price	Spread	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity	
1275386347944	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348101	126050	100	126100	291	126000	2800	126200	300	125900	1120	...

$$p_t = \frac{p_a^{(1)}(t) + p_b^{(1)}(t)}{2}$$

Definitions and Concepts

Limit Order Book Value

- The limit order that is executed last is what determines the “price” of an asset:
  - For example, AAPL stock price at 150USD means that the last limit order that was executed had a limit price of 150USD
- Analysis of transaction history of the LOB can lead to information about possible future movements of the price (in our case mid-price)

ML for LOB data

Data representation

› LOB data can be naturally described as an D-dimensional time series, where D = 4L

			Level 1				Level 2				...
			Ask		Bid		Ask		Bid		
Timestamp	Mid-price	Spread	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity	
1275386347944	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348101	126050	100	126100	291	126000	2800	126200	300	125900	1120	...

ML for LOB data

Data representation

› LOB data can be naturally described as an D-dimensional time series, where D = 4L

depth of L levels

time

			Level 1				Level 2				...
			Ask		Bid		Ask		Bid		
Timestamp	Mid-price	Spread	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity	
1275386347944	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348101	126050	100	126100	291	126000	2800	126200	300	125900	1120	...

ML for LOB data

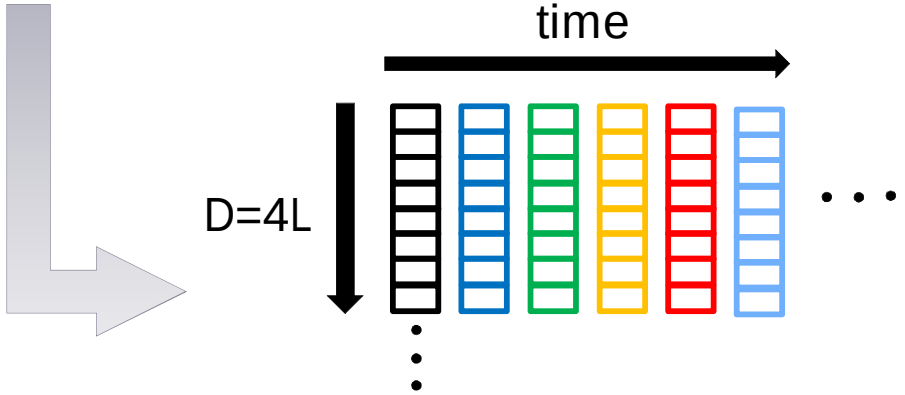
Data representation

› LOB data can be naturally described as an D-dimensional time series, where D = 4L

depth of L levels

time

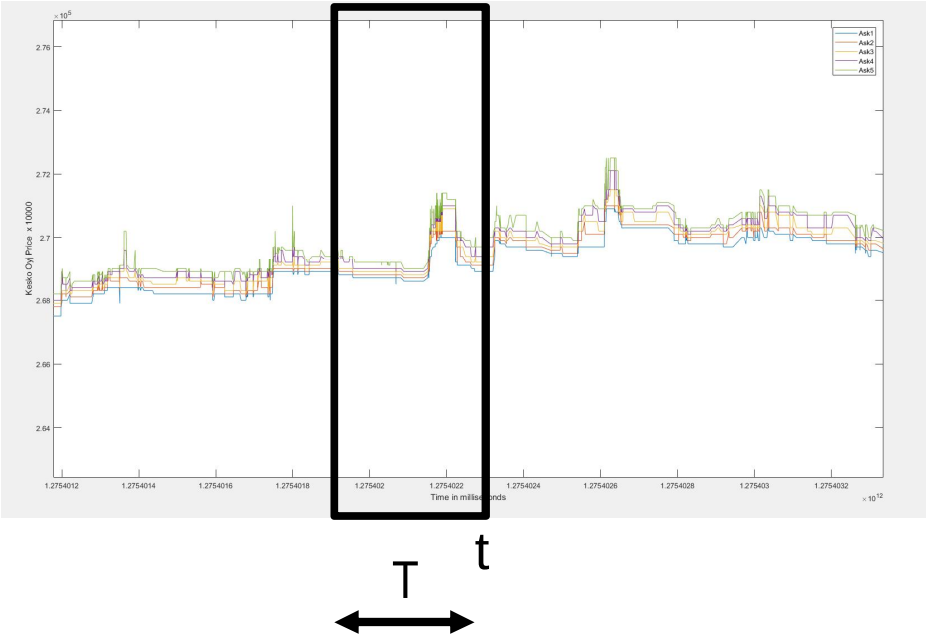
			Level 1				Level 2				...
			Ask		Bid		Ask		Bid		
Timestamp	Mid-price	Spread	Price	Quantity	Price	Quantity	Price	Quantity	Price	Quantity	
1275386347944	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386347981	126200	200	126300	300	126100	17	126400	4765	126000	2800	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348070	126050	100	126100	291	126000	2800	126200	300	125900	1120	...
1275386348101	126050	100	126100	291	126000	2800	126200	300	125900	1120	...



ML for LOB data

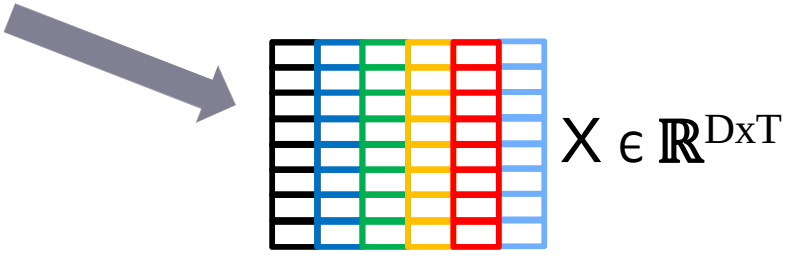
Data representation

- LOB data can be naturally described as an D-dimensional time series, where D = 4L
- Thus, LOB data corresponding to a time window with T events can be represented by a matrix with D x T elements



t à Current time instance

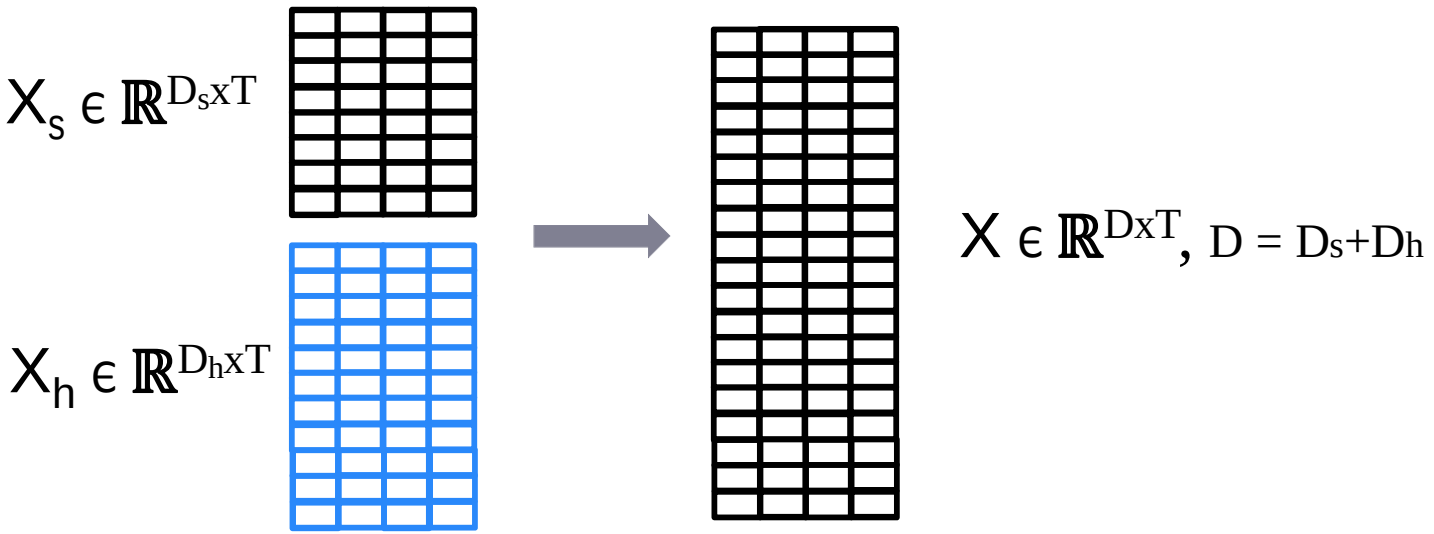
T à Time window (time period or number of orders)



ML for LOB data

Data representation

› The above LOB data representation can be augmented by appending handcrafted data representations, e.g. describing short/long-term trends.



ML for LOB data

Additional features

Feature Set	Description	Details
Basic	$u_1 = \{P_i^{ask}, V_i^{ask}, P_i^{bid}, V_i^{bid}\}_{i=1}^{n}$	10(=n)-level LOB Data
Time-Insensitive	$u_2 = \{(P_i^{ask} - P_i^{bid}), (P_i^{ask} + P_i^{bid})/2\}_{i=1}^{n}$	Spread & Mid-Price
	$u_3 = \{P_n^{ask} - P_1^{ask}, P_1^{bid} - P_n^{bid}, \lvert P_{i+1}^{ask} - P_i^{ask}\rvert, \lvert P_{i+1}^{bid} - P_i^{bid}\rvert\}_{i+1}^{n}$	Price differences
	$u_4 = \left\{\frac{1}{n}\sum_{i=1}^{n} P_i^{ask}, \frac{1}{n}\sum_{i=1}^{n} P_i^{bid}, \frac{1}{n}\sum_{i=1}^{n} V_i^{ask}, \frac{1}{n}\sum_{i=1}^{n} V_i^{bid}\right\}$	Price & Volume means
	$u_5 = \left\{\sum_{i=1}^{n}(P_i^{ask} - P_i^{bid}), \sum_{i=1}^{n}(V_i^{ask} - V_i^{bid})\right\}$	Accumulated differences
Time-Sensitive	$u_6 = \left\{dP_i^{ask}/dt, dP_i^{bid}/dt, dV_i^{ask}/dt, dV_i^{bid}/dt\right\}_{i=1}^{n}$	Price & Volume derivation
	$u_7 = \left\{\lambda_{\Delta t}^1, \lambda_{\Delta t}^2, \lambda_{\Delta t}^3, \lambda_{\Delta t}^4, \lambda_{\Delta t}^5, \lambda_{\Delta t}^6\right\}$	Average intensity per type
	$u_8 = \left\{\mathbf{1}_{\lambda_{\Delta t}^1 > \lambda_{\Delta T}^1}, \mathbf{1}_{\lambda_{\Delta t}^2 > \lambda_{\Delta T}^2}, \mathbf{1}_{\lambda_{\Delta t}^3 > \lambda_{\Delta T}^3}, \mathbf{1}_{\lambda_{\Delta t}^4 > \lambda_{\Delta T}^4}, \mathbf{1}_{\lambda_{\Delta t}^5 > \lambda_{\Delta T}^5}, \mathbf{1}_{\lambda_{\Delta t}^6 > \lambda_{\Delta T}^6}\right\}$	Relative intensity comparison
	$u_9 = \{d\lambda^1/dt, d\lambda^2/dt, d\lambda^3/dt, d\lambda^4/dt, d\lambda^5/dt, d\lambda^6/dt\}$	Limit activity accelaration

A.N. Kercheval and Y. Zhang, “Modeling high-frequency limit order book dynamics with support vector machines”, Quantitative Finance, 2015

ML for LOB data

Analysis problems

- Mid-price (direction) prediction.
- Mid-price jump detection
- Mid-price jump prediction
- ...

ML for LOB data

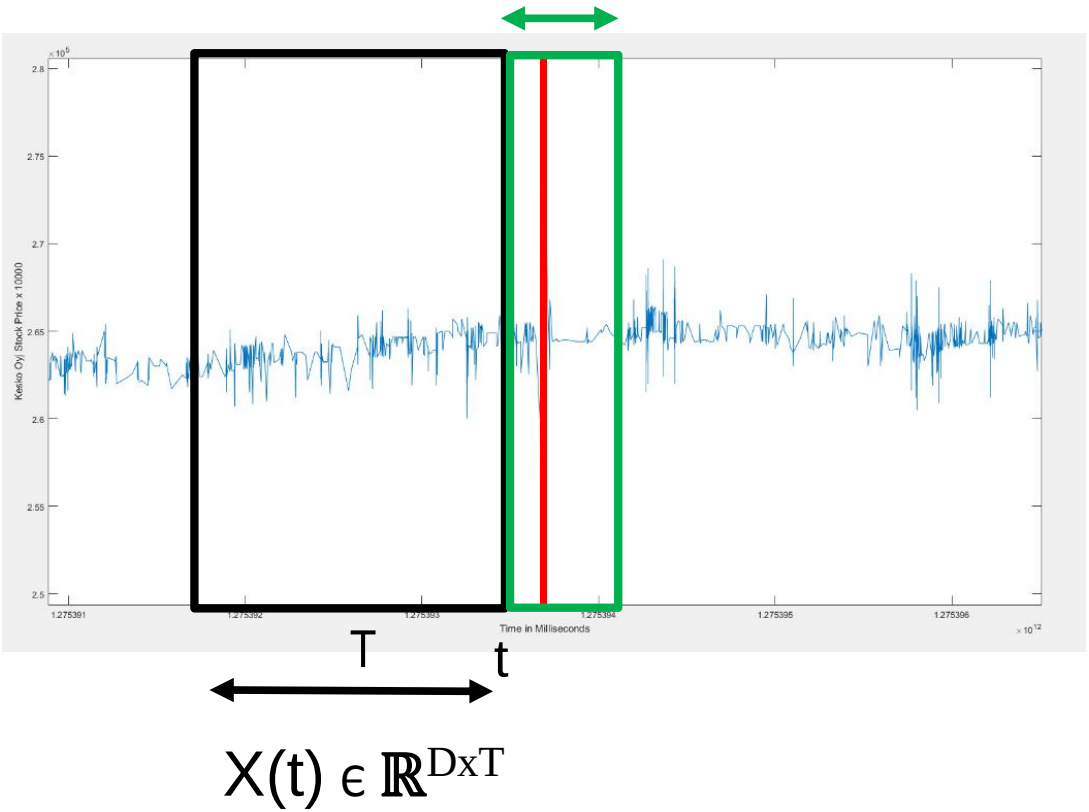
Analysis problems

- Mid-price (direction) prediction.
- Mid-price jump detection
- Mid-price jump prediction
- ...

ML for LOB data

Mid-price direction prediction

- Given X(t) predict if p_{t+1} will:
  - Increase
  - Stay stationary
  - Decrease



ML for LOB data

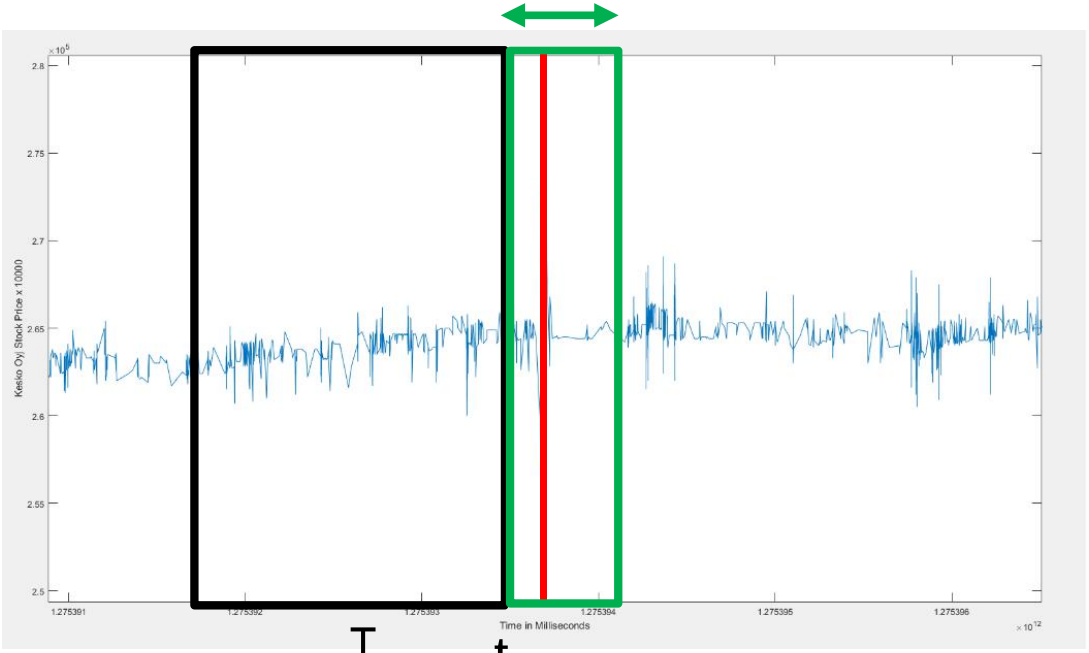
Mid-price direction prediction

- Given X(t) predict if p_{t+1} will:
  - Increase
  - Stay stationary
  - Decrease
- In order to predict trends, we predict changes in a smoothed version of:

$$\Delta p_t = (p_{t+1} - p_t) / \alpha$$

scaling factor

Δt à prediction horizon



$$X(t) \in \mathbb{R}^{DxT}$$

ML for LOB data

Mid-price direction prediction as a regression problem

- Based on the above, mid-price prediction can be transformed to a Tensor-to-Vector regression problem:

ML for LOB data

Mid-price direction prediction as a regression problem

- Based on the above, mid-price prediction can be transformed to a Tensor-to-Vector regression problem:

  X(t) à y(t), where $y(t) \in \mathbb{R}^3$ is a (probability-like) target vector denoting the desired class:

  - $y(t) = [1\ 0\ 0]^T$, if $\Delta p_t < -1$
  - $y(t) = [0\ 1\ 0]^T$, if $-1 \leq \Delta p_t \leq 1$
  - $y(t) = [0\ 0\ 1]^T$, if $\Delta p_t > 1$

ML for LOB data

Mid-price direction prediction as a regression problem

- Based on the above, mid-price prediction can be transformed to a Tensor-to-Vector regression problem:

  X(t) à y(t), where $y(t) \in \mathbb{R}^3$ is a (probability-like) target vector denoting the desired class:

  - $y(t) = [1\ 0\ 0]^T$, if $\Delta p_t < -1$
  - $y(t) = [0\ 1\ 0]^T$, if $-1 \leq \Delta p_t \leq 1$
  - $y(t) = [0\ 0\ 1]^T$, if $\Delta p_t > 1$

  } Competitive training

ML for LOB data

Mid-price direction prediction as a regression problem

- Based on the above, mid-price prediction can be transformed to a Tensor-to-Vector regression problem:

  $X(t)$ à $y(t)$, where $y(t) \in \mathbb{R}^3$ is a (probability-like) target vector denoting the desired class:

  - $y(t) = [1\ 0\ 0]^T$, if $\Delta p_t < -1$
  - $y(t) = [0\ 1\ 0]^T$, if $-1 \leq \Delta p_t \leq 1$
  - $y(t) = [0\ 0\ 1]^T$, if $\Delta p_t > 1$

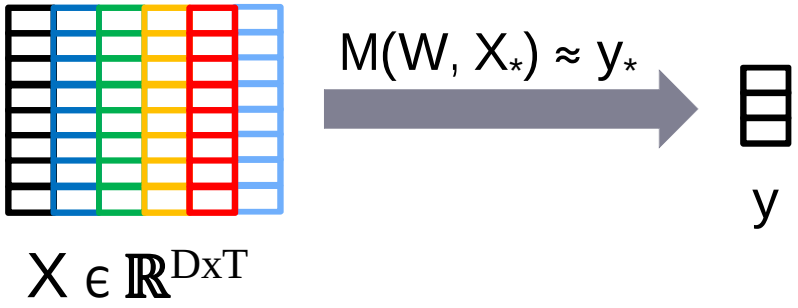
  } Competitive training

- Regression model $M(W, \cdot)$ such that: $M(W, X_*) \approx y_*$
- The model's parameters W are estimated by training on known examples

Mid-price direction prediction

Tensor-to-Vector regression:

› Time information plays its own role in this regression



Mid-price direction prediction

Three types of Tensor-to-Vector regression:

- Tensor-based regression
- Convolutional Neural Network (CNN)-based regression
- Recurrent Neural Network (RNN)-based regression

D. Thanh, M. Magris, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Tensor Representation in High-Frequency Financial Data for Price Change Prediction”, IEEE SSCI, 2017

A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Forecasting Stock Prices from the Limit Order Book using Convolutional Neural Networks", CBI, 2017

A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Using Deep Learning to Detect Price Change Indications in Financial Markets", EUSIPCO, 2017

Mid-price direction prediction

Tensor-based regression

› Estimation of regression parameters by minimizing:

N training samples

$$J(\mathbf{W}_1, \mathbf{w}_2) = \sum_{i=1}^{N} s_i \|\mathbf{W}_1^T \mathcal{X}_i \mathbf{w}_2 - \mathbf{y}_i\|_F^2 + \lambda_1 \|\mathbf{W}_1\|_F^2 + \lambda_2 \|\mathbf{w}_2\|_F^2$$

Feature weighting

Time weighting

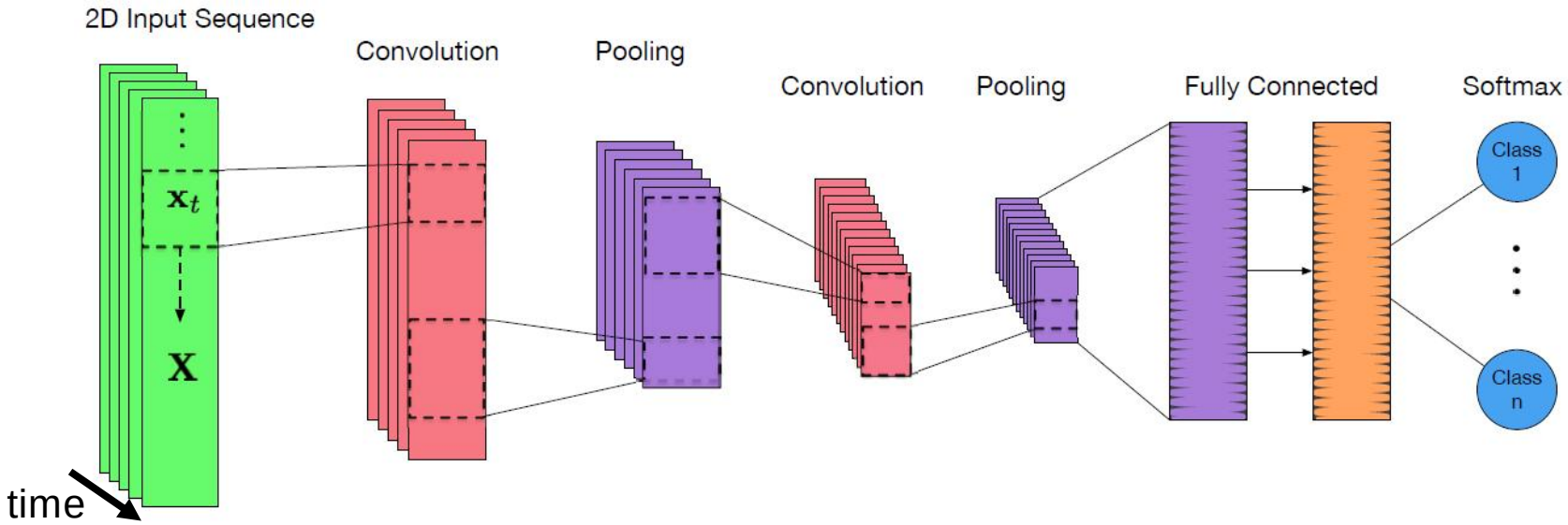
› Iterative optimization, until convergence

D. Thanh, M. Magris, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Tensor Representation in High-Frequency Financial Data for Price Change Prediction”, IEEE SSCI, 2017

Mid-price direction prediction

CNN-based regression

› X(t) is introduced to the network as a T-channel sequence:



› Iterative optimization, until convergence

A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Forecasting Stock Prices from the Limit Order Book using Convolutional Neural Networks", CBI, 2017

Mid-price direction prediction

CNN-based regression

› Estimation of regression parameters by maximizing:

$$\mathcal{L}(\mathbf{W}) = -\sum_{i=1}^{L} y_i \cdot \log \hat{y}_i$$

› Backpropagation-based optimization:

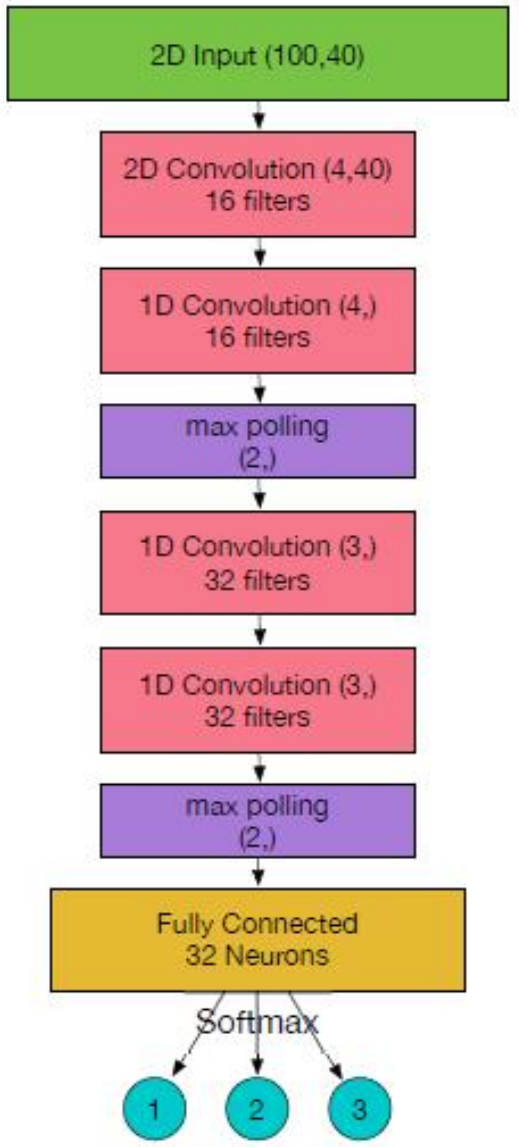
$$\mathbf{W}' = \mathbf{W} - \eta \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{W}}$$

A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Forecasting Stock Prices from the Limit Order Book using Convolutional Neural Networks", CBI, 2017

Mid-price direction prediction

CNN-based regression

› Adopted topology in our experiments

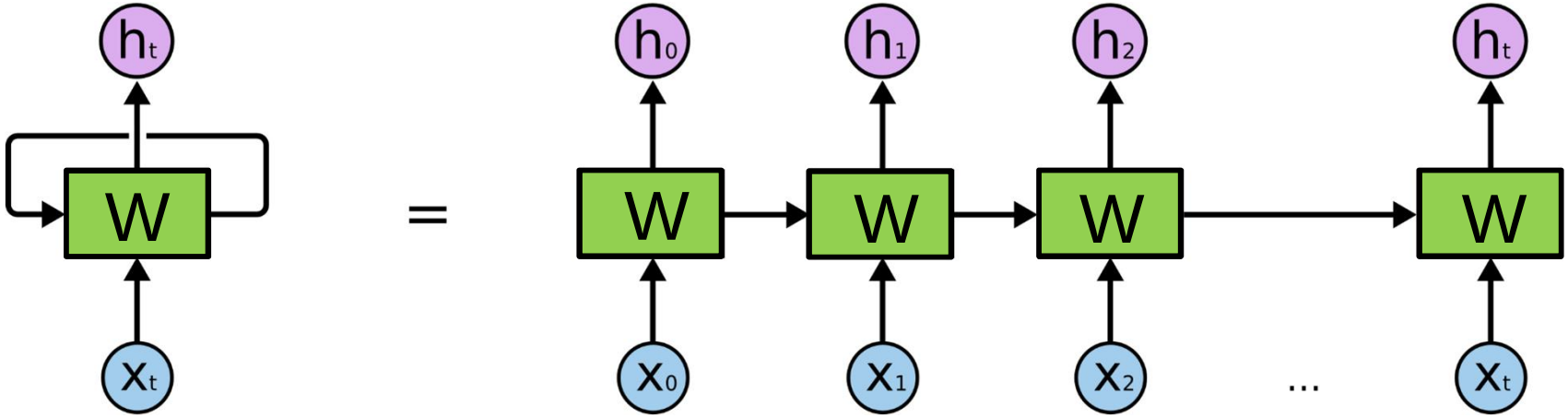


Mid-price direction prediction

RNN-based regression

$h_t \approx y_t$

› Neural network topology exploiting time progression:



A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Using Deep Learning to Detect Price Change Indications in Financial Markets", EUSIPCO, 2017

Mid-price direction prediction

RNN-based regression

- We used the Long-Short Term Memory (LSTM) network topology
- Estimation of regression parameters by maximizing:

$$\mathcal{L}(\mathbf{W}) = -\sum_{i=1}^{L} y_i \cdot \log \hat{y}_i$$

- Backpropagation-based optimization:

$$\mathbf{W}' = \mathbf{W} - \eta \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{W}}$$

A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Using Deep Learning to Detect Price Change Indications in Financial Markets", EUSIPCO, 2017

Some results

Dataset

- The stock data, provided by Nasdaq Nordic, come from the Finnish companies
  - Kesko Oyj
  - Outokumpu Oyj
  - Sampo, Rautaruukki
  - Wärtsilä Oyj
- The dataset is made up of 10 days for 5 different stocks and the total number of depth changes is 4.5 million
- The depth of the Orderbook we use has size of 10 for each side (bid/ask)
- The database is publicly available at:
  https://etsin.avointiede.fi/dataset/urn-nbn-fi-csc-kata20170601153214969115

A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Benchmark Dataset for Mid-Price Prediction of Limit Order Book data", arXiv:1705.03233, 2017

Some results

Performance (%) using the anchored forward cross-validation protocol

Method	Precision	Recall	F1
RR	43.30 ± 9.9	43.54 ± 5.20	42.52 ± 1.22
SLFN	49.60 ± 3.81	41.28 ± 4.04	38.24 ± 5.66
LDA	37.93 ± 6.00	45.80 ± 4.07	36.28 ± 1.02
MDA	44.21 ± 1.35	60.07 ± 2.10	46.06 ± 2.22
BoFs	39.26 ± 0.94	51.44 ± 2.53	36.28 ± 2.85
N-BoFs	42.28 ± 0.87	61.41 ± 3.68	41.63 ± 1.90
MTR	51.68 ± 7.54	40.81 ± 6.18	40.14 ± 5.26
WMTR	46.25 ± 1.90	51.29 ± 1.88	47.87 ± 1.91

Some results

Performance (%) using the 7 days for training and 3 days for testing

Method	Precision	Recall	F1
SVM	44.92	39.62	35.88
MLP	60.78	47.81	48.27
CNN	65.54	50.98	55.21
RNN	75.92	60.77	66.33

D. Thanh, M. Magris, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Tensor Representation in High-Frequency Financial Data for Price Change Prediction”, IEEE SSCI, 2017

A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Forecasting Stock Prices from the Limit Order Book using Convolutional Neural Networks", CBI, 2017

A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Using Deep Learning to Detect Price Change Indications in Financial Markets", EUSIPCO, 2017

N. Passalis, A. Tsantekidis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Time-series Classification using Neural Bag-of-Features, EUSIPCO, 2017

A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj and A. Iosifidis, "Benchmark Dataset for Mid-Price Prediction of Limit Order Book data", arXiv:1705.03233, 2017

Current work

Dataset

- US data corresponding to 375 days (April 8th 2014 to September 30th 2015), containing information of trades through NASDAQ

- Stocks:
  - AAPL
  - INTC
  - GOOG
  - MSFT
  - FB.

- The data is split into 7 sets of 50 days for training and the following 10 days for testing

Conclusion

Limit Order Book data analysis:

- Interesting ML research problem

Conclusion

Limit Order Book data analysis:

- Interesting ML research problem
- Basic models have shown promising results

Conclusion

Limit Order Book data analysis:

- Interesting ML research problem
- Basic models have shown promising results
- Need for ML models exploiting domain knowledge

Conclusion

Limit Order Book data analysis:

- Interesting ML research problem
- Basic models have shown promising results
- Need for ML models exploiting domain knowledge
- Many more problems to be attacked!

Conclusion

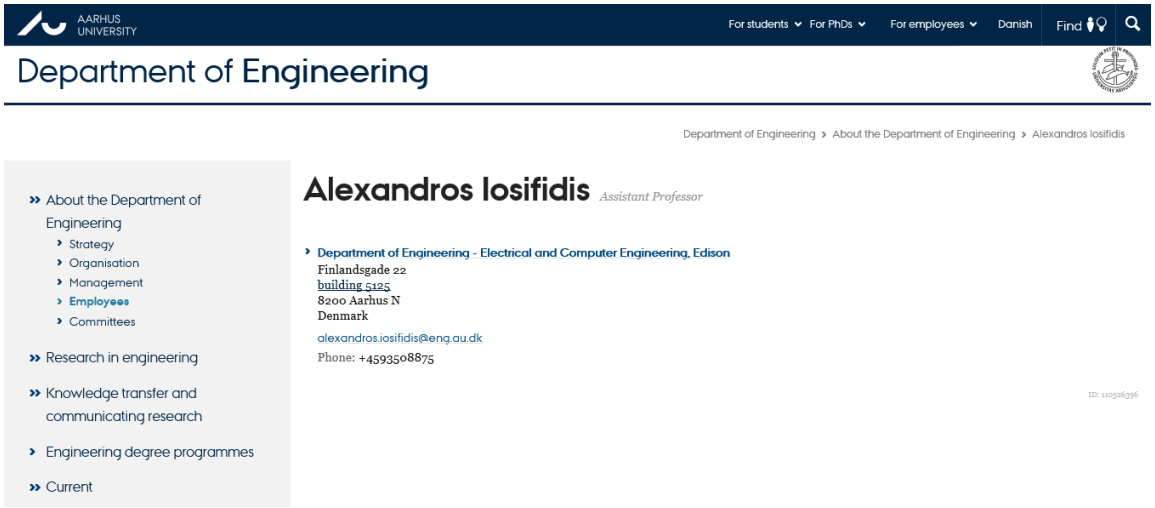
Limit Order Book data analysis:

- Interesting ML research problem
- Basic models have shown promising results
- Need for ML models exploiting domain knowledge
- Many more problems to be attacked!
- Need for more publicly available datasets!

Thank you for your attention!

More Info:

Pure profile in AU



former website in TUT

(to be moved to AU)