Dealer Markets: a Reinforcement Learning Mean Field Approach

Mar 2021

Martino Bernasconi de Luca, Edoardo Vittori, Francesco Trovò, Marcello Restelli

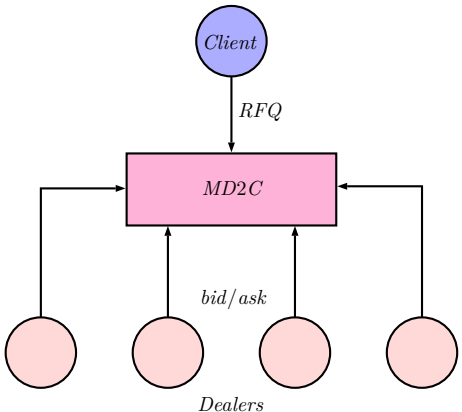
POLITECNICO
MILANO 1863

Outline

Game Theory and Market Making

Automation and Multi Dealer to Client Platforms

Multi Dealers to Client



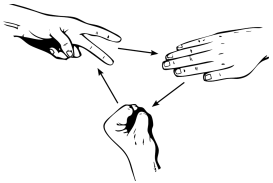
Objective

- Obtain a positive PnL with low risk by doing a large number of transactions, buying at the bid and selling at the ask
- Keep inventory low in order to have low capital requirements/risk

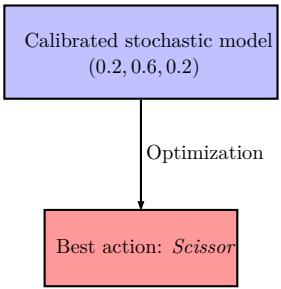
Approaches to Dealer Markets

- Classical approach (single agent)
  1. Model the behavior of the other agents
  2. Collect data on the behaviour of other dealers
  3. Fit the model parameters (see [Fermanian et al., 2016])
  4. Solve the optimization problem (see [Ganesh et al., 2019])
- Game Theoretic (multi-agent)
  1. Implement rules
  2. ~~Solve the optimization problem~~ Learn an *Equilibrium*

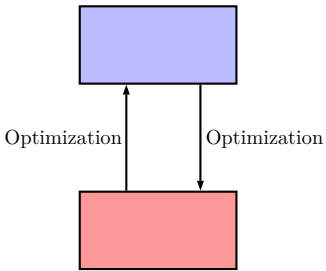
Motivating Example



Fixed Opponent

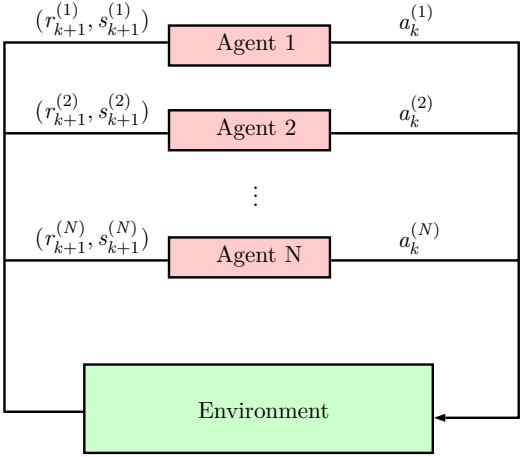


Learning Opponent/Opponents



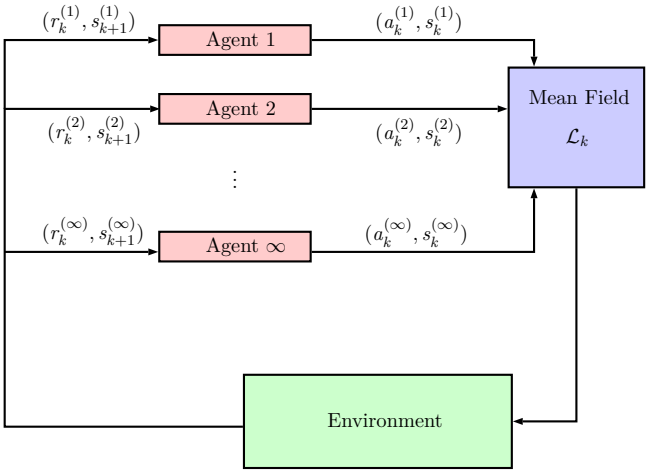
Mean Field Games

N-player stochastic games



- Combinatorial complexity in N
- Transitions depend on the aggregate action of all the players

Generalized Mean Field Games [Lasry and Lions, 2007]



- Assume homogeneity/anonymity
- Continuum number of players
- Transition depends only on the mean field $\mathcal{L}_k$

Notation for Generalized Mean Field Games

1. $\mathcal{A}$ is the action space
2. $\mathcal{S}$ is the state space
3. $\mathcal{L} \in \Delta(\mathcal{A} \times \mathcal{S})$
4. $\mu = \int_{\mathcal{A}} \mathcal{L}(a, \cdot) da \in \Delta(\mathcal{S})$
5. $r(s, a, \mathcal{L})$ is the reward

Fixing $\mathcal{L}$

Generalized Mean Field Game with fixed Mean Field are Markov Decision Process (single agent)

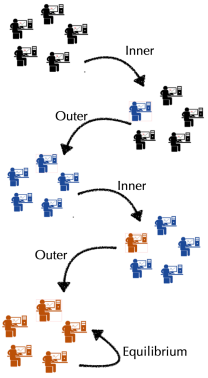
Solving MDP

1. π is the policy
2. $Q(s, a|\pi, \mathcal{L}) = \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t, \mathcal{L}) \Big| (s_0, a_0) = (s, a) \right]$ state-action value function
3. $V(\pi, \mathcal{L}) = \mathbb{E} \left[\sum_{t=1}^{+\infty} \gamma^t r(s_t, a_t, \mathcal{L}) \right]$ value function
4. $Q^*(s, a|\mathcal{L}) = \sup_\pi Q(s, a|\pi, \mathcal{L})$ optimal state-action value function

Learning in Generalized Mean Field Games

Definition of NE in GMFG

Is a policy $\pi^* : \mathcal{S} \to \Delta(\mathcal{A})$ and a mean field $\mathcal{L}^* \in \Delta(\mathcal{S} \times \mathcal{A})$ s.t $V(\pi^*, \mathcal{L}^*) \geq V(\pi, \mathcal{L}^*)$, $\forall \pi$ and $\mathcal{L}^*$ is *consistent* with policy π^*.



Algorithm 1 Model Free GMFG [Guo et al., 2019]

Require: Initial state-action distribution $\mathcal{L}_0$ simulator $\mathcal{E}_{\mathcal{L}}$[a]

1: **for** $k \in [K]$ **do**
2: Solve the MDP with fixed state-action distribution $\mathcal{L}_k$ and obtain Q_k^*
3: Update $\mathcal{L}_{k+1}$ using $\mathcal{E}_{\mathcal{L}_k}$
4: **end for**

[a] $\mathcal{E}_{\mathcal{L}}$ is a simulator for a fixed Mean Field $\mathcal{L}$.

Solving the Inner Loop

Q-learning [Watkins, 1989]

$$\hat{Q}_{N+1}(s_t, a_t) = (1-\alpha)\,\hat{Q}_N(s_t, a_t) + \alpha \left[r(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} \hat{Q}_N(s_{t+1}, a) \right]$$

Pros

Strong theoretical guarantees of convergence

Cons

Discontinuity between states

FQI [Ernst et al., 2005]

$$\mathcal{D} := \{(s_t^k, a_t^k, r_{t+1}^k, s_{t+1}^k)\}_{k=1}^K$$

Do regression of the function:

$$(s_t^k, a_t^k) \mapsto r_{t+1}^k + \gamma \max_{a \in \mathcal{A}} \hat{Q}_N(s_{t+1}^k, a)$$

Pros

Continuity is inherited from the regression

Cons

Computationally intensive

Problem Formulation

Mean Field Game model of Dealer Markets

We model an environment where the clients see an indicative market price $\tilde{P}_{t,buy}(v), \tilde{P}_{t,sell}(v)$ and then put a M market makers in competition for the firm price trough a RFQ. The M dealers are extracted from a population of $\mathfrak{M} \to \infty$ market makers.

State

- price of the asset: P_t (exogenous)
- the inventory: $z_t = z_{t-1} + v_t \mathbb{I}\{won_t\}$

Actions (a_1, a_2) where:

- $a_1 : P^i_{t,buy}(v) = \tilde{P}_{t,buy}(v)(1 + a_1)$
- $a_2 : P^i_{t,sell}(v) = \tilde{P}_{t,sell}(v)(1 + a_2)$

We assume that the market maker only decides how much to differ from a spread which is a function of the size of the trade v

The **reward** is defined as:

$$r_t = \underbrace{\mathbb{I}\{won_t\}|v_t(P_{t,buy/sell}(v_t) - P_t)|}_{\text{spread PnL}} + \underbrace{z_{t-1}(P_t - P_{t-1})}_{\text{inventory PnL}} - \underbrace{\phi(z_t)}_{\text{inventory penalty}} , \tag{1}$$

where v_t is the size of the trade, $P_{t,buy/sell}(v_t)$ is the quote published by the market maker, z_t is the inventory, $\phi : \mathbb{R} \to \mathbb{R}^+$ is the penalty of owning a net inventory

Experimental Evaluation

Experimental Setup

$$P_{t+dt} = P_t \exp\left\{\left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma Z_t \sqrt{dt}\right\}$$

$$\tilde{P}_{t,buy/sell}(v) = P_t[1 \pm \delta(|v| + 0.01v^2)]$$

$$v_t \sim \mathcal{U}(-1, 1)$$

$\phi(z) = z^2/2$, inventory penalization

$dt = 1/250, \mu = 0, \sigma = 0.2, P_0 = 100$

$\delta = 0.01$

$M \in \{2, 4\}$

Considered Metrics

- R: Mean reward ($\sum_{t=1}^{T} r_t/T$)
- L: Mean dollar reward ($\phi = 0$)
- S: Sharpe ratio $L/std(L)$
- Z: Standard deviation of the inventory

Agents

- FQI_2: GMFG-FQI trained with $M = 2$
- FQI_4: GMFG-FQI trained with $M = 4$
- Q_2: GMFG-Q trained with $M = 2$
- Q_4: GMFG-Q trained with $M = 4$

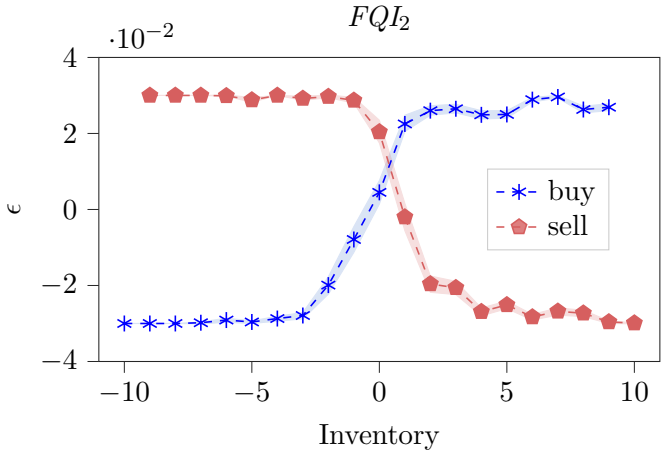
- P: plays $(a_1, a_2) = (0, 0)$
- U: plays $(a_1, a_2) \sim \mathcal{U}([0, 1]^2)$
- N: plays $(a_1, a_2) \sim \mathcal{N}(0, I)$

Exploitability

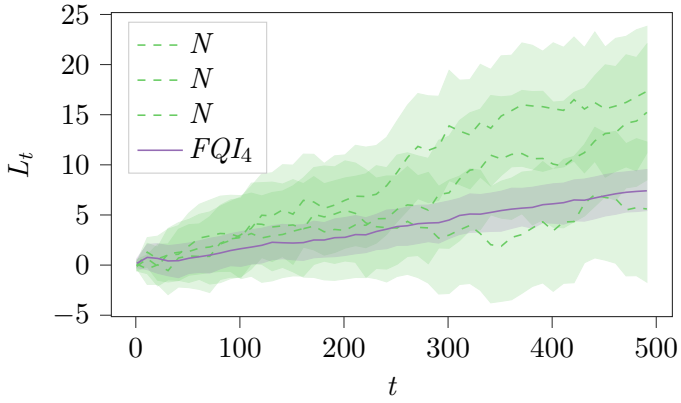
For each agent $\mathfrak{U}$ we trained an agent $E(\mathfrak{U})$ which is trained as a RL agent (PPO[1] with 2 layer Neural Network of 252 parameters each) on a MD2C platform with only the agent $\mathfrak{U}$ (already trained) and the agent $E(\mathfrak{U})$. Exploitability measures the robustness/safeness of the policies.

[1] [Schulman et al., 2017]

Results - Learned Policy



Results



Results - Exploitability

	FQI_4	FQI_2	Q_2	Q_4	P0	U	N
Dollar Reward	**0.049**	0.048	-0.002	0.021	0.009	0.018	0.026
Sharpe ratio	**0.008**	0.008	-0.0	0.002	0.001	0.002	0.002
Inventory std dev	**0.01**	0.01	0.019	0.019	0.02	0.019	0.019
Reward	-15.871	**-13.858**	-49.19	-48.955	-49.812	-48.275	-45.943

Conclusions

Contributions and Future Works

Contributions

- Introduced a truly multi-agent setting in Dealer Markets
- Proposed the use of learning in GMFG to find the equilibrium profile
- Empirical validation of the methodology

Future Works

- Consider a portfolio of correlated assets
- Use other RL techniques (such as DeepRL) to solve the inner loop
- How to incorporate data into the framework?

References

References

[Ernst et al., 2005] Ernst, D., Geurts, P., and Wehenkel, L. (2005).
Tree-based batch mode reinforcement learning.
Journal of Machine Learning Research, 6.

[Fermanian et al., 2016] Fermanian, J.-D., Guéant, O., and Pu, J. (2016).
The behavior of dealers and clients on the european corporate bond market: the case of multi-dealer-to-client platforms.
Market microstructure and liquidity, 2(03n04):1750004.

[Ganesh et al., 2019] Ganesh, S., Vadori, N., Xu, M., Zheng, H., Reddy, P., and Veloso, M. (2019).
Reinforcement learning for market making in a multi-agent dealer market.
arXiv preprint arXiv:1911.05892.

[Guo et al., 2019] Guo, X., Hu, A., Xu, R., and Zhang, J. (2019).
Learning mean-field games.
arXiv preprint arXiv:1901.09585.

[Lasry and Lions, 2007] Lasry, J.-M. and Lions, P.-L. (2007).
Mean field games.
Japanese journal of mathematics, 2(1):229–260.

[Schulman et al., 2017] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017).
Proximal policy optimization algorithms.
arXiv preprint arXiv:1707.06347.

[Watkins, 1989] Watkins, C. J. C. H. (1989).
Learning from delayed rewards.

Appendix

FQI-GMFG

Algorithm 2 FQI for GMFG

Require: Initial state-action distribution $\mathcal{L}_0$, simulator $\mathcal{E}_{\mathcal{L}}$

for $k \in [K]$ **do**
  Initialize $\hat{Q}_k(s,a) = 0 \ \forall a \in \mathcal{A}, s \in \mathcal{S}$
  Generate dataset $\mathcal{D}_k = \{(s_i, a_i, \mathcal{R}(s_i,a_i), s_i')\}_{i\in[D]}$
  for $j \in [J]$ **do**
    $\hat{Q}_{k,j+1} = \arg\min_{f\in\mathcal{F}} \sum_{i\in D} \left(f(s_i,a_i) - r_i - \gamma \min_{a\in\mathcal{A}} \hat{Q}_{k,j}(s',a) \right)^2$
  end for
  Obtain $\hat{Q}_k(s,a) = \hat{Q}_{k,J}(s,a)$ from FQI algorithm.
  $\pi_k(s) = \phi_\tau(\hat{Q}_k(s,\cdot))$
  $\mu_k \leftarrow \int_{\mathcal{A}} \mathcal{L}_k(s,a)\,da$
  Initialize $\mathcal{L}_{k+1}(s,a) = 0 \ \forall a \in \mathcal{A}, s \in \mathcal{S}$
  for $i \in [N]$ **do**
    $s_i \sim \mu_k, \ a_i \sim \pi_k(s_i)$
    $s_i' \leftarrow \mathcal{E}_{\mathcal{L}_k}(s_i, a_i)$
    $\mathcal{L}_{k+1}(s_i', a_i) = \mathcal{L}_{k+1}(s_i', a_i) + \frac{1}{N}$
  end for
end for

Q-GMFG

Algorithm 3 Q-learning for GMFG

Require: Initial state-action distribution $\mathcal{L}_0$, simulator $\mathcal{E}_{\mathcal{L}}$

for $k \in [K]$ **do**
  Initialize $\hat{Q}_k(s, a) = 0 \ \forall a \in \mathcal{A}, s \in \mathcal{S}$
  for $j \in [J]$ **do**
    Update the $\hat{Q}_k(s, a)$ with Q-learning on the MDP defined by $\mathcal{E}_{\mathcal{L}_k}$
  end for
  $\pi_k(s) = \phi_\tau(\hat{Q}_k(s, \cdot))$
  $\mu_k \leftarrow \int_{\mathcal{A}} \mathcal{L}_k(s, a)\, da$
  Initialize $\mathcal{L}_{k+1}(s, a) = 0 \ \forall a \in \mathcal{A}, s \in \mathcal{S}$
  for $i \in [N]$ **do**
    $s_i \sim \mu_k, \ a_i \sim \pi_k(s_i)$
    $s_i' \leftarrow \mathcal{E}_{\mathcal{L}_k}(s_i, a_i)$
    $\mathcal{L}_{k+1}(s_i', a_i) = \mathcal{L}_{k+1}(s_i', a_i) + \frac{1}{N}$
  end for
end for

Additional results

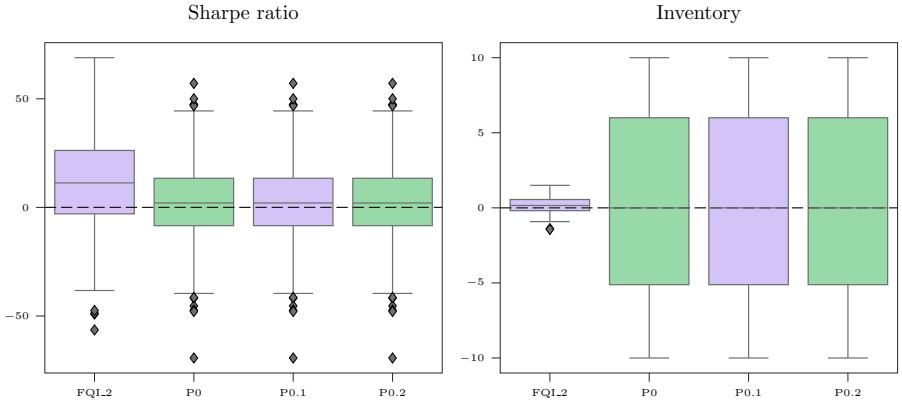
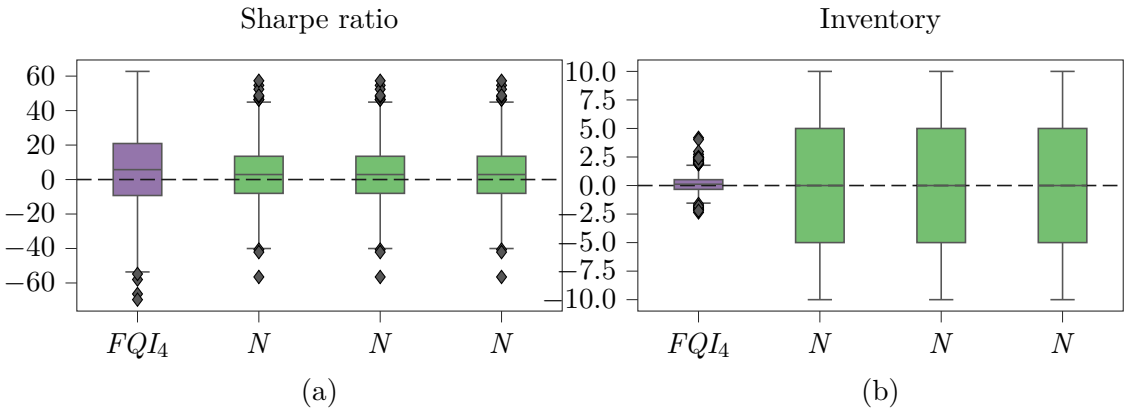


Figure 1: Box-plot of the distribution of 1000 episodes of the Sharpe ratio S (a) and the inventory z_t (b).

Results


Sharpe ratio
60
40
20
0
−20
−40
−60
FQI_4
N
N
N
(a)
Inventory
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
FQI_4
N
N
N
(b)

Additional results on Mean Dollar Reward L

	FQI_4	FQI_2	Q_2	Q_4	P0	U	N
FQI_4	0.058	0.034	0.068	0.071	0.089	0.039	0.057
FQI_2	0.068	0.051	0.069	0.071	0.08	0.057	0.068
Q_2	0.036	0.031	0.047	0.041	0.053	0.043	0.041
Q_4	0.035	0.028	0.046	0.044	0.049	0.046	0.044
FQI_4,FQI_4,FQI_4	0.03	0.014	0.029	0.031	0.031	0.031	0.037
Q_4,Q_4,Q_4	0.013	0.007	0.016	0.021	0.009	0.023	0.033
U,U,U	0.01	0.007	0.023	0.013	0.015	0.019	0.037
P0,P0,P0	0.019	0.014	0.039	0.045	0.026	0.035	0.066
N,N,N	0.015	0.007	0.011	0.016	0.008	0.017	0.021
Max	0.068	0.051	0.069	0.071	0.089	0.057	0.068
Min	0.01	0.007	0.011	0.013	0.008	0.017	0.021
Mean	0.032	0.021	0.039	0.039	0.04	0.034	0.045

Additional results on Mean Sharpe ratio S

	FQI_4	FQI_2	Q_2	Q_4	P0	U	N
FQI_4	0.011	0.008	0.009	0.009	0.012	0.006	0.008
FQI_2	0.011	0.01	0.009	0.01	0.011	0.007	0.009
Q_2	0.026	0.025	0.007	0.006	0.008	0.006	0.006
Q_4	0.025	0.023	0.007	0.006	0.007	0.007	0.006
FQI_4,FQI_4,FQI_4	0.008	0.006	0.004	0.004	0.004	0.004	0.004
Q_4,Q_4,Q_4	0.014	0.008	0.003	0.003	0.002	0.003	0.005
U,U,U	0.011	0.008	0.004	0.002	0.003	0.002	0.006
P0,P0,P0	0.021	0.017	0.006	0.007	0.004	0.005	0.009
N,N,N	0.011	0.005	0.002	0.003	0.001	0.003	0.004
Max	0.026	0.025	0.009	0.01	0.012	0.007	0.009
Min	0.008	0.005	0.002	0.002	0.001	0.002	0.004
Mean	0.015	0.012	0.006	0.006	0.006	0.005	0.006

Additional results inventory standard deviation Z

	FQI_4	FQI_2	Q_2	Q_4	P0	U	N
FQI_4	0.01	0.008	0.013	0.013	0.012	0.013	0.013
FQI_2	0.011	0.01	0.013	0.012	0.012	0.012	0.013
Q_2	0.002	0.002	0.012	0.012	0.012	0.012	0.012
Q_4	0.002	0.002	0.012	0.012	0.012	0.012	0.012
FQI_4,FQI_4,FQI_4	0.008	0.005	0.013	0.013	0.013	0.013	0.013
Q_4,Q_4,Q_4	0.002	0.002	0.011	0.012	0.012	0.012	0.012
U,U,U	0.002	0.002	0.012	0.012	0.012	0.012	0.012
P0,P0,P0	0.002	0.002	0.012	0.012	0.012	0.012	0.012
N,N,N	0.003	0.002	0.012	0.012	0.012	0.012	0.012
Max	0.011	0.01	0.013	0.013	0.013	0.013	0.013
Min	0.002	0.002	0.011	0.012	0.012	0.012	0.012
Mean	0.005	0.004	0.012	0.012	0.012	0.012	0.012

Additional results on Mean Reward R

	FQI_4	FQI_2	Q_2	Q_4	P0	U	N
FQI_4	-12.86	-7.845	-20.703	-19.897	-18.704	-20.908	-20.246
FQI_2	-15.542	-11.008	-19.668	-19.247	-19.109	-20.038	-20.931
Q_2	-0.408	-0.3	-17.995	-16.805	-18.332	-19.823	-18.28
Q_4	-0.559	-0.347	-16.288	-16.411	-17.977	-19.703	-18.434
FQI_4,FQI_4,FQI_4	-8.617	-2.357	-21.361	-21.076	-20.473	-19.996	-21.521
Q_4,Q_4,Q_4	-0.127	-0.137	-17.117	-18.001	-19.951	-18.028	-17.909
U,U,U	-0.141	-0.116	-16.997	-18.438	-17.66	-17.611	-18.085
P0,P0,P0	-0.162	-0.117	-15.928	-17.355	-19.137	-17.999	-17.998
N,N,N	-0.333	-0.322	-16.394	-17.562	-18.776	-17.62	-18.638
Max	-0.127	-0.116	-15.928	-16.411	-17.66	-17.611	-17.909
Min	-15.542	-11.008	-21.361	-21.076	-20.473	-20.908	-21.521
Mean	-4.305	-2.505	-18.05	-18.31	-18.902	-19.081	-19.116

Why Game Theory in Market Making?

Single agent optimization problems

- Environment is stochastic and independent
- Environment does not adapt to your actions

Pros

Easy/Close form solutions

Cons

- Interaction is reduced to background noise
- Past "data" of the environment can not describe future environments

Multi-agent optimization problems

- Environment is generated by the actions of all the players
- Players adapt to the changes of behaviour of all the players

Pros

Less assumptions/more realistic

Cons

- No concept of "best" action, because best action depends on the aggregate behaviour
- Computationally complex to "solve"