# Distributional Temporal Difference Learning for Finance: Dealing with Leptokurtic Rewards

Brain, Mind and Markets Laboratory

Shijie Huang[1], Nitin Yadav[1], Peter Bossaerts[1,2]

[1]Brain, Mind and Markets Laboratory, Department of Finance, University of Melbourne, Parkville, Australia; [2]Florey Institute of Neuroscience and Mental Health, Parkville, Australia.

THE UNIVERSITY OF MELBOURNE

THE FLOREY INSTITUTE OF NEUROSCIENCE & MENTAL HEALTH

## Motivation

- The domain of finance remains a challenge for RL. A characteristic statistical feature of financial data is the presence of *leptokurtosis*, caused by frequent outliers ("tail risks").
- Leptokurtosis emerges in the credit assignment problem [4], and it is also an issue in human learning [3].
- We exploit distributional RL [2], and introduce efficient estimation of the action-values. The results vastly improve.

## Core Issues

**Outliers swamp TD error and push policy off optimal path even with a small learning rate.**

- Outliers overwhelm action-value distributions.
- Outliers make it hard to determine boundaries in Categorical Distributional RL (d-RL-Categorical).
- TD error: sum of a leptokurtic term and asymptotically non-leptokurtic terms.

TD-error: $\underline{R(s,a)} + \underline{\gamma Q(s',a') - Q(s,a)}$

*Leptokurtic Distribution* — *Non-Leptokurtic Distribution*

## Our Solution

1. $R(s,a)$ & $\gamma Q(s',a') - Q(s,a)$ should be dealt with separately.
2. Exploiting distributional RL, we incorporate Maximum Likelihood Estimator (MLE) principle into RL.

## References

**[1]** Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
**[2]** Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 449-458). JMLR. org.
**[3]** d'Acremont, M., & Bossaerts, P. (2016). Neural mechanisms behind identification of leptokurtic noise and adaptive behavioral response. Cerebral Cortex, 26(4), 1818-1830.
**[4]** Singh, S., & Dayan, P. (1998). Analytical mean squared error curves for temporal difference learning. *Machine Learning*, *32*(1), 5-40.
**[5]** Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279-292.

## Solutions to Core Issues

### SARSA

$$Q(s,a) \leftarrow Q(s,a) + \alpha * [R(s,a) + \gamma Q(s',a') - Q(s,a)]$$

Update a single action-value

Update action-value distribution

### d-RL-MLE & d-RL

$$Q(s,a) \leftarrow Q(s,a) + \alpha * [\hat{E}(R(s,a)) + \gamma Q(s',a') - Q(s,a)]$$

$\hat{E}(R(s,a))$

Maximum Likelihood Estimator → Most efficient estimator that reaches Cramér-Rao lower bound → **d-RL-MLE**

Integration over reward distribution → **d-RL**

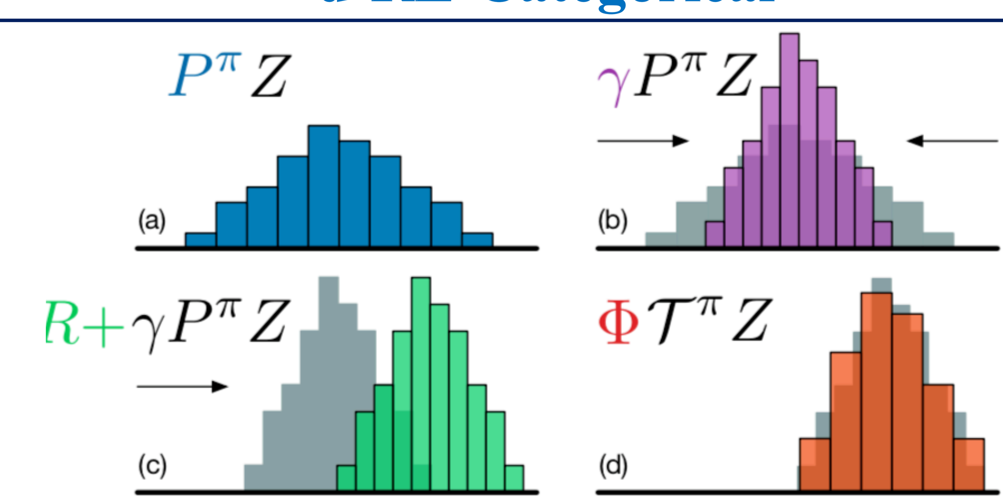
### Results (Post Exploration) **

| Optimal policy convergence rate | Gaussian N(0, 1) | Student-t (dof=1.1) | Empirical S&P500 |
|---|---|---|---|
| SARSA | 82% | 2% | 47% |
| d-RL-Categorical | 80% | 4% | - |
| d-RL | 100% | 33% | 95% |
| d-RL-MLE | 100% | **95%** | **97%** |

**"d-RL-MLE" outperforms both in simulations and field data**

**(numbers in the results table represent: in how many games out of 100 does the agent manage to obtain optimal policy at the end of all 200 episodes per game?)

### d-RL-Categorical

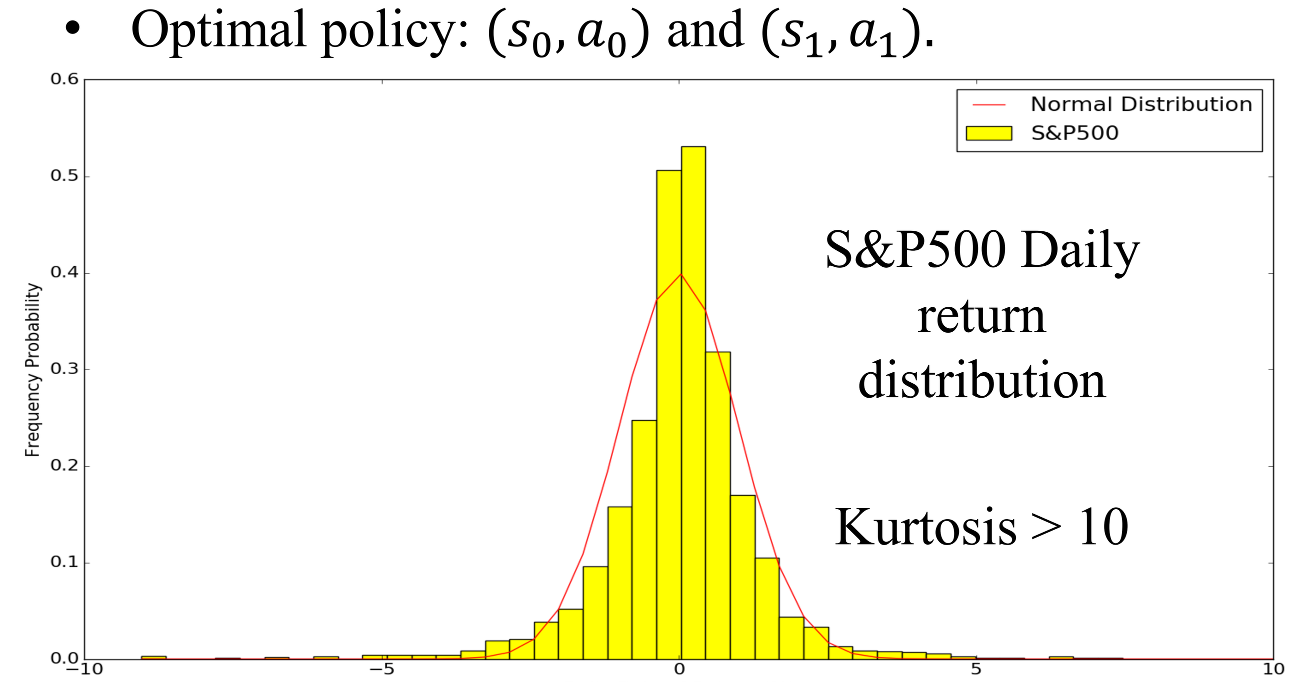



Action–value is a weighted average of possible values $Q(s,a) = \sum_i z_i p_i(s,a)$

## Environment: financial decision making task

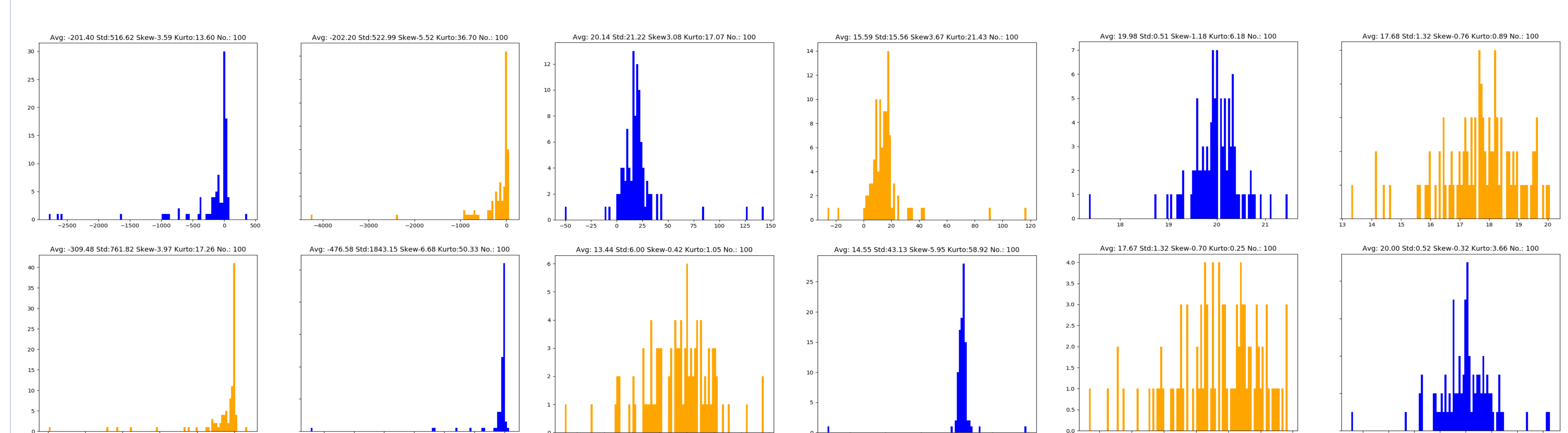
| $R(s,a) \sim$ | $s_0$ | $s_1$ | $P(s'\mid s)$ | $s_0$ | $s_1$ |
|---|---|---|---|---|---|
| $a_0$ | + 1.5 | + 1.0 | $s_0$ | 0.7 | 0.3 |
| $a_1$ | + 1.0 | + 1.5 | $s_1$ | 0.3 | 0.7 |

- Optimal policy: $(s_0, a_0)$ and $(s_1, a_1)$.




- Return > 2 std.dev.: 40 times in 25 years
- Fitted value of degree of freedom (dof) using Student-t distribution: 3.29

## Illustration: single game behavior of action-value distributions



SARSA action-values show high kurtosis

In d-RL, action-values are less extreme

In d-RL-MLE, the frequent outliers are eliminated