Reinforcement Learning and Double Auctions

Performance, Strategies, and Market Design

Pranjal Rawat

December 26, 2023

Georgetown University

Introduction

Findings

I conduct experiments with reinforcement learning (RL) in dynamic double auctions (DA):

- Reinforcement learning can outperform simple trading rules.
- Reinforcer competition is efficient and prices are stable.
- Prices do not show quick reversals or corrections.
- There is no fight to hold the current bid (ask).

However,

- Reinforcers can learn to collude, but can also be vulnerable to it.
- Increasing disclosures can, paradoxically, worsen market outcomes.

Research Outline

- Motivation:
  - Breakthrough in learning algorithms for dynamic problems, especially with high dimensional states and granular action spaces.
  - Rise of high frequency, computerized markets driven by algorithms not humans; in sectors like finance, advertising, energy, e-commerce.
- Questions:
  - How well can reinforcement learning perform in double auctions?
  - How do we design auctions when traders use reinforcement learning?
- Directions:
  - Experimental study of the Santa Fe double auction tournament.
  - Monte Carlos with Q-learning and one-sided auctions.
  - Study of Q-learning's replicator dynamic and mean field game.

Literature

Experiments with discrete or continuous double auctions:

Period	Authors	Research Focus
1960-1990	Smith, Williams, Porter	• Human Behaviour • Efficiency
1980-2010	Easley, Ledyard, Gode, Sunder, Rust, Friedman, Dickhaut, Gerstaud, Cliff, Tesuaro, Das	• Strategies • Performance • Price Formation
2000-2025	Andrews, Prager, Wellman, Hu, Tesfatsion, Chen, Tai	• Learning and Evolution • Evolutionary Stability

Theory: Chatterjee-Samuleson (1983), Myerson-Satterthwaite (1983), Wilson (1987), Satterthwaite-Williams (1989).

Synchronized Double Auction

Rules

Santa Fe Discrete DA: (Rust, Palmer, Friedman 1992/1993).

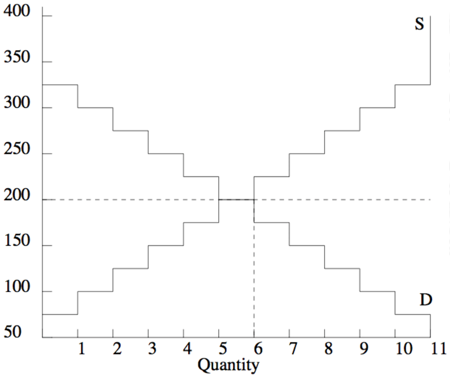
Round $\Rightarrow$ Period $\Rightarrow$ Step

- Round: Draw Token Values / Costs
- Period: Replenish Tokens
- Trading Step:
  - Bid and Ask
  - Buy and Sell
    - Price: (Bid + Ask)/2
    - Seller Reward: Price - TokenCost
    - Buyer Reward: TokenValue - Price

Game Parameters: $nRounds, nPeriods, nSteps, nTokens, nBuyers, nSellers$

Token Values

Tokens values (costs) are randomly generated for buyers (sellers).



Gives us market demand and supply, and market clearing prices.

Trading Strategies

To benchmark reinforcement learning performance, I use the following trading strategies as opponents:

- Zero-Intelligence Constrainted (ZIC) - bids randomly while respecting a budget. (Gode and Sunder 1993)
- Easley-Ledyard (EL) - human-like bluffing at first, then adjusts profit margin according to performance. (Easley and Ledyard 1983)
- Zero-Intelligence Plus (ZIP) - bids randomly in the range of an adjustable profit margin. (Cliff and Bruten 1997)
- Gjerstad-Dickhaut (GD) - forecasts winning bids and bids if profit is maximized. (Gjerstad and Dickhaut 1998)
- Kaplan-Ringouette (KR) - does not bid until the bid-ask gap closes, then jumps in and steals the deal. (Rust, Palmer, Friedman 1992/1993)

Reinforcement Learning

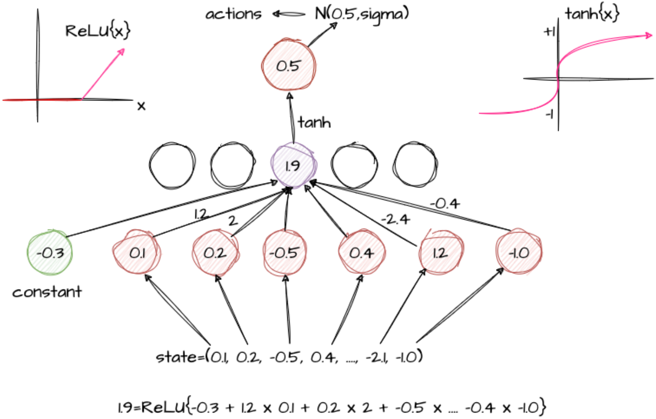
Notation

Variables	Functions
State: $s \in \mathbb{R}^N$	Round: $\tau = (s_0, a_1, r_1, ...a_T, r_T, s_T)$
Action[1]: $a \in [-1, 1]$	Policy: $\pi_\theta(a\|s) = \mathbb{P}(a_t = a\|s_t = s; \theta)$
Reward: $r \in \mathbb{R}$	Return: $G(\tau) = \sum_{t=0}^{T} \gamma^t r_t$
Discounting: $\gamma \in (0, 1)$	Exp. Return: $J^\pi = E_{\tau \sim \pi}[G(\tau)]$

[1] Are linked to bids (asks) by normalization $frac = (a + 1)/2$
$\text{bid} = \text{bid}_{\text{min}} frac + \text{bid}_{\text{max}}(1 - frac)$

Policies $\pi(a|s;\theta)$

Policies are parametrized through neural networks: $a_t \sim \mathbb{N}(\mu(s_t;\theta),\sigma)$



This permits *continuous stochastic actions* and *high dimensional states*.

Algorithm

REINFORCE is a popular policy gradient algorithm. (Williams 1992)

- **Objective**: Improve policy π_θ.
- While not converged, do:
  - Create dataset of rounds $\mathbb{D}$ using π_θ
  - Compute return: $G(\tau)$ for τ in $\mathbb{D}$
  - Backpropagation: $\frac{d\mu(s_t;\theta)}{d\theta}$
  - Compute log-probability gradient: $\frac{d \log(\pi_\theta(a_t|s_t))}{d\theta}$
  - Compute policy gradient:

$$\frac{dJ(\theta)}{d\theta} = |\mathbb{D}|^{-1} \sum_{\mathbb{D}} [\sum_{t=0}^{T-1} \frac{d \log \pi_\theta(a_t|s_t)}{d\theta} G(\tau)]$$

  - Update policy parameters: $\theta \leftarrow \theta + \alpha \frac{dJ(\theta)}{d\theta}$

Experiments

Experimental Design

A standard series of experiments:

Single Agent RL:

A1: Baseline

A2: vs Particular Trading Strategy

Multi-Agent RL (Main):

B1: Baseline

B2: Inelastic Supply

B3: Few Buyers

B4: Single Token Only

B5: Non-Random Tokens

B6: High Discount Factor

B7: Reduced Disclosures

B8: Zero Disclosures

B9: Conditional Disclosures

B10: Second-Price DA

B11: NYSE Rule

B12: Offer Fees

B13: Reserve Prices

Measuring Performance

Performance is measured across Rounds, not Periods or Steps.

Individual Performance

- Avg. Profit in last 100 rounds
- Std. Profit in last 100 rounds
- Speed of Learning

Market Performance

- Efficiency: fraction of total possible surplus obtained.
- Price Dispersion around Market Clearing Prices
- Speed of Convergence of Prices to Clearing Levels

Game Parameters

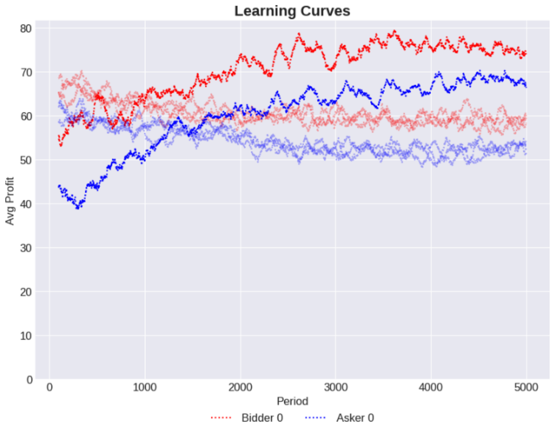
These parameters stay fixed in all experiments.

- *nRounds*: 5,000
- *nPeriods*: 1
- *nSteps*: 16
- *nTokens*: 4
- *nBuyers*: 4
- *nSellers*: 4

Token values are drawn from a fixed distribution (normal).

Experiment I - Single Agent RL

Buyer 1 and Seller 1 are Reinforcers, rest are ZIC. **There are no public disclosures**. We look at average profit over 100 rounds.



The reinforcers, with minimal information, outsmart the ZIC agents.

Experiment I - Single Agent RL

Prices are volatile but neither side seems to enjoy market power.

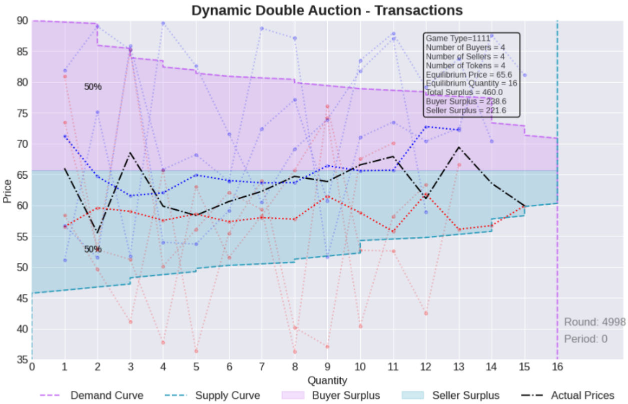
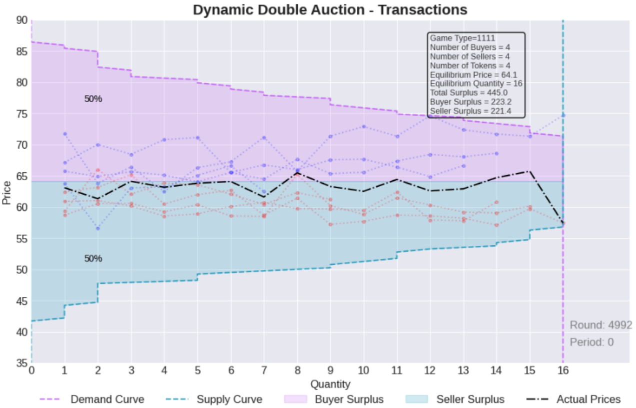


Figure 1: Red (Bids), Blue (Asks), Black (Prices)

Unlike ZIC agents, reinforcers are able to bid close to prices.

Experiment II - Multi-Agent RL (No Disclosure)

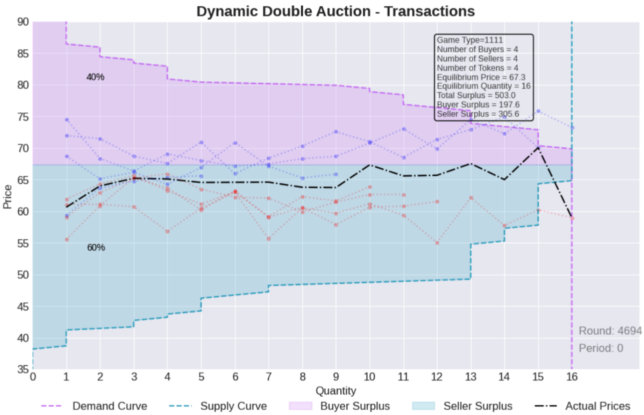
All agents are reinforcers, there are **no public disclosures**.



Prices are not volatile, and efficiency is very high - but there is noticeable buyer power. Offers are also closer together.

Experiment III - Multi-Agent RL (Full Disclosure)

All agents are reinforcers, there is **full public disclosure**.



Prices continue to be less volatile and efficiency remains high, but buyer power is pronounced. Red (Bids), Blue (Asks), Black (Prices).

Summary of Experimental Results

Criterion	Humans Only[2]	ZIC Only	Single-RL (1B,1S)	Multi-RL (No Disc)	Multi-RL (Full Disc)
Efficiency as % of realized vs possible	Higher than ZIC	98.7 (0.02)	98.6 (0.02)	99.4 (0.01)	0.99 (0.06)
Buyer Efficiency as % of realized vs possible	Close to 100%.	1.03 (0.15)	1.03 (0.12)	1.05 (0.12)	1.07 (0.15)
Mean Absolute Deviation of Prices from Clearing Levels	Lower than ZIC	4.63 (0.96)	4.51 (0.91)	1.53 (0.56)	2.28 (0.99)
Price volatility in Std Dev	Lower than ZIC	5.41 (0.98)	5.11 (0.89)	1.99 (0.99)	2.15 (0.65)
1st order Auto correlation in Prices	Close to ZIC (-0.5 to -0.25)	-0.04 (0.24)	-0.03 (0.25)	+0.09 (0.29)	+0.019 (0.32)
Avg. % Current Bid Handovers	Higher than ZIC (nearer to 100%)	72%	67%	60%	64%

[2] Gode and Sunder 1993, Cason and Friedman 1996.

Summary of Findings

I study reinforcement learning (RL) in dynamic double auctions (DA):

- Reinforcement learning can outperform simple trading rules.
- Reinforcer competition is efficient and prices are stable.
- Prices do not show quick reversals or corrections.
- There is no fight to hold the current bid (ask).

However,

- Reinforcers can learn to collude, but can also be vulnerable to it.
- Increasing disclosures can, paradoxically, worsen market outcomes.

Next Steps

- Conduct the full experiment.
- Ensure valid inference.
- Find which disclosures improve outcomes.
- Test reinforcers against humans.

Appendix

FAQ-I

- What is the economic motivation?
  - To study the effect of information disclosures on market outcomes when traders use reinforcement learning.
- Why not use a theoretical approach?
  - Because attempts to reduce reinforcement learning to a differential equation have only been done for a single state (Banachio et al, Asker-Pakes).
  - No general characterization of Bayesian Nash Equilibria for the dynamic double auction. Wilson (1987) provides a single example, but that is rejected by human data (Cason & Friedman 1996).
- Is this a computer science project?
  - No, it's a computational experiment. Any computer science is confined to the agent's learning process.

FAQ-II

- Why should we care about this research?
  - It demonstrates the possibility of algorithmic collusion even in a market widely considered to be highly efficient.
  - It offers some policy advise on market design which the current theoretical approach cannot address.
- How generalizable are these results?
  - I use a very standardized double auction setup and a classic reinforcement learning algorithm; so this study generalizes as well as most papers in this field.
  - I collect data over multiple trials to ensure valid inference.
- Can experiments have a wider appeal than theorem proving?
  - The famed efficiency of the double auction was establised in experiments such as Smith (1962), Gode-Sunder (1993). In contrast, theoretical analysis of the double auction highlights inefficiencies (e.g. Myserson-Satterthwaite 1983).

References I (Experimental)

- *Chen, S.-H., & Tai, C.-C. (2010).* The Agent-Based Double Auction Markets: 15 Years On. *World Congress on Social Simulation.*
- *Cason, T. N., & Friedman, D. (1996).* Price formation in double auction markets. *Journal of Economic Dynamics and Control.*
- *De Luca, M., & Cliff, D. (2011).* Agent-Human Interactions in the Continuous Double Auction, Redux - Using the OpEx Lab-in-a-Box to explore ZIP and GDX. *International Conference on Agents and Artificial Intelligence.*
- *Erev, I., & Roth, A. E. (1998).* Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review.*
- *Friedman, D. (2018).* The Double Auction Market Institution: A Survey.
- *Gjerstad, S., & Dickhaut, J. (2001).* Price Formation in Double Auctions. *Lecture Notes in Computer Science.*

References II (Experimental)

- *Holt, C. A., & Holt, C. A. (2021).* An experimental study of competitive market behavior (by Vernon L. Smith). *The Art of Experimental Economics.*
- *Hu, J., & Wellman, M. P. (1998).* Online learning about other agents in a dynamic multiagent system. *International Conference on Autonomous Agents.*
- *Nicolaisen, J., Petrov, V., & Tesfatsion, L. (2000).* Market Power and Efficiency in a Computational Electricity Market with Discriminatory Double-Auction Pricing. *Computational Economics.*
- *Rust, J., Miller, J. H., & Palmer, R. G. (2018).* Behavior of Trading Automata in a Computerized Double Auction Market. *Lecture Notes in Computer Science.*
- *Tesauro, G., & Bredin, J. (2002).* Strategic sequential bidding in auctions using dynamic programming. *Adaptive Agents and Multi-Agent Systems.*
- *Tesauro, G., & Das, R. (2001).* High-performance bidding agents for the continuous double auction. *ACM Conference on Economics and Computation.*

References III (Theoretical)

- *Chatterjee, K., & Samuelson, W. (1983).* Bargaining under Incomplete Information. *Operations Research, vol. 31, issue 5, 835-851.*
- *Myerson, R. B., & Satterthwaite, M. A. (1983).* Efficient Mechanisms for Bilateral Trading. *Journal of Economic Theory, 29, 265-281.*
- *Sobel, J., & Takahashi, I. (1983).* A Multistage Model of Bargaining. *Review of Economic Studies, vol. 50, issue 3, 411-426.*
- *Satterthwaite, M. A., & Williams, S. R. (1989).* Bilateral trade with the sealed bid k-double auction: Existence and efficiency. *Journal of Economic Theory, Jun 1989;48(1):107-133.*
- *Bulow, J. I., & Klemperer, P. (1994).* Auctions vs. Negotiations. *NBER Working Paper No. w4608, January 1994.*

References IV (Theoretical)

- *Bulow, J., & Klemperer, P. (1996).* Auctions Versus Negotiations. *The American Economic Review, Vol. 86, No. 1. (Mar., 1996), pp. 180-194.*
- *Pesendorfer, W., & Swinkels, J. M. (1997).* The Loser's Curse and Information Aggregation in Common Value Auctions. *Econometrica, Econometric Society, vol. 65(6), 1997.*
- *Pesendorfer, W., & Swinkels, J. M. (2000).* Efficiency and Information Aggregation in Auctions. *American Economic Review, vol. 90, no. 3, June 2000 (pp. 499-525).*
- *Cripps, M. W., & Swinkels, J. M. (2006).* Efficiency of Large Double Auctions. *Econometrica, Econometric Society, vol. 74 (1), pages 47-92, January 2006.*

References V (Reinforcement Learning)

- *Watkins, C. J. C. H., & Dayan, P. (1992).* Q-learning. *Machine Learning, 8, 279-292.*
- *Sutton, R. S., & Barto, A. G. (1998).* Reinforcement Learning: An Introduction. *MIT Press, Cambridge, Mass., ISBN 0-262-19398-1.*
- *Littman, M. L. (1994).* Markov games as a framework for multi-agent reinforcement learning. *In Proceedings of the Eleventh International Conference on Machine Learning, 157-163, Morgan Kaufmann.*
- *Mnih, V. et al. (2015).* Human-level control through deep reinforcement learning. *Nature, 2015.*

References VI (Reinforcement Learning)

- *Mnih, V. et al. (2016).* Asynchronous Methods for Deep Reinforcement Learning. *2016.*
- *Brockman, G., Cheung, V., Pettersson, L., et al. (2016).* OpenAI Gym. *CoRR, arXiv:1606.01540.*
- *Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017).* Proximal Policy Optimization Algorithms.
- *Silver, D. et al. (2018).* A general reinforcement learning algorithm that masters chess, shogi, and Go. *2018.*
- *Lillicrap, T. P. et al. (2019).* Continuous control with deep reinforcement learning. *Google Deepmind, London, UK*

Research Motivation

Algorithms	Applications
Reinforcement Learning	Stock Trading, Real-Time Bidding, Chess, Go, Starcraft, Atari, Self-driving Cars, Robotics, Physical Control
Multi-Armed Bandits	Dynamic Pricing, Website Personalization, Digital Marketing, Portfolio Optimization

Sector	% World GDP	Computerized Markets
Financial	20-25	NYSE, Chicago Ex, Forex, Cryptocurrencies
Energy	6	Electricity, Natural Gas
E-Commerce	2.5	Retail, Resale
Advertising	2	Sponsored Search, Display Advertising

Research Outline

Research Questions

- How well does reinforcement learning perform in auctions?
- How to design auctions for multi-agent reinforcement learning?

Research Directions

- Experiments: Reinforcement Learning and Double Auctions.
- Experiments: Q-learning in First and Second Price Auctions.
- Q-learning and its Replicator Dynamics / Mean Field Games.

Algorithmic Collusion

A few experiments with reinforcement learning show **algorithmic collusion** and **market inefficiency**:

Year	Market	Authors	Methodology
2006	Electricity Auction	Tellidou-Bakirtzis	Experiments
2008	Cournot Oligopoly	Waltman-Kaymak	Theory + Experiments
2020	Bertrand Oligopoly	Calvano et al.	Experiments
2020	Multi-sided Platforms	Johnson et al.	Experiments
2021	One-sided Auction	Banchio-Skrzypacz	Theory + Experiments
2022	Prisioners' Dilemma	Dolgopolov	Theory

Theoretical Insights

Key highlights from theoretical literature:

- Uncertainty about valuations $\Rightarrow$ bluffing $\Rightarrow$ market inefficiency (Myerson-Satterthwaite 1983).
- No. of traders $\uparrow$ $\Rightarrow$ honesty $\Rightarrow$ market efficiency (Satterthwaite-Williams 1989).
- Wilson's 1987 example of Dynamic Bayesian Nash Equilibrium:
  - High-value traders "wait out" low-value traders.
  - Non-serious offers are just not believed, so nobody makes them.
  - Every serious offer is led to completion in a descending "Dutch" way.
  - Each event is used to update assessments.

Double Auctions I

There are many closely related types of auctions:

- Double Auction - traders (buyers/sellers) message the bid/ask offer, and decide whether to buy/sell.
- Single Auction - Buyers post bids in single or multiple rounds, and the seller chooses a winner and a payment amount from the bids.
- Posted Price - Sellers (buyers) announce ask (bid) prices and then buyers (sellers) accept or reject.

Double Auctions II

Auctions can vary along other dimensions as well:

Auction Type	Examples
Single-dimensional vs multi-dimensional	Auction based on price vs one based on price, date, quality
One-sided or multi-sided	Art auction vs Call market (buyers and sellers)
Open-cry or sealed-bid	Bids (winning or otherwise) are revealed or they are not
First-price, second-price or k-th price	Winner pays their bid, the second-highest bid or the k-th highest bid
Single-unit or multi-unit	Auction for one barrel of wine vs for X barrels of wine in one go
Single-item or multi-item / combinatorial	Single item vs Bundles of products (e.g. 10 barrels of wine, 1 box of fish, etc.)

Double Auctions III

- Double auction is where buyers place **bids** and sellers place **asks**.
- Types:
  - Periodic - bids and asks are recieved for a fixed duration, quantity demanded and supplied for each price is computed, and market clearing price is determined. e.g. NYSE Call Market
  - Continuous - the market does not close, but the auctioneer immediately matches bids and asks as many as it can in a continuous fashion. e.g. Comodity trading at Chicago
- These are most commonly used in stock markets where buyers and sellers try to sell blocks of shares (multi-unit auctions).

Double Auctions IV

- At any time the prevailing bids and asks can be tallied up to find the quantity demanded and quantity supplied at any given price.
- A range of prices may clear the market, in the figure it is 20-20$.

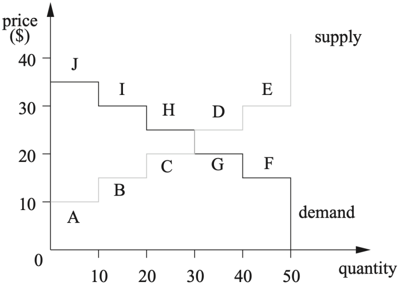


Fig. 2. Illustrative supply and demand curves for a double auction.

Double Auctions V

- The main benefit of double auctions is that they economize on information and lead to market clearing prices.
- If an auctioneer wanted to clear this market, she would have to compute the demand and supply curves from everybody's reservation prices. This is infeasible.
- But double auctions have shown that even with extremely sparse information and only a few traders, prices quickly converge to market clearing levels.
- They have also been found to be more efficient than one-sided auctions or posted pricing.
- The mechanism ensures that even with silly trading strategies, prices converge and allocation is efficient.

Policy Gradient Theorem I

Here I show how the Policy Gradient theorem can converge to local optima when the environment is stationary.

- Probability of Episode:

$$\mathbb{P}(\tau|\pi) = \mathbb{P}(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t)\mathbb{P}(s_{t+1}|s_t, a_t)$$

- Global Expected Return:

$$J(\pi) = E_{\tau\sim\pi}\left[G(\tau)\right] = \int_{\tau} \mathbb{P}(\tau|\pi)G(\tau)$$

Then the problem of Reinforcement Learning is to find the optimal policy,

$$\pi^* = \text{argmax}_{\pi} J(\pi)$$

Policy Gradient Theorem II

Since policy π_θ is parametrized by θ, we can incrementally improve $J(\theta)$ by gradient ascent:

$$\theta \leftarrow \theta + \alpha \frac{dJ(\theta)}{d\theta}$$

where,

$$\frac{dJ(\theta)}{d\theta} = \int_\tau \frac{d\mathbb{P}(\tau|\pi_\theta)}{d\theta} G(\tau)$$

$$= \int_\tau \frac{d \log \mathbb{P}(\tau|\pi)}{d\theta} \mathbb{P}(\tau|\pi) G(\tau)$$

$$= E[\frac{d \log \mathbb{P}(\tau|\pi)}{d\theta} G(\tau)]$$

Policy Gradient Theorem III

Taking logs on the probability of an episode,

$$\log \mathbb{P}(\tau|\pi) = \log \mathbb{P}(s_0) + \sum_{t=0}^{T-1} [\log \pi(a_t|s_t) + \log \mathbb{P}(s_{t+1}|s_t, a_t)]$$

And taking derivative,

$$\frac{d \log \mathbb{P}(\tau|\pi)}{d\theta} = \sum_{t=0}^{T-1} \frac{d \log(\pi_\theta(a_t|s_t))}{d\theta}$$

we get the policy gradient,

$$\frac{dJ(\theta)}{d\theta} = E[\sum_{t=0}^{T-1} \frac{d \log(\pi_\theta(a_t|s_t))}{d\theta} G(\tau)]$$

Policy Gradient Theorem IV

Which can be approximated via sampling from $\mathbb{D}$ set of episodes:

$$\frac{dJ(\theta)}{d\theta} = |\mathbb{D}|^{-1} \sum_{\mathbb{D}} [\sum_{t=0}^{T-1} \frac{d \log \pi_\theta(a_t|s_t)}{d\theta} G(\tau)]$$

Compare with the gradient to maximize the log-likelihood of observing these trajectories from this policy,

$$\frac{dJ(\theta^{ML})}{d\theta^{ML}} = |\mathbb{D}|^{-1} \sum_{\mathbb{D}} [\sum_{t=0}^{T-1} \frac{d \log \pi_\theta(a_t|s_t)}{d\theta} G(\tau)]$$

So *policy gradient is an adjusted ML gradient but moves policy towards trajectories that bring higher rewards!*.

Policy Gradient Theorem V

We enable **continuous actions** through neural network f,

$$a_t \sim \mathbb{N}(\mu(s_t;\theta), \sigma)$$

Then log-probability is,

$$\log \pi_\theta(a_t|s_t) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{(a_t - \mu(s_t;\theta))^2}{2\sigma^2}$$

And its derivative,

$$\frac{d \log \pi_\theta(a_t|s_t)}{d\theta} = -\frac{1}{2}\sigma^{-2}(a_t - \mu(s_t;\theta))\frac{d\mu(s_t;\theta)}{d\theta}$$

The last term is obtained via backpropagation.

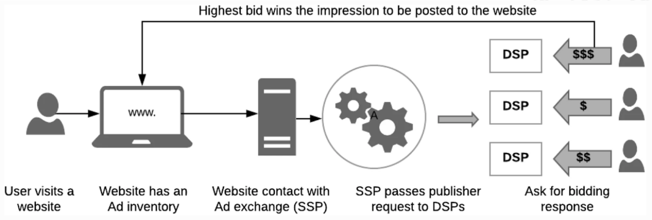
Policy Gradient Theorem VI

Demo: Teaching a robot how to walk.

Example: Google Ad Exchange[3]

- 2 million websites
- 90% of internet users
- 70% of impressions

- 90% publisher ad share
- 30 cents per ad $
- 150-300ms per auction



[3] Google recently moved from a Second Price Auction to a First Price Auction. Apart from reserve prices, winner pays a 20% fee and the winning bid is revealed. The US Display Advertising market supports 13 billion ads daily and 20 billion $ annual revenue.

First Price Auction

- Player Index: $k \in \{1, 2\}$
- Bids: $a_k \in \{0, 0.5, 1\}$
- Identical Private Value: 1
- Winner Fees: $\epsilon = 0.25$
- k-th Payoff $R(a_k, a_{-k})$:

$$= \begin{cases} 1 - a_k - \epsilon & \text{if } a_k > a_{-k} \\ \frac{1 - a_k - \epsilon}{2} & \text{if } a_k = a_{-k} \\ 0 & \text{if } a_k < a_{-k} \end{cases}$$

- Payoff Matrices: A, B

	0	0.5	1
0	$0.5, 0.5$	$0, 0.5$	$0, 0$
0.5	$0.5, 0$	$0.25, 0.25$	$0, 0$
1	$0, 0$	$0, 0$	$0, 0$

- PNE: (0,0), (0.5, 0.5)
- Mixed Strategy:
  - $\mathbb{P}(a_1 = 0) = \pi$
  - $\mathbb{P}(a_1 = 0.5) = 1 - \pi$
  - $\mathbb{P}(a_2 = 0) = \sigma$
  - $\mathbb{P}(a_2 = 0.5) = 1 - \sigma$

Replicator Dynamics: EGT

Replicator Dynamics[4] for Evolutionary Game Theory (EGT):

$$\dot{\pi}_i = \pi_i \left[\underbrace{(A\sigma)_i}_{\text{Fitness of action } i \text{ against } \sigma} - \underbrace{\pi' A\sigma}_{\text{Avg. Fitness for } \pi} \right]$$

$$\dot{\sigma}_i = \sigma_i \left[\underbrace{(\pi' B)_i}_{\text{Fitness of action } i \text{ against } \pi} - \underbrace{\pi' B\sigma}_{\text{Avg. Fitness for } \sigma} \right]$$

π_i : Prob of playing action i

$\pi : (\pi_1, \pi_2, \ldots, \pi_N)$

σ_i : Prob of playing action i

$\sigma : (\sigma_1, \sigma_2, \ldots, \sigma_M)$

[4] Borgers and Sarin 1997 show that the replicator dynamics for EGT can be derived from cross-learning, which updates π based on reward r from action j:

$$\Delta\pi_i = \begin{cases} r - \pi_i r & \text{if } i = j \\ -\pi_i r & \text{if } i \neq j \end{cases}$$

Replicator Dynamics: Q-Learning

$$\dot{Q}(i) = \pi_i^{-1}\alpha\left[R(a_1, a_2) + \max_j Q(j) - Q(i)\right]$$

$$\pi_i = \frac{e^{Q(i)}/\tau}{\sum_j e^{Q(j)/\tau}}$$

Replicator Dynamics[5]:

$$\dot{\pi}_i = \underbrace{\frac{\alpha\pi}{\tau}\left[(A\sigma)_i - \pi' A\sigma\right]}_{\text{Exploitation}} + \underbrace{\alpha\pi_i\left[\sum_j \pi_j \log\pi_j - \log\pi_i\right]}_{\text{Exploration}}$$

Q : "Long Run" Values

$Q(i)$: Value of action i

R : Payoff function

α : Learning Rate

τ : Temperature

a_k : Action taken by player k

[5] Kaisers and Tulys 2010. Action i is explored more when the entropy (uncertainty) of overall policy is high relative to π_i. And τ balances exploration vs exploitation.

Mean Field Games: Q-Learning

The PDE[6] for fraction of agents with $Q_t = (Q_t^{a_1}, Q_t^{a_2} .... Q_t^{a_N})$:

$$\dot{p}(Q_t, t) = -\sum_j \frac{d[p(Q_t, t)V_j(Q_t, \bar{\pi}_t)]}{dQ_t^{a_j}}$$

Expected change in $Q_t^{a_j}$:

$$V_j(Q_t, t) = E[\frac{dQ_t^{a_j}}{dt}] = \alpha \pi_t(a_j) E[r_t(a_j, \bar{\pi}_t) - Q_t^{a_j}]$$

and mean policy $\bar{\pi}_t$:

$$\bar{\pi}_t = \int \int ... \int \pi_t(a_j) p(Q_t, t) dQ_t^{a_1} ... dQ_t^{a_N}$$

[6]Hu et al., 2019 reduce infinite agent Q-learning to a Fokker-Plank equation without diffusion.