An Introduction to Reinforcement Learning

Yuan YAO

HKUST

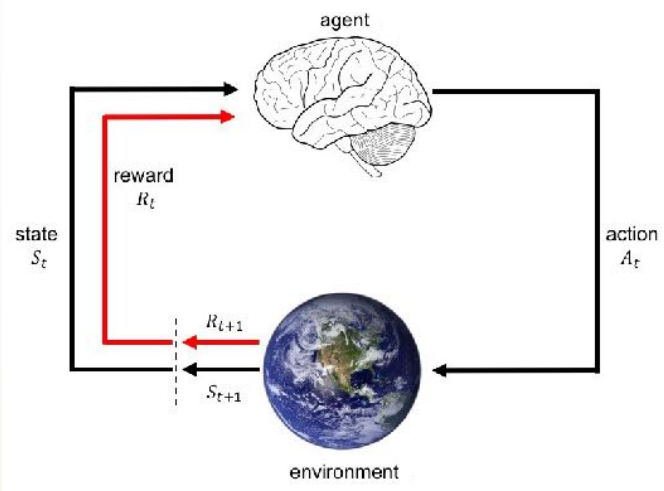
Supervised Learning

- **Data**: (x, y)
  x is input, y is output/response (label)
- **Goal**: Learn a *function* to map x -> y
- **Examples**:
  - Classification,
  - regression,
  - object detection,
  - semantic segmentation,
  - image captioning, etc.

Cat

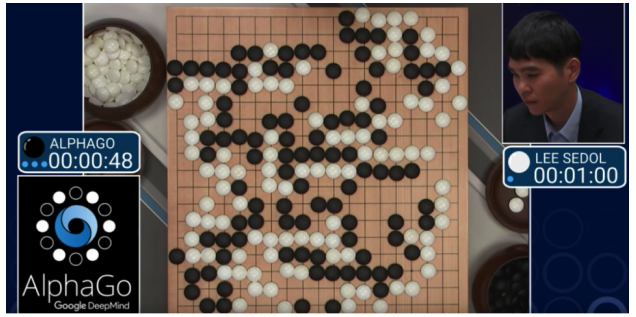
Today: Reinforcement Learning

- Problems involving an **agent**
- interacting with an **environment**,
- which provides numeric **reward** signals
- **Goal**:
  - Learn how to take actions in order to maximize reward in dynamic scenarios

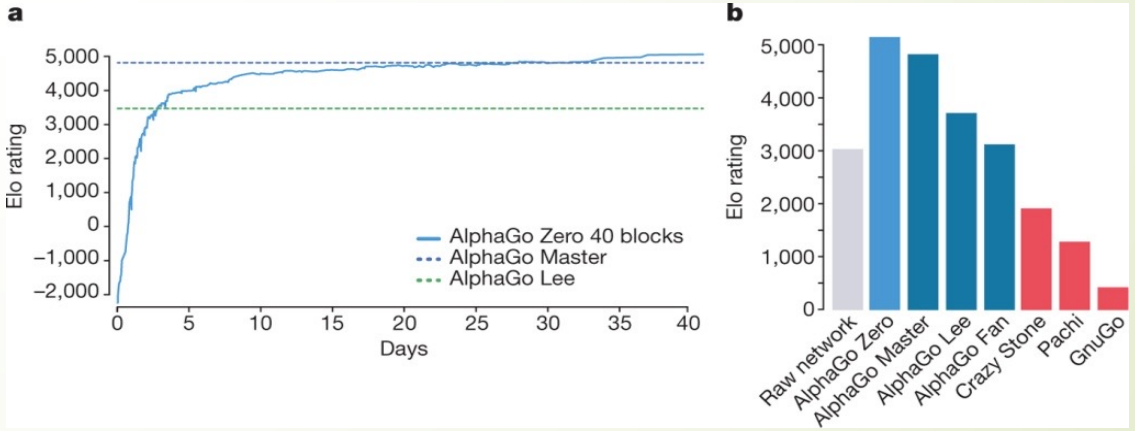


Playing games against human champions

Deep Blue in 1997

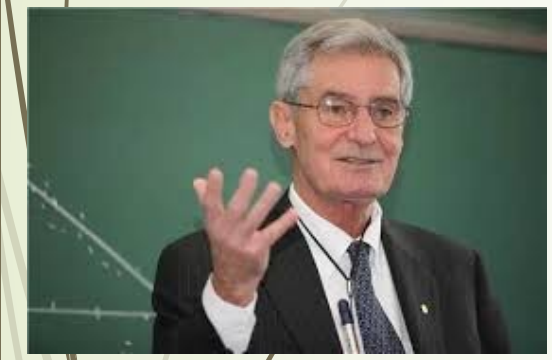


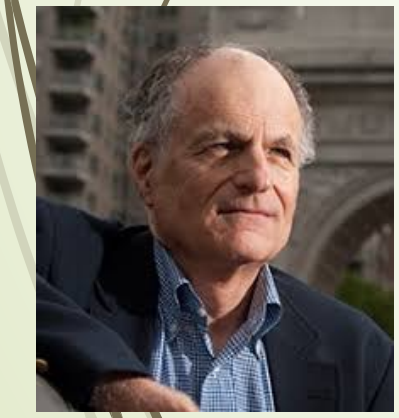
AlphaGo "LEE" 2016



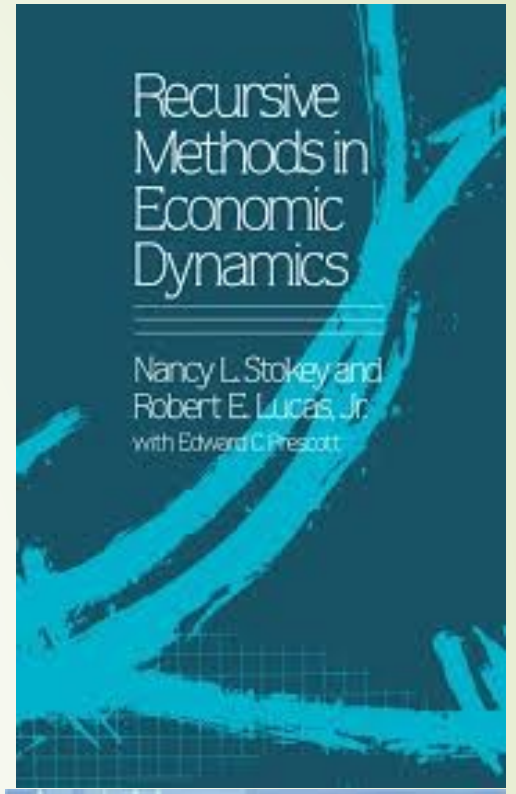
AlphaGo "ZERO" D Silver *et al. Nature* **550,** 354–359 (2017) doi:10.1038/nature24270

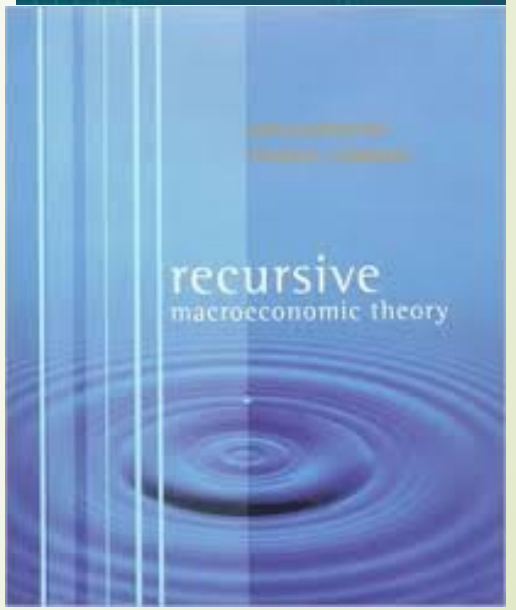
Markov Decision Process /Dynamic Programming in Economics





- The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1995 was awarded to **Robert E. Lucas Jr**. **"for having developed and applied the hypothesis of rational expectations, and thereby having transformed macroeconomic analysis and deepened our understanding of economic policy".**
- **Thomas John Sargent** was awarded the Nobel Memorial Prize in Economics in 2011 together with Christopher A. Sims for their **"empirical research on cause and effect in the macroeconomy"**





What supervision does an agent need to learn purposeful behaviors in dynamic environments?

- **Rewards**:
  - sparse feedback from the environment whether the desired goal is achieved e.g., game is won, car has not crashed, agent is out of the maze etc.
  - Rewards can be intrinsic, i.e., generated by the agent and guided by its curiosity as opposed to an external task
- Learning from **rewards**
  - Reward: jump as high as possible: It took years for athletes to find the right behavior to achieve this
- Learns from **demonstrations**
  - It was way easier for athletes to perfection the jump, once someone showed the right general trajectory
- Learns from specifications of optimal behavior
  - For novices, it is much easier to replicate this behavior if additional guidance is provided based on specifications: where to place the foot, how to time yourself etc.

How learning goal-seeking behaviors is different to supervised learning paradigms?

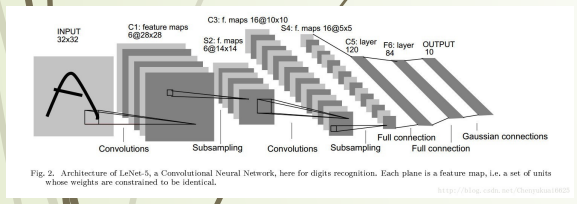


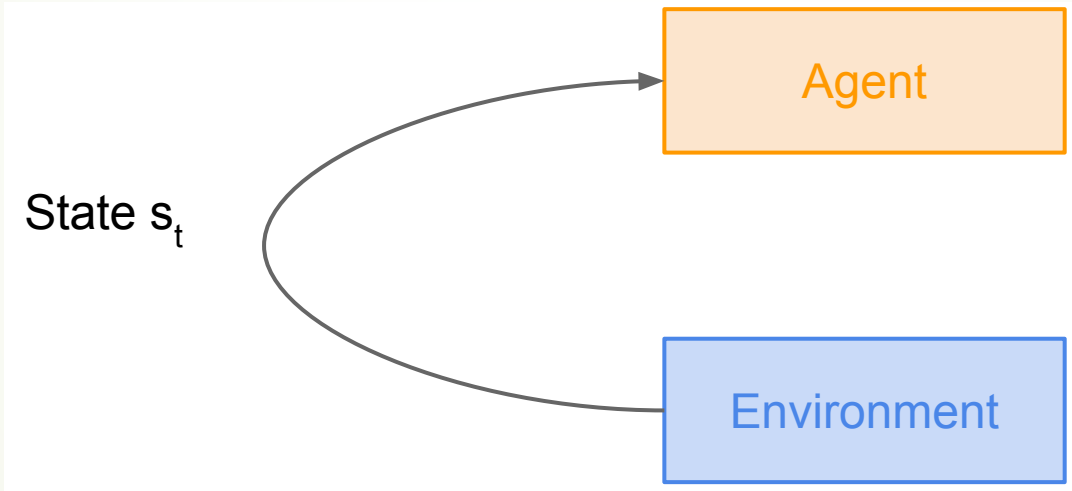
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

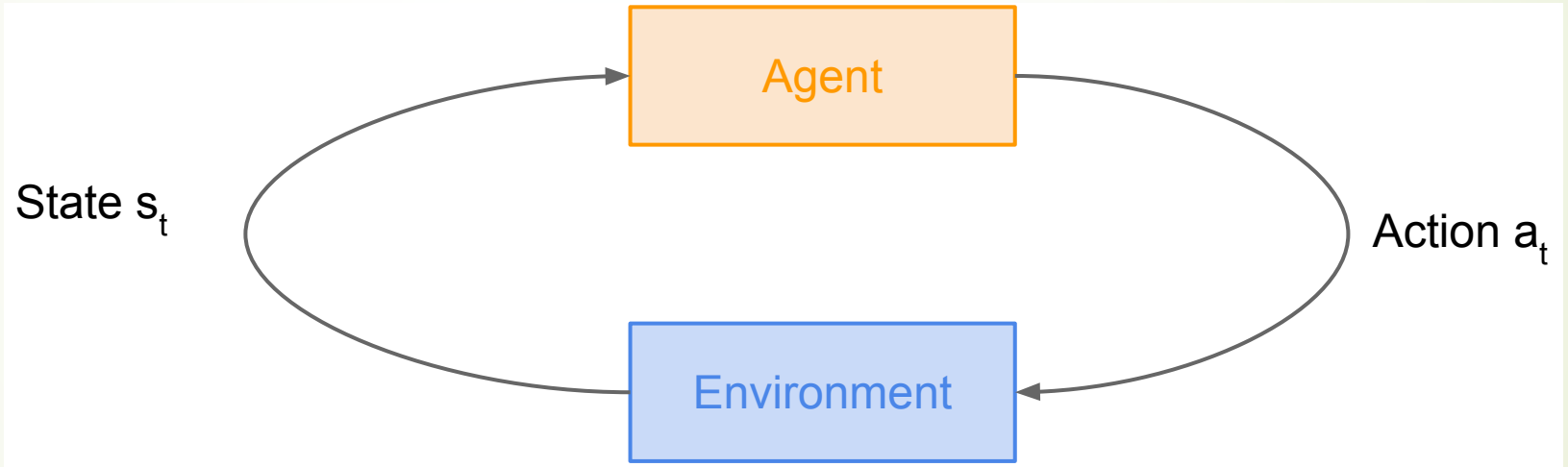
- The agent's **actions** affect the data she will receive in the future
- The **reward** (whether the goal of the behavior is achieved) is far in the future:
  - Temporal credit assignment: which actions were important and which were not, is hard to know
  - **Isn't it the same with loss of multi-layer deep networks?**
  - **No: here the horizon involves acting in the environment, rather than going from one neural layer to the next, we cannot apply chain rule to back propagate the gradient of rewards.**
  - But another way of "**Back Propagation**": **Bellman's Dynamic Programing** principle
- Actions take time to carry out in the real world, and thus this may limit the amount of experience
  - We can use simulated experience with multiple agents.

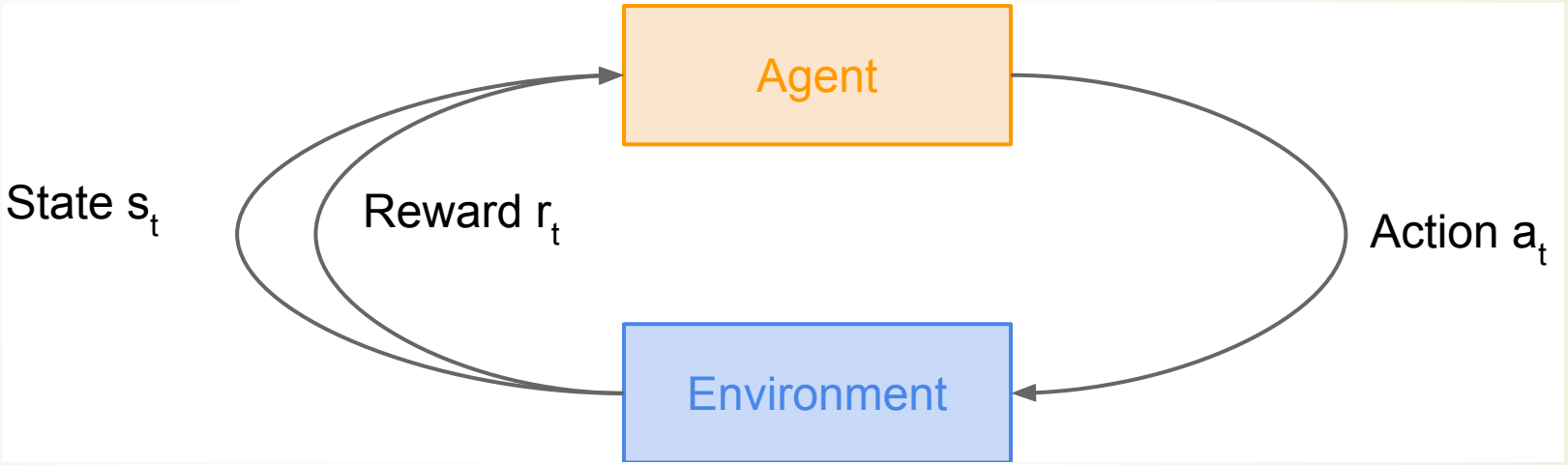
Outline

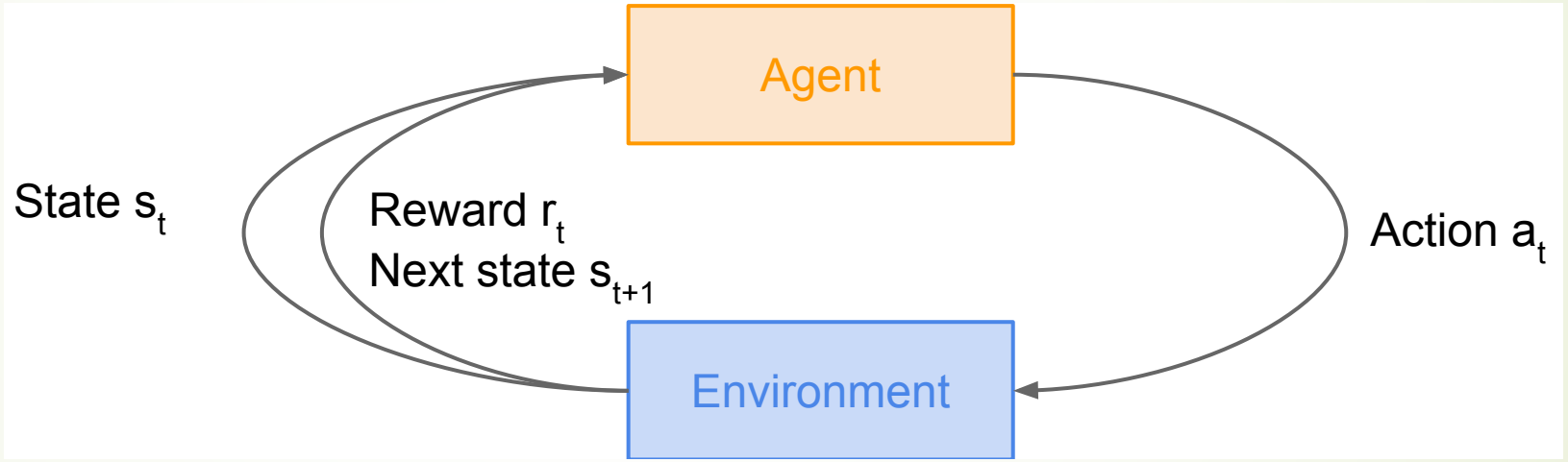
- What is Reinforcement Learning?
- Markov Decision Processes
- Bellman Equation as Linear Programming
- Q-Learning
- Policy Gradients
- Actor-Critics (Q-learning+Policy gradient)
- Examples:
  - Deep RL for quantitative trading
  - Order Book Optimization via Discrete Q-Learning by Prof. Michael Kearns

Reinforcement Learning

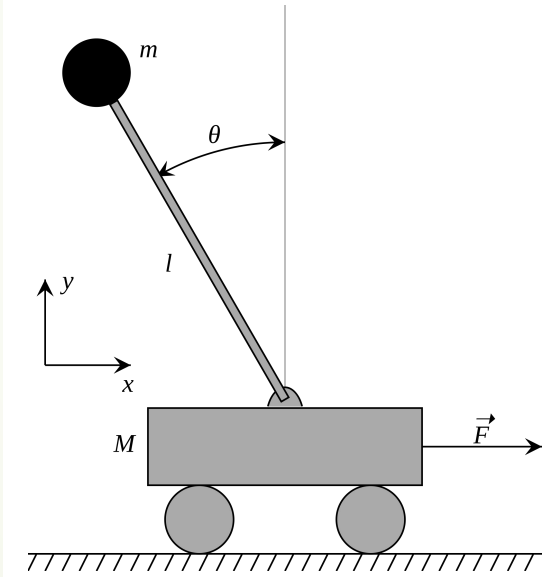

Agent
State s_t
Environment


Agent
State s_t
Action a_t
Environment


Agent
State s_t
Reward r_t
Action a_t
Environment


Agent
State s_t
Reward r_t
Next state s_{t+1}
Action a_t
Environment

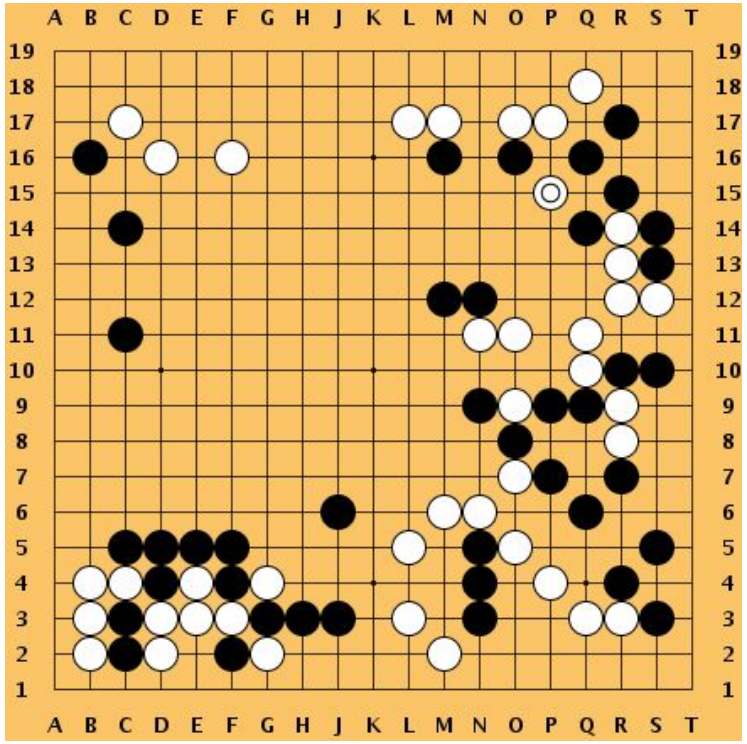
Car-Pole Control Problem



Objective: Balance a pole on top of a movable cart

State: angle, angular speed, position, horizontal velocity
Action: horizontal force applied on the cart
Reward: 1 at each time step if the pole is upright

Go Game



Objective: Win the game!

State: Position of all pieces
Action: Where to put the next piece down
Reward: 1 if win at the end of the game, 0 otherwise

Mathematical Formulation of Reinforcement Learning

A Markov Decision Process is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbb{P}, \gamma)$

- $\mathcal{S}$ is a set of states
- $\mathcal{A}$ is a set of actions
- $\mathcal{R}$ is a distribution of reward given (state, action) pair

$$R_{t+1} \sim \mathcal{R}\left[\cdot \mid S_t = s, A_t = a\right]$$

- $\mathbb{P}$ is a state transition probability function, satisfying the **Markov Property**:

$$\begin{aligned} & \mathbb{P}\left[R_{t+1} = r, S_{t+1} = s' \mid S_t, A_t\right] \\ = \quad & \mathbb{P}\left[R_{t+1} = r, S_{t+1} = s' \mid S_0, A_0, R_1, \ldots, S_{t-1}, A_{t-1}, R_t, S_t, A_t\right] \end{aligned}$$

- γ is a discount factor $\gamma \in [0, 1]$

Dynamics:

- At time step $t=0$, environment samples initial state $s_0 \sim p(s_0)$
- Then, for $t=0$ until done:
  - Agent selects **action** a_t
  - Environment samples **reward** $r_t \sim R(\cdot \mid s_t, a_t)$
  - Environment samples next **state** $s_{t+1} \sim P(\cdot \mid s_t; a_t)$
  - Agent receives reward r_t and next state s_{t+1}
- A **policy** π:S->A is a map from S to A that specifies what action to take in each state, which might be stochastic as a distribution on A
- **Objective**: find policy that maximizes the cumulated discounted reward

Rewards

- They are scalar values (not vector rewards) provided by the environment to the agent that indicate whether goals have been achieved, e.g., ***1 if goal is achieved, 0 otherwise, or -1 for overtime step the goal is not achieved***
- **Episodic tasks**: A sequence of interactions based on which the reward will be judged at the end is called episode. Interaction breaks naturally into episodes, e.g., plays of a game, trips through a maze.
- Goal-seeking behavior of an agent can be formalized as the behavior that seeks maximization of the expected value of the ***cumulative sum of (potentially time discounted) rewards, we call it return. We want to maximize returns.***
  - Return in Finite horizon:
  - Return (discounted) in infinite horizon:

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T$$

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \qquad \gamma \in [0, 1]$$

$$r(s, a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$$

Dynamics of Environment or Model

- How the states and rewards change given the actions of the agent

$$\mathrm{p}(s', r \mid s, a) = \mathbb{P}\{S_t = s', \mathrm{R_t} = r \mid \mathrm{S}_{t-1} = s, \mathrm{A}_{t-1} = a\}$$

- Transition function or next step function:

$$\mathrm{T}(s' \mid s, a) = \mathrm{p}(s' \mid s, a) = \mathbb{P}\{S_t = s' \mid \mathrm{S}_{t-1} = s, \mathrm{A}_{t-1} = a\} = \sum_{r \in \mathbb{R}} \mathrm{p}(s', r \mid s, a)$$

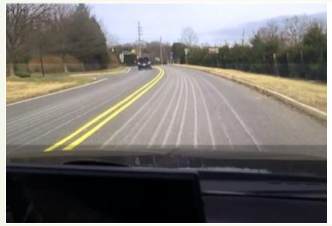
- Model-based RL: dynamics are known or are estimated, and are used for learning the policy
- Model-free RL: we do not know the dynamics, and we do not attempt to estimate them

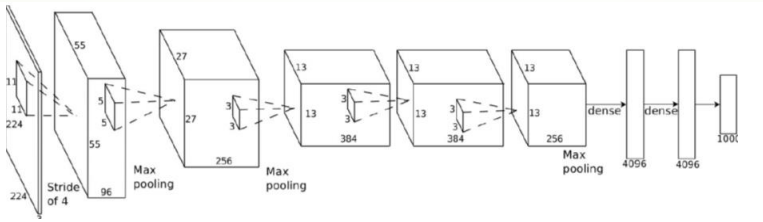
Policy

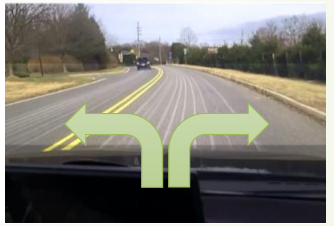
- A mapping function from states to actions of the end effectors, e.g. stochastic actions:

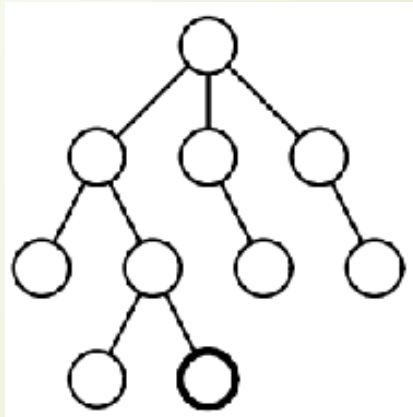
$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$$

- It can be a shallow or deep **network**, or involving a **tree** look-ahead search









The optimal policy π^*

We want to find optimal policy π^* that maximizes the sum of rewards.

How do we handle the randomness (initial state, transition probability…)?
Maximize the **expected sum of rewards!**

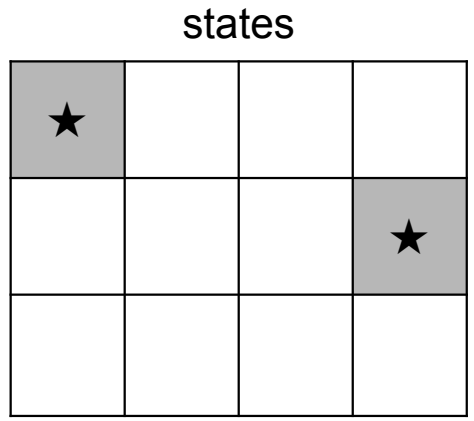
Formally: $\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t | \pi\right]$ with $s_0 \sim p(s_0), a_t \sim \pi(\cdot|s_t), s_{t+1} \sim p(\cdot|s_t, a_t)$

A simple MDP: Grid World

actions = {

1. right
2. left
3. up
4. down

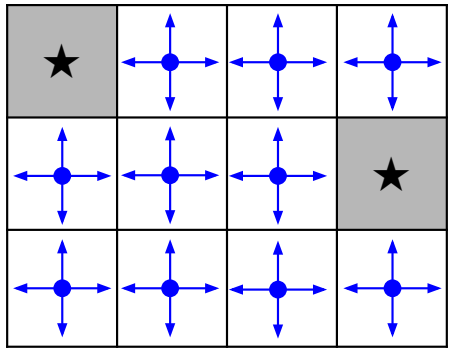
}



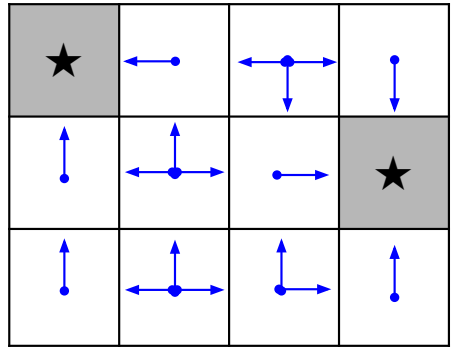
Set a negative “reward” for each transition (e.g. r = -1)

Objective: reach one of terminal states (greyed out) in least number of actions

A simple MDP: Grid World



Random Policy



Optimal Policy

- Finding the optimal policy: ***Bellman's Principle of Dynamic Programming***
  - Begin with the terminal states, find the nearest neighbors (depth-1) states with their optimal move (policy);
  - From depth-1 neighbor cells, find the optimal move (policy) of depth-2 neighbor cells;
  - And so on recursively…

Definitions: Value function and Q-value function

Following a policy produces sample trajectories (or paths) $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$

How good is a state?

The **value function** at state s, is the expected cumulative reward from following the policy from state s:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t | s_0 = s, \pi\right]$$

How good is a state-action pair?

The **Q-value function** at state s and action a, is the expected cumulative reward from taking action a in state s and then following the policy:

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t | s_0 = s, a_0 = a, \pi\right]$$

Bellman Equation of Optimal Value:

finite states and actions

Optimal Value Function $V^* : \mathcal{S} \to R = x^*$ satisfied the following nonlinear fixed point equation

$$x^*(i) = \max_{a \in \mathcal{A}} \left\{ r_a(i) + \gamma \sum_{j \in \mathcal{S}} P_a(i,j) x^*(j) \right\}$$

where a policy π^* is an optimal policy if and only if it attains the optimality of the Bellman equation.

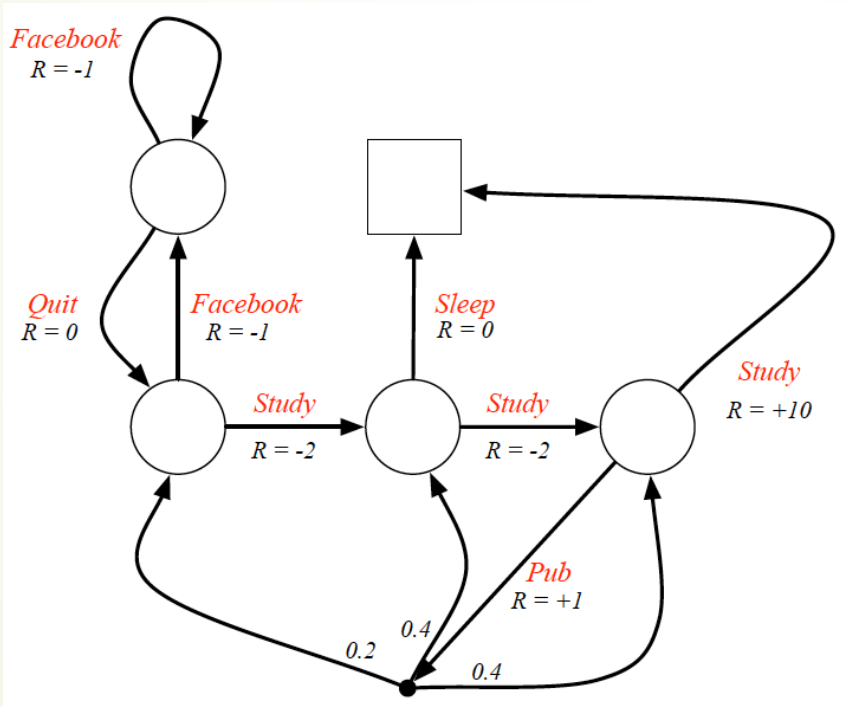
Remarks

- In the continuous-time analog of MDP, i.e., stochastic optimal control, the Bellman equation is the Hamilton-Jacobi-Bellman (HJB)
- Exact solution methods: value iteration, policy iteration, variational analysis
- What makes things hard:

Curse of dimensionality + Modeling Uncertainty

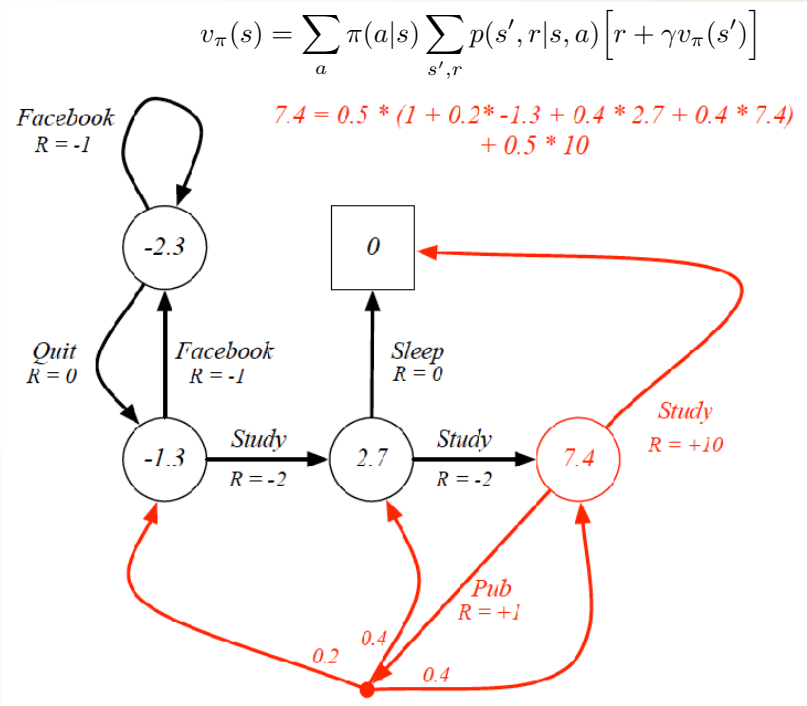
Example: the student MDP

The Student MDP



Value function

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma v_\pi(s')\right]$$



Bellman Equation as LP (Farias and Van Roy, 2003)

The Bellman equation is equivalent to

$$\begin{aligned} \text{minimize} \quad & e^T x \\ \text{subject to} \quad & (I - \gamma P_a)x - r_a \geq 0, \quad a \in \mathcal{A}, \sum_{i \in \mathcal{S}} e(i) = 1, e > 0. \end{aligned}$$

- Exact policy iteration is a form of simplex method and exhibits strongly polynomial performance (Ye 2011)
- Again, curse of dimensionality:
- Variable dimension $= |\mathcal{S}|$.
- Number of constraints $= |\mathcal{S}| \times |\mathcal{A}|$.

Duality between Value Function and Policy

Let $\lambda_{i,a} \geq 0$ be the multiplier associated with the i-th row of the primal constraint $\gamma P_a x + r_a \leq x$. The dual problem is

$$\begin{aligned} \text{maximize} \quad & \lambda_a^T r_a, \quad a \in \mathcal{A} \\ \text{subject to} \quad & \sum_{a \in \mathcal{A}} (I - \gamma P_a^T)\lambda_a = e, \quad \lambda_a \geq 0, \quad a \in \mathcal{A} \end{aligned}$$

where the dual variable is high-dimensional $\lambda = (\lambda_a)_{a \in \mathcal{A}} \in \mathbb{R}^{|\mathcal{A}||\mathcal{S}|}$.

Theorem

The optimal dual solution $\lambda^* = (\lambda^*_{i,a})_{i \in \mathcal{S}, a \in \mathcal{A}}$ is sparse and has exact $|S|$ nonzeros. It satisfies

$$\left(\lambda^*_{i,\mu^*(i)}\right)_{i \in \mathcal{S}} = (I - \alpha P_{\mu^*}^T)^{-1} e,$$

and $\lambda^*_{i,a} = 0$ if $a \neq \mu^*(i)$.

Finding the optimal policy μ^* = Finding the basis of the dual solution λ^*

Stochastic Primal-Dual Value-Policy Iteration

(Mengdi Wang (2019), Mathematics of Operations Research, 45(2):517-546. arXiv:1704.01869)

Stochastic primal-dual (value-policy) algorithm

- **Input:** Simulation Oracle $\mathcal{M}$, $n = |\mathcal{S}|$, $m = |\mathcal{A}|$, $\alpha \in (0, 1)$.
- Initialize $x^{(0)}$ and $\lambda = (\lambda_u^{(0)} : u \in \mathcal{A})$ arbitrarily.
- For$k = 1, 2, \ldots, T$
  - Sample i_k uniformly from $\mathcal{S}$ and sample u_k uniformly from $\mathcal{A}$.
  - Sample next state j_k and immediate reward $g_{i_k j_k u_k}$ conditioned on (i_k, u_k) from $\mathcal{M}$.
  - Update the iterates by

$$x^{(k-\frac{1}{2})} = x^{(k-1)} - \gamma_k \Big(-e + m\lambda_{u_k}^{(k-1)} - \alpha mn \left(\lambda_{u_k}^{(k-1)} \cdot e_{i_k}\right) e_{j_k} \Big),$$

$$\lambda_{u_k}^{(k-\frac{1}{2})} = \lambda_{u_k}^{(k-1)} + m\gamma_k \Big(x^{(k-1)} - \alpha n \left(x^{(k-1)} \cdot e_{j_k}\right) e_{i_k} - n g_{i_k j_k u_k} e_{i_k} \Big),$$

$$\lambda_u^{(k-\frac{1}{2})} = \lambda_u^{(k-1)}, \qquad \forall\ u \neq u_k,$$

  - Project the iterates orthogonally to some regularization constraints

$$x^{(k)} = \Pi_X x^{(k-\frac{1}{2})}, \qquad \lambda^{(k)} = \Pi_\Lambda \lambda^{(k-\frac{1}{2})}.$$

- **Ouput:** Averaged dual iterate $\hat{\lambda} = \frac{1}{T}\sum_{k=1}^{T} \lambda^{(k)}$

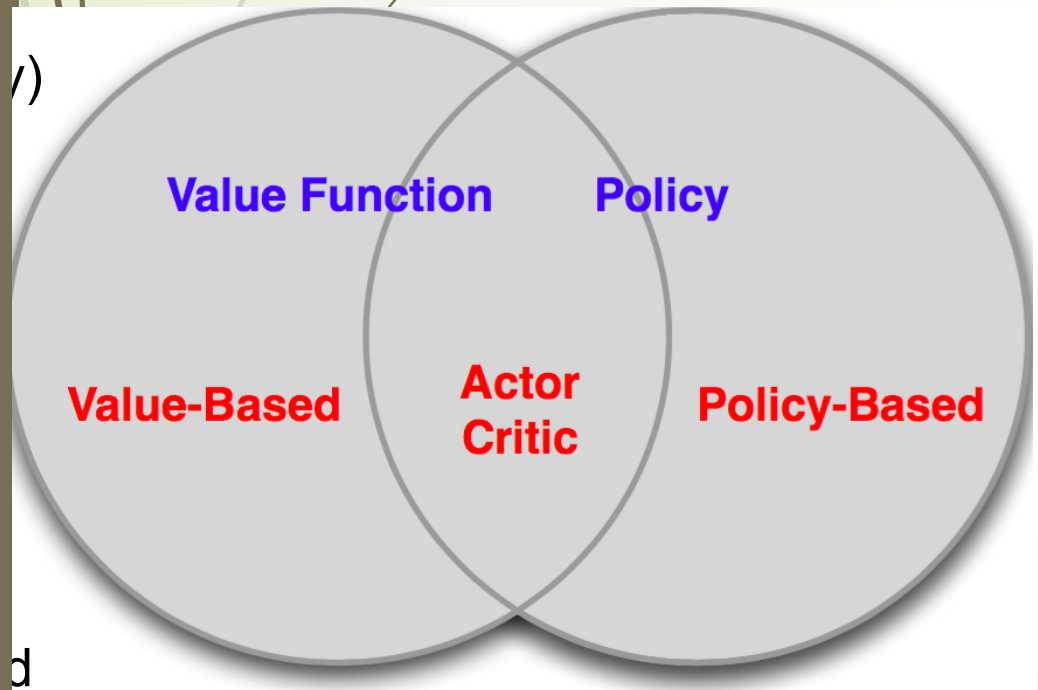
Near Optimal Primal-Dual Algorithms

Method	Setting	Sample Complexity	Run-Time Complexity	Space Complexity	Reference
Phased Q-Learning	γ discount factor, ϵ-optimal value	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\ln\frac{1}{\delta}$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\ln\frac{1}{\delta}$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert$	[17]
Model-Based Q-Learning	γ discount factor, ϵ-optimal value	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\ln\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\delta}$	NA	$\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert$	[1]
Randomized P-D	γ discount factor, ϵ-optimal policy	$\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^6\epsilon^2}$	$\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^6\epsilon^2}$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert$	[25]
Randomized P-D	γ discount factor, τ-stationary, ϵ-optimal policy	$\tau^4\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}$	$\tau^4\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert$	[25]
Randomized VI	γ discount factor, ϵ-optimal policy	$\frac{\lvert S\rvert\lvert A\rvert\cdot}{(1-\gamma)^4\epsilon^2}$	$\frac{\lvert S\rvert\lvert A\rvert\cdot}{(1-\gamma)^4\epsilon^2}$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert$	[23]
Primal-Dual π Learning	τ-stationary, t^*_{mix}-mixing, ϵ-optimal policy	$\frac{(\tau\cdot t^*_{mix})^2\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\epsilon^2}$	$\frac{(\tau\cdot t^*_{mix})^2\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\epsilon^2}$	$\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert$	This Paper

Table 1: Complexity Results for Sampling-Based Methods for MDP. The sample complexity is measured by the number of queries to the $\mathcal{SO}$. The run-time complexity is measured by the total run-time complexity under the assumption that each query takes $\tilde{\mathcal{O}}(1)$ time. The space complexity is the additional space needed by the algorithm in addition to the input.

Mengdi Wang, Primal-Dual π Learning, arXiv:1710.0610

Approaches of Deep RL: approximate dynamic programming



- Value-based RL
  - Learn an optimal value function $Q_*(s,a)$ or $V_*(s)$
  - Implicit derivation of policy
  - Deep Q-Learning (DQN), Double DQN, Dueling DQN
- Policy-based RL
  - Learn directly an optimal policy π_*
  - This is the policy achieving maximum future reward
  - Policy Gradient (PG)
- Actor-Critic RL
  - Learn a value function and a policy
  - A2C, SAC
- Model-based RL (not here)
  - Build a model of the environment
  - Plan (e.g. by look-ahead) using model

Q-Learning

Bellman equation

The optimal Q-value function Q* is the maximum expected cumulative reward achievable from a given (state, action) pair:

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t | s_0 = s, a_0 = a, \pi\right]$$

Q* satisfies the following **Bellman equation**:

$$Q^*(s,a) = \mathbb{E}_{s' \sim \mathcal{E}}\left[r + \gamma \max_{a'} Q^*(s',a') | s, a\right]$$

Intuition: if the optimal state-action values for the next time-step Q*(s',a') are known, then the optimal strategy is to take the action that maximizes the expected value of $r + \gamma Q^*(s',a')$

The optimal policy **π*** corresponds to taking the best action in any state as specified by Q*

Solving for the optimal policy

Value iteration algorithm: Use Bellman equation as an iterative update

$$Q_{i+1}(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q_i(s',a') | s,a\right]$$

Q_i will converge to Q* as i -> infinity

What's the problem with this?
Not scalable. Must compute Q(s,a) for every state-action pair. If state is e.g. current game state pixels, computationally infeasible to compute for entire state space!

Solution: use a function approximator to estimate Q(s,a). E.g. a neural network!

Solving for the optimal policy: Q-learning

Q-learning: Use a function approximator to estimate the action-value function

$$Q(s, a; \theta) \approx Q^*(s, a)$$

If the function approximator is a deep neural network => **deep q-learning**!

Solving for the optimal policy: Q-learning

Remember: want to find a Q-function that satisfies the Bellman Equation:

$$Q^*(s,a) = \mathbb{E}_{s' \sim \mathcal{E}} \left[r + \gamma \max_{a'} Q^*(s',a') | s, a \right]$$

Forward Pass

Loss function: $L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(\cdot)} \left[(y_i - Q(s,a;\theta_i))^2 \right]$

where $y_i = \mathbb{E}_{s' \sim \mathcal{E}} \left[r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) | s, a \right]$

Backward Pass

Gradient update (with respect to Q-function parameters θ):

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(\cdot); s' \sim \mathcal{E}} \left[r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) - Q(s,a;\theta_i)) \nabla_{\theta_i} Q(s,a;\theta_i) \right]$$

Yet, such a training might be unstable …

- Learning from batches of consecutive samples is problematic:
  - Samples are correlated => inefficient learning
  - Current Q-network parameters determines next training samples (e.g. if maximizing action is to move left, training samples will be dominated by samples from left-hand size) => can lead to bad feedback loops
- Experience replay will help!

DQN: Experience Replay

- To remove correlations, build a replay memory data-set D from agent's own experience

s_1, a_1, r_2, s_2
s_2, a_2, r_3, s_3
s_3, a_3, r_4, s_4
...
$s_t, a_t, r_{t+1}, s_{t+1}$

$\rightarrow \quad s, a, r, s'$

- Sample random mini-batch of transitions (s,a,r,s') from D, instead of consecutive samples
- Compute Q-learning targets w.r.t. old, fixed parameters w−
- Optimize MSE between Q-network and Q-learning target by SGD, where each transition can also contribute to multiple weight updates => greater data efficiency

$$\mathcal{L}_i(w_i) = \mathbb{E}_{s,a,r,s' \sim \mathcal{D}_i}\left[\left(\underbrace{r + \gamma \max_{a'} Q(s', a'; w_i^-)}_{\text{Q-learning target}} - \underbrace{Q(s, a; w_i)}_{\text{Q-network}}\right)^2\right]$$

[Mnih et al. NIPS Workshop 2013; Nature 2015]

Putting it together: Deep Q-Learning with Experience Replay

Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

[Mnih et al. NIPS Workshop 2013; Nature 2015]

Case Study: Playing Atari Games

Objective: Complete the game with the highest score

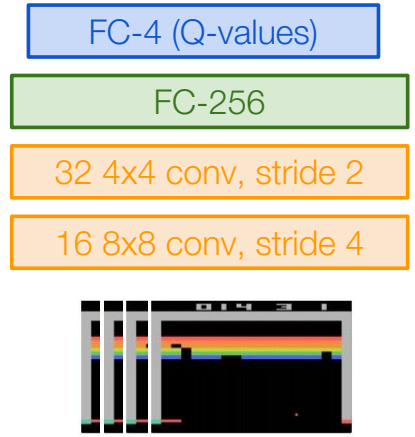
State: Raw pixel inputs of the game state
Action: Game controls e.g. Left, Right, Up, Down
Reward: Score increase/decrease at each time step

[Mnih et al. NIPS Workshop 2013; Nature 2015]

Q-network Architecture

$Q(s, a; \theta)$:
neural network
with weights θ

A single feedforward pass to compute Q-values for all actions from the current state => efficient!



Last FC layer has 4-d output (if 4 actions), corresponding to $Q(s_t, a_1)$, $Q(s_t, a_2)$, $Q(s_t, a_3)$, $Q(s_t, a_4)$

Number of actions between 4-18 depending on Atari game

Current state s_t: 84x84x4 stack of last 4 frames
(after RGB->grayscale conversion, downsampling, and cropping)

Example

- Google DeepMind's Deep Q-learning playing Atari Breakout:
  - https://www.youtube.com/watch?v=V1eYniJ0Rnk
  - Google DeepMind created an artificial intelligence program using deep reinforcement learning that plays Atari games and improves itself to a superhuman level. It is capable of playing many Atari games and uses a combination of deep artificial neural networks and reinforcement learning. After presenting their initial results with the algorithm, Google almost immediately acquired the company for several hundred million dollars, hence the name Google DeepMind. Please enjoy the footage and let me know if you have any questions regarding deep learning!

Prioritized Replay: importance sampling

[Schaul, Quan, Antonoglou, Silver, ICLR 2016]

- Current Q-network w is used to select actions
- Older Q-network w− is used to evaluate actions

Action evaluation: w−

$$I = \left(r + \gamma Q(s', \underbrace{\operatorname{argmax}_{a'} Q(s', a', \mathbf{w})}_{\text{Action selection: w}}, \mathbf{w}^-) - Q(s, a, \mathbf{w}) \right)^2$$

$$\left| r + \gamma \max_{a'} Q(s', a', \mathbf{w}^-) - Q(s, a, w) \right|$$

- Importance Weight experience according to ``surprise'' (or error):
- Store experience in priority according to DQN error:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$$

- α determines how much prioritization is used, with α = 0 corresponding to the uniform case.

Maximization Bias

- We often need to maximize over our value estimates. The estimated maxima suffer from maximization bias
- Consider a state for which all ground-truth $Q_*(s,a)=0$. Our estimates $Q(s,a)$ are uncertain, some are positive and some negative. $Q(s,\text{argmax}_a(Q(s,a))$ is positive while $Q_*(s,\text{argmax}_a(Q_*(s,a))=0$.

Double Q-Learning (DDQN)

- Train 2 **action-value** functions, Q_1 and Q_2
- Do Q-learning on both, but
  - never on the same time steps (Q_1 and Q_2 are independent)
  - pick Q_1 or Q_2 at random to be updated on each step
- If updating Q_1, use Q_2 for the value of the next state:

$$Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) + \\ + \alpha\Big(R_{t+1} + Q_2\big(S_{t+1}, \operatorname*{arg\,max}_a Q_1(S_{t+1}, a)\big) - Q_1(S_t, A_t)\Big)$$

- Action selections are with respect to the sum of Q_1 and Q_2

Double DQN:

Initialize $Q_1(s,a)$ and $Q_2(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily
Initialize $Q_1(\textit{terminal-state},\cdot) = Q_2(\textit{terminal-state},\cdot) = 0$
Repeat (for each episode):
    Initialize S
    Repeat (for each step of episode):
        Choose A from S using policy derived from Q_1 and Q_2 (e.g., ε-greedy in $Q_1 + Q_2$)
        Take action A, observe R, S'
        With 0.5 probabilility:

$$Q_1(S,A) \leftarrow Q_1(S,A) + \alpha\Big(R + \gamma Q_2\big(S', \arg\max_a Q_1(S',a)\big) - Q_1(S,A)\Big)$$

        else:

$$Q_2(S,A) \leftarrow Q_2(S,A) + \alpha\Big(R + \gamma Q_1\big(S', \arg\max_a Q_2(S',a)\big) - Q_2(S,A)\Big)$$

        $S \leftarrow S'$;
    until S is terminal

Summary of Q-Learning

- We have introduced Q-learning with several variants:
  - DQN, Double DQN, and Dueling DQN (next)
  - Experience replay, prioritization
- What is a problem with Q-learning?
  - The Q-function can be very complicated!
  - Example: a robot grasping an object has a very high-dimensional state => hard to learn exact value of every (state, action) pair
- But the policy can be much simpler: just close your hand
- Can we learn a policy directly, e.g. finding the best policy from a collection of policies?

Policy Gradients

Policy Gradients

Formally, let's define a class of parametrized policies: $\Pi = \{\pi_\theta, \theta \in \mathbb{R}^m\}$

For each policy, define its value:

$$J(\theta) = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t | \pi_\theta\right]$$

We want to find the optimal policy $\theta^* = \arg\max_\theta J(\theta)$

How can we do this?

Gradient ascent on policy parameters!

REINFORCE algorithm

Mathematically, we can write:

$$J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\theta)} [r(\tau)]$$
$$= \int_{\tau} r(\tau) p(\tau;\theta) \mathrm{d}\tau$$

Where r(τ) is the reward of a trajectory $\tau = (s_0, a_0, r_0, s_1, \ldots)$

Expected reward: $$J(\theta) = \mathbb{E}_{\tau \sim p(\tau;\theta)}\left[r(\tau)\right] = \int_{\tau} r(\tau)p(\tau;\theta)\mathrm{d}\tau$$

Now let's differentiate this: $$\nabla_\theta J(\theta) = \int_{\tau} r(\tau)\nabla_\theta p(\tau;\theta)\mathrm{d}\tau$$

Intractable! Gradient of an expectation is problematic when p depends on θ

However, we can use a nice trick: $$\nabla_\theta p(\tau;\theta) = p(\tau;\theta)\frac{\nabla_\theta p(\tau;\theta)}{p(\tau;\theta)} = p(\tau;\theta)\nabla_\theta \log p(\tau;\theta)$$

If we inject this back:

$$\nabla_\theta J(\theta) = \int_{\tau} \left(r(\tau)\nabla_\theta \log p(\tau;\theta)\right) p(\tau;\theta)\mathrm{d}\tau$$
$$= \mathbb{E}_{\tau \sim p(\tau;\theta)}\left[r(\tau)\nabla_\theta \log p(\tau;\theta)\right]$$

Can estimate with Monte Carlo sampling

REINFORCE algorithm

$$\nabla_\theta J(\theta) = \int_\tau \left(r(\tau)\nabla_\theta \log p(\tau;\theta)\right) p(\tau;\theta)\mathrm{d}\tau$$
$$= \mathbb{E}_{\tau\sim p(\tau;\theta)}\left[r(\tau)\nabla_\theta \log p(\tau;\theta)\right]$$

Can we compute those quantities without knowing the transition probabilities?

We have: $p(\tau;\theta) = \prod_{t\geq 0} p(s_{t+1}|s_t,a_t)\pi_\theta(a_t|s_t)$

Thus: $\log p(\tau;\theta) = \sum_{t\geq 0} \log p(s_{t+1}|s_t,a_t) + \log \pi_\theta(a_t|s_t)$

And when differentiating: $\nabla_\theta \log p(\tau;\theta) = \sum_{t\geq 0} \nabla_\theta \log \pi_\theta(a_t|s_t)$

Doesn't depend on transition probabilities!

Therefore when sampling a trajectory τ, we can estimate J(θ) with

$$\nabla_\theta J(\theta) \approx \sum_{t\geq 0} r(\tau)\nabla_\theta \log \pi_\theta(a_t|s_t)$$

Intuition

Gradient estimator: $\nabla_\theta J(\theta) \approx \sum_{t \geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t)$

Interpretation:

- If r(τ) is high, push up the probabilities of the actions seen
- If r(τ) is low, push down the probabilities of the actions seen

Might seem simplistic to say that if a trajectory is good then all its actions were good. But in expectation, it averages out!

However, this also suffers from high variance because credit assignment is really hard. Can we help the estimator?

Variance reduction

Gradient estimator: $\nabla_\theta J(\theta) \approx \sum_{t \geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t)$

First idea: Push up probabilities of an action seen, only by the cumulative future reward from that state

$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} \left(\sum_{t' \geq t} r_{t'} \right) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

Second idea: Use discount factor γ to ignore delayed effects

$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} \left(\sum_{t' \geq t} \gamma^{t'-t} r_{t'} \right) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

Variance reduction: Baseline

Problem: The raw value of a trajectory isn't necessarily meaningful. For example, if rewards are all positive, you keep pushing up probabilities of actions.

What is important then? Whether a reward is better or worse than what you expect to get

Idea: Introduce a baseline function dependent on the state.
Concretely, estimator is now:

$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} \left(\sum_{t' \geq t} \gamma^{t'-t} r_{t'} - b(s_t) \right) \nabla_\theta \log \pi_\theta(a_t | s_t)$$

How to choose the baseline?

$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} \left(\sum_{t' \geq t} \gamma^{t'-t} r_{t'} - b(s_t) \right) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

A simple baseline: constant moving average of rewards experienced so far from all trajectories

Variance reduction techniques seen so far are typically used in “Vanilla REINFORCE”

How to choose the baseline?

A better baseline: Want to push up the probability of an action from a state, if this action was better than the **expected value of what we should get from that state**.

Q: What does this remind you of?

A: Q-function and value function!

Intuitively, we are happy with an action a_t in a state s_t if $Q^{\pi}(s_t, a_t) - V^{\pi}(s_t)$ is large. On the contrary, we are unhappy with an action if it's small.

Using this, we get the estimator: $\nabla_\theta J(\theta) \approx \sum_{t \geq 0} (Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t)) \nabla_\theta \log \pi_\theta(a_t | s_t)$

Actor-Critic Algorithm

Problem: we don't know Q and V. Can we learn them?

Yes, using Q-learning! We can combine Policy Gradients and Q-learning by training both an **actor** (the policy) and a **critic** (the Q-function).

- The actor decides which action to take, and the critic tells the actor how good its action was and how it should adjust
- Also alleviates the task of the critic as it only has to learn the values of (state, action) pairs generated by the policy
- Can also incorporate Q-learning tricks e.g. experience replay
- **Remark:** we can define by the **advantage function** how much an action was better than expected

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

Actor-Critic Model

- Learn both **actor** (policy **π**) and **critic** (value Q and V)
  - Actor decides which action to take $\pi_\theta(a|s)$
  - **Advantage** function in critic tells how much an action might be better than expected:

$$A^{\pi_\theta}(s,a;w) = Q^{\pi_\theta}(s,a;w) - V^{\pi_\theta}(s;w)$$

- Policy gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) A^{\pi_\theta}(s,a)\right]$$

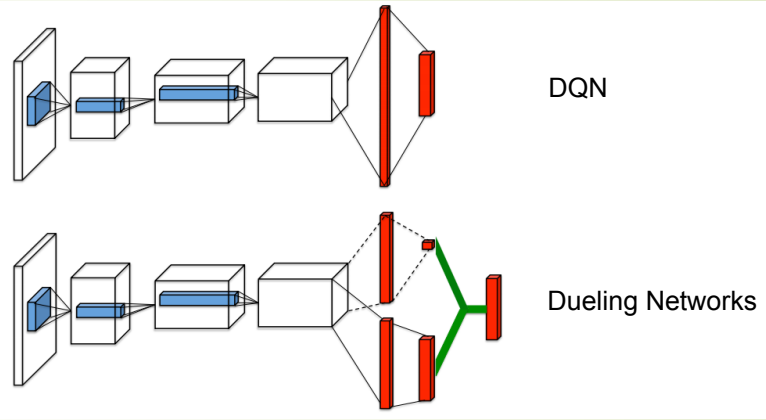
  - Stochastic Advantage can be approximated by TD-error (Temporal-Difference error)

$$\delta^{\pi_\theta} = r + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$$

One-step Actor–Critic (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
    Initialize S (first state of episode)
    $I \leftarrow 1$
    Loop while S is not terminal (for each time step):
        $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
        Take action A, observe S', R
        $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S,\boldsymbol{\theta})$
        $I \leftarrow \gamma I$
        $S \leftarrow S'$

Dueling DQN
[Wang et.al., ICML, 2016]



- Split Q-network into two channels:
  - Action-independent value function $V(s; \mathbf{w})$
  - Action-dependent advantage function $A(s, a; \mathbf{w})$

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

- Dueling DQN learns Q-function using

$$Q(s,a;\mathbf{w}) = V(s;\mathbf{w}) + \left(A(s,a;\mathbf{w}) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s,a';\mathbf{w}) \right)$$

PG Summary

- Policy Gradient:

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

- Policy Gradient with Baseline:

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \Big(G_t - b(S_t)\Big) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

- Actor-Critic Policy Gradient:

$$\theta_{t+1} = \theta_t + \alpha(R_t + \gamma \hat{v}(S_{t+1}) - \hat{v}(S_t)) \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)}$$

Maximal Entropy RL

- Promoting the stochastic policies

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \underbrace{R(s_t, a_t)}_{\text{reward}} + \alpha \underbrace{\mathrm{H}(\pi(\cdot \mid s_t))}_{\text{entropy}} \right]$$

- Why?
  - Better exploration
  - Learning alternative ways of accomplishing the task
  - Better generalization, e.g., in the presence of obstacles a stochastic policy may still succeed.

- "Soft" Bellman Equation:

$$Q^{\pi}(s,a) = r(s,a) + \mathbb{E}_{s',a'}\left[Q^{\pi}\left(s',a'\right) - \log\left(\pi\left(a' \mid s'\right)\right)\right]$$

- "Soft" Value function:

$$V\left(s\right) = \mathbb{E}_{a\sim\pi}\left[Q\left(s,a\right) - \log\pi\left(a \mid s\right)\right]$$

Soft version of actor-critic model

- Learn the following value and policy functions: $V_\psi(s_t)$ $Q_\theta(s_t, a_t)$ $\pi_\phi(a_t \mid s_t)$
  - Gradient for the **state**-value function V:

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\frac{1}{2} \left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi} \left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t | \mathbf{s}_t) \right] \right)^2 \right]$$

$$\hat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(\mathbf{s}_t) \left(V_\psi(\mathbf{s}_t) - Q_\theta(\mathbf{s}_t, \mathbf{a}_t) + \log \pi_\phi(\mathbf{a}_t | \mathbf{s}_t) \right)$$

  - Gradient for the **state-action** value Q-function:

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t) \right)^2 \right]$$

$$\hat{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \, \mathbb{E}_{\mathbf{s}_{t+1} \sim p} \left[V_{\bar{\psi}}(\mathbf{s}_{t+1}) \right]$$

$$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(\mathbf{a}_t, \mathbf{s}_t) \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - r(\mathbf{s}_t, \mathbf{a}_t) - \gamma V_{\bar{\psi}}(\mathbf{s}_{t+1}) \right)$$

- "Soft" Policy gradient:

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi_\phi(\cdot|\mathbf{s}_t) \,\middle\|\, \frac{\exp\left(Q_\theta(\mathbf{s}_t, \cdot)\right)}{Z_\theta(\mathbf{s}_t)} \right) \right]$$

$$\nabla_\phi J_\pi(\phi) = \nabla_\phi \mathbb{E}_{s_t \in D} \mathbb{E}_{a_t \sim \pi_\phi(a|s_t)} \log \frac{\pi_\phi(a_t \mid s_t)}{\exp(Q_\theta(s_t, a_t))}$$

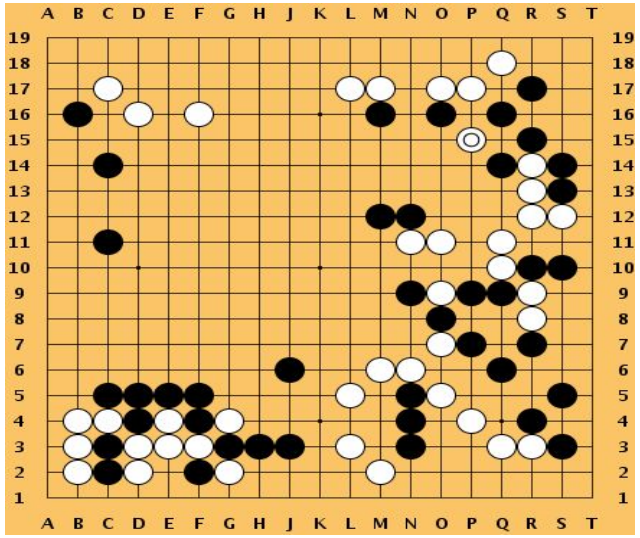
Soft Actor-Critic

- Different to openAI implementation which is essentially SoftDDQN:
  - https://spinningup.openai.com/en/latest/algorithms/sac.html

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

More policy gradients: AlphaGo



Overview:

- Mix of supervised learning and reinforcement learning
- Mix of old methods (Monte Carlo Tree Search) and recent ones (deep RL)

How to beat the Go world champion:

- Featurize the board (stone color, move legality, bias, …)
- Initialize policy network with supervised training from professional go games, then continue training using policy gradient (play against itself from random previous iterations, +1 / -1 reward for winning / losing)
- Also learn value network (critic)
- Finally, combine combine policy and value networks in a Monte Carlo Tree Search algorithm to select actions by lookahead search

[Silver et al., Nature 2016]

Summary

- **Q-learning**: does not always work but when it works, usually more sample-efficient. **Challenge**: exploration
- **Policy gradients**: very general but suffer from high variance so requires a lot of samples. **Challenge**: sample-efficiency
- Guarantees:
  - **Policy Gradients**: Converges to a local minima, often good enough!
  - **Q-learning**: Zero guarantees since you are approximating Bellman equation with a complicated function approximater

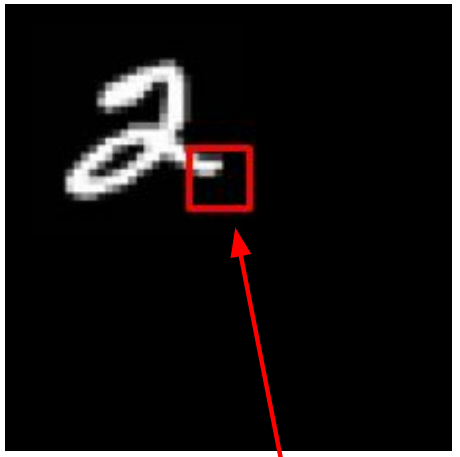
REINFORCE in action: Recurrent Attention Model (RAM)

Objective: Image Classification

Take a sequence of “glimpses” selectively focusing on regions of the image, to predict class

- Inspiration from human perception and eye movements
- Saves computational resources => scalability
- Able to ignore clutter / irrelevant parts of image

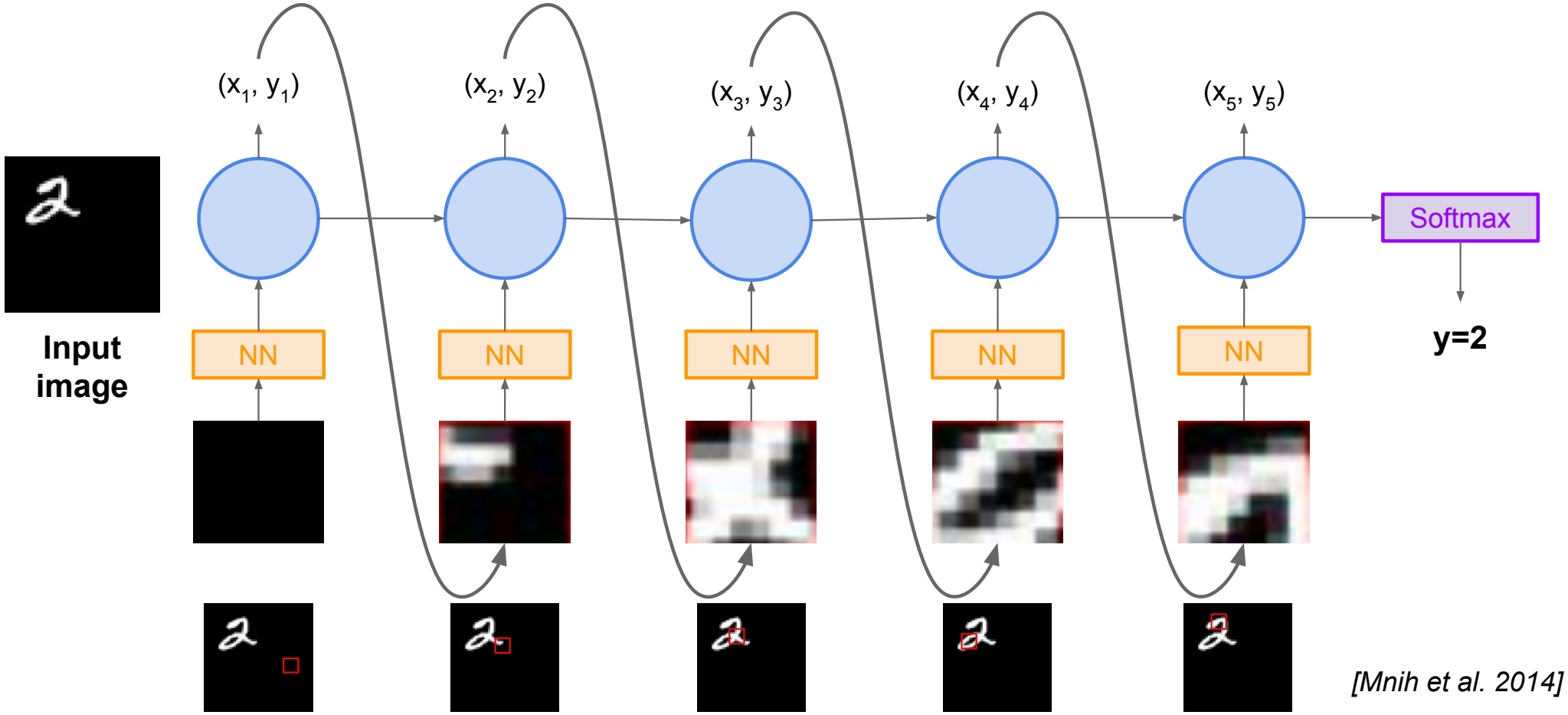
State: Glimpses seen so far
Action: (x,y) coordinates (center of glimpse) of where to look next in image
Reward: 1 at the final timestep if image correctly classified, 0 otherwise



Glimpsing is a non-differentiable operation => learn policy for how to take glimpse actions using REINFORCE
Given state of glimpses seen so far, use RNN to model the state and output next action

[Mnih et al. 2014]

REINFORCE in action: Recurrent Attention Model (RAM)



[Mnih et al. 2014]

Pytorch Implementation

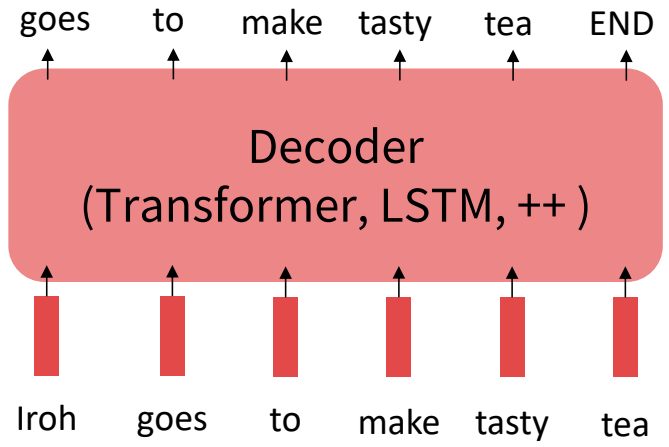
- https://github.com/kevinzakka/recurrent-visual-attention
- A Pytorch implementation for the paper, Recurrent Models of Visual Attention by *Volodymyr Mnih, Nicolas Heess, Alex Graves and Koray Kavukcuoglu*, NIPS 2014.

Alignment of LLMs: Reinforcement Learning from Human Feedback (RLHF)

Finetuning of language models as transformer decoder

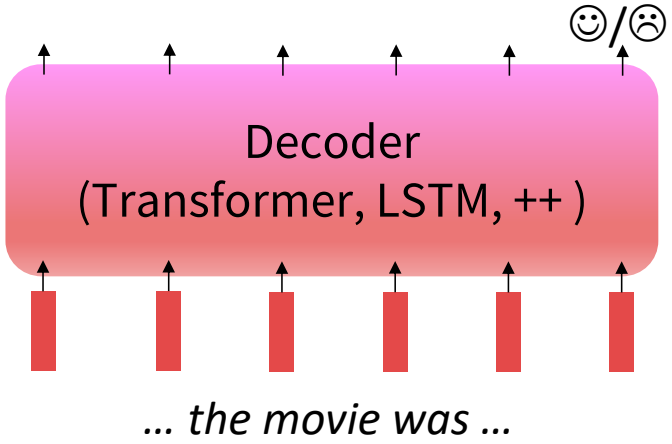
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



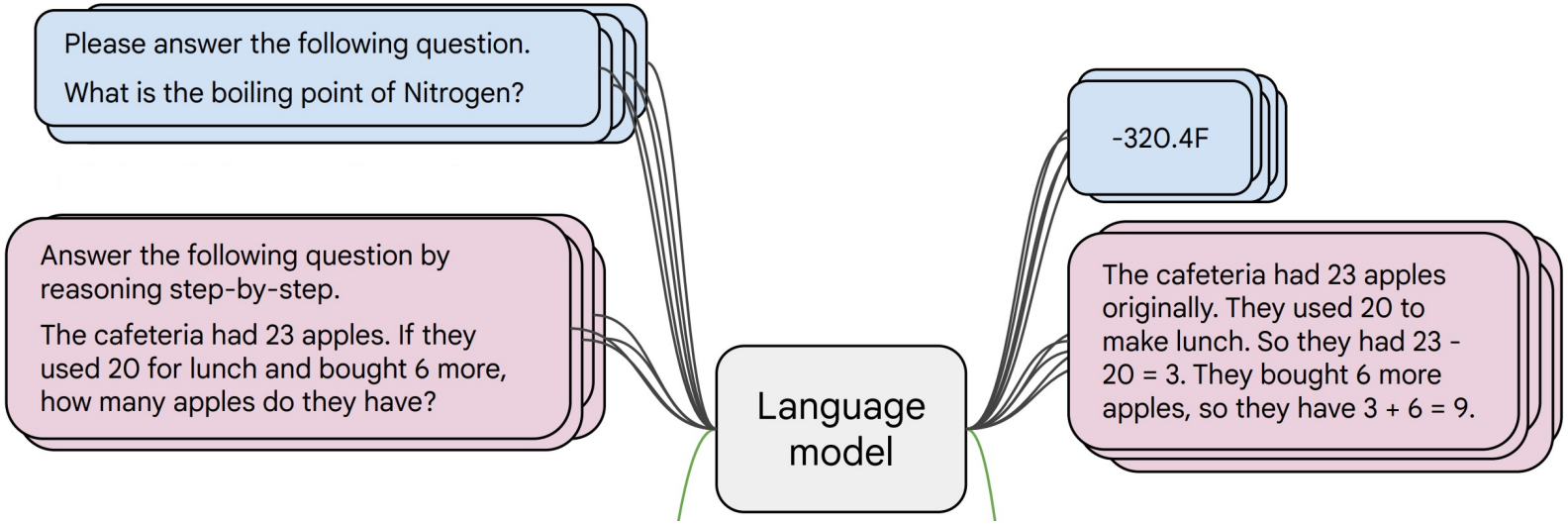
Step 2: Finetune (on many tasks)

~~Not~~ many labels; adapt to the tasks!



Instruction Finetuning

- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



- Evaluate on **unseen tasks**

[FLAN-T5; Chung et al., 2022]

Limitations of instruction finetuning

- One limitation of instruction finetuning is obvious: it's **expensive** to collect ground-truth data for tasks.
- But there are other, subtler limitations too. Can you think of any?
- **Problem 1:** tasks like open-ended creative generation have no right answer.
  - *E.g. Write me a story about a dog and her pet grasshopper.*
- **Problem 2:** language modeling penalizes all token-level mistakes equally, but some errors are worse than others.
- Even with instruction finetuning, there a mismatch between the LM objective and the objective of "satisfy human preferences"!
- Can we **explicitly attempt to satisfy human preferences? (alignment)**

Reward maximization from human

- Let's say we were training a language model on some task (e.g. summarization).
- For each LM sample s, imagine we had a way to obtain a *human reward* of that summary: $R(s) \in$ R, higher is better.
- Now we want to maximize the expected reward of samples from our LM:

$$\mathbb{E}_{\hat{s} \sim p_\theta(s)}[R(\hat{s})]$$

SAN FRANCISCO, California (CNN) -- A magnitude 4.2 earthquake shook the San Francisco ... overturn unstable objects.

An earthquake hit San Francisco. There was minor property damage, but no injuries.

s_1

$R(s_1) = 8.0$

The Bay Area has good weather but is prone to earthquakes and wildfires.

s_2

$R(s_2) = 1.2$

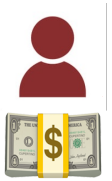
How do we model human preferences?

- Awesome: now for any **arbitrary, non-differentiable reward function** $R(s)$, we can train our language model to maximize expected reward.
- Not so fast! (Why not?)
- **Problem 1:** human-in-the-loop is expensive!
  - **Solution:** instead of directly asking humans for preferences, **model their preferences** as a separate (NLP) problem! [Knox and Stone, 2009]

s_1: `An earthquake hit San Francisco. There was minor property damage, but no injuries.`

$R(s_1) = 8.0$ 

s_2: `The Bay Area has good weather but is prone to earthquakes and wildfires.`

$R(s_2) = 1.2$ 

Train an LM $RM_\phi(s)$ to predict human preferences from an annotated dataset, then optimize for RM_ϕ instead.

How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
  - **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [Phelps et al., 2015; Clark et al., 2018; Jiang et al. 2011; Xu et al. 2011]

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.

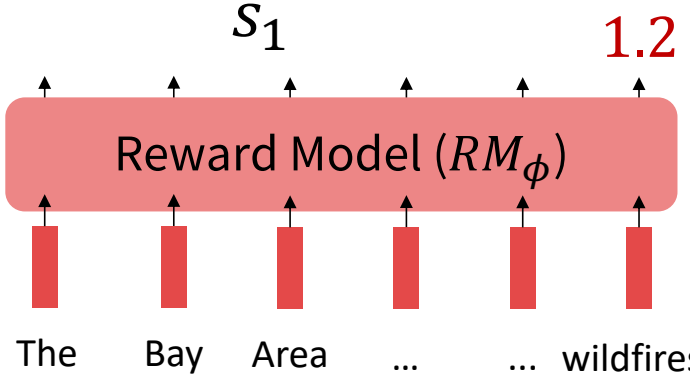
$$s_3$$

$$R(s_3) = \quad 4.1? \quad 6.6? \quad 3.2?$$

How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
  - **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [Phelps et al., 2015; Clark et al., 2018; Jiang et al. 2011; Xu et al. 2011]

`An earthquake hit San Francisco. There was minor property damage, but no injuries.` > `A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.` > `The Bay Area has good weather but is prone to earthquakes and wildfires.`



s_3 s_2

Bradley-Terry [1952] paired comparison model

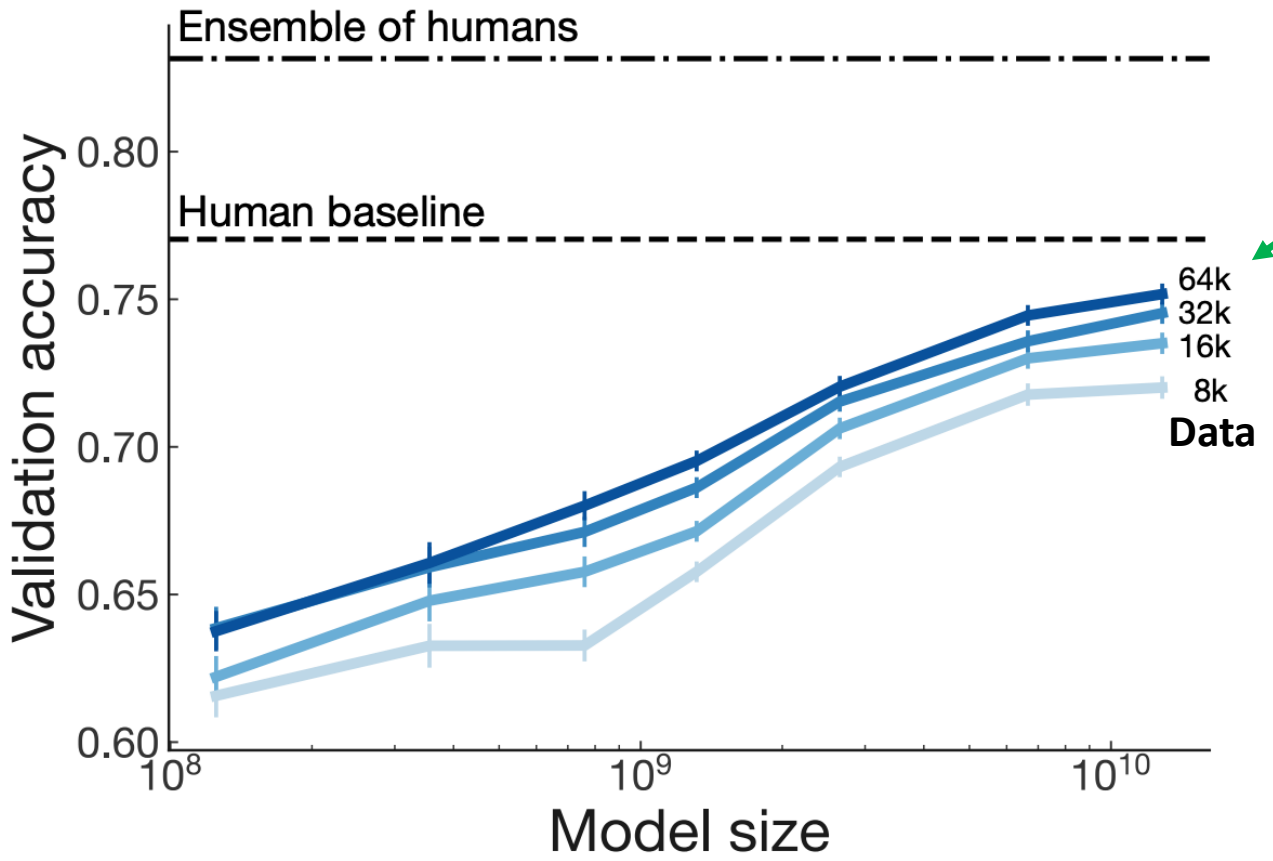
$$J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l)\sim D}\left[\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))\right]$$

"winning" sample — "losing" sample

s^w should score higher than s^l

Training a reward model first

Evaluate RM on predicting outcome of held-out human judgments

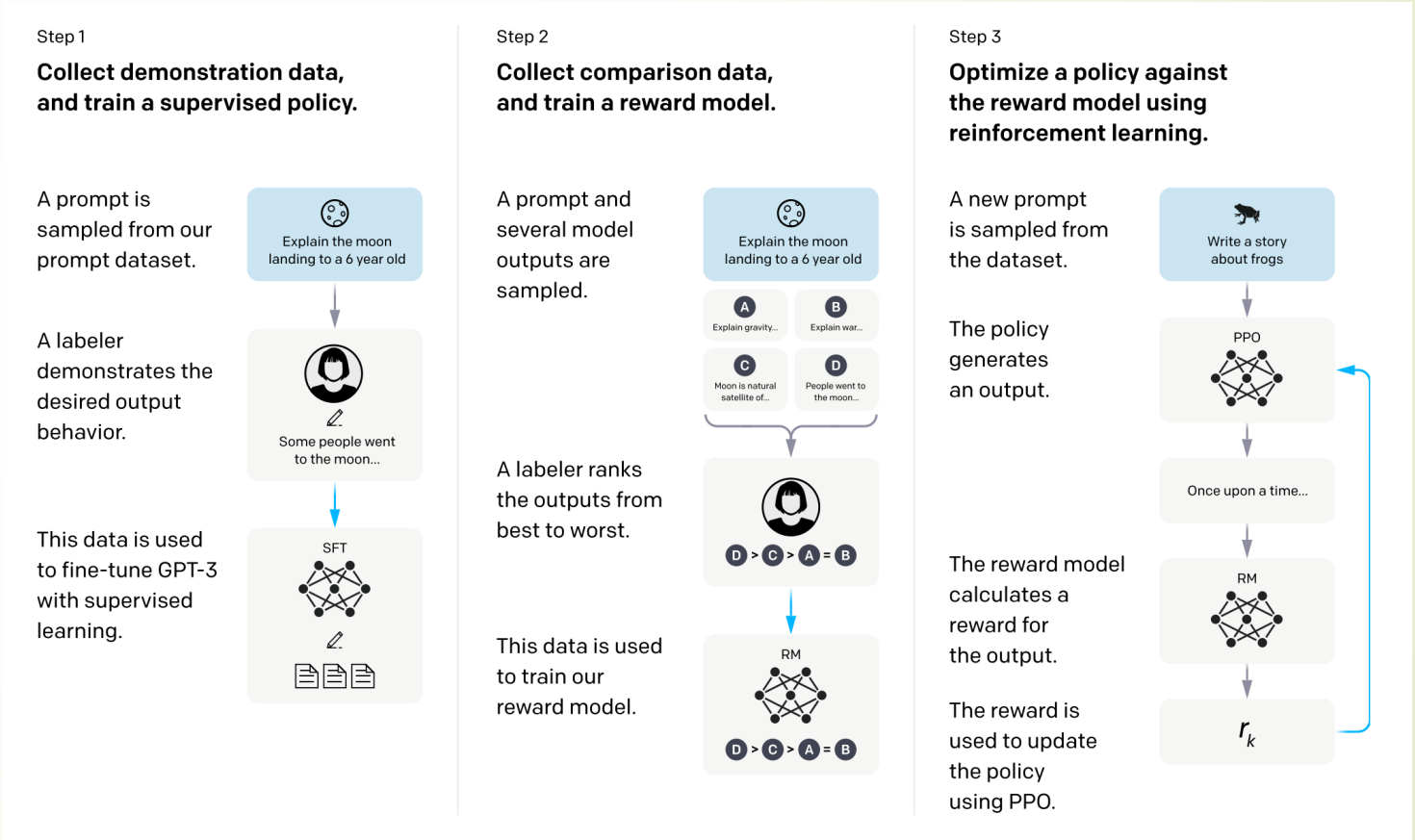


Large enough RM trained on enough data approaching single human perf

[Stiennon et al., 2020]

High-level instantiation: ‘RLHF’ pipeline

- First step: instruction tuning!
- Second + third steps: maximize reward (but how??)



Reinforcement learning to the rescue

- The field of **reinforcement learning (RL)** has studied these (and related) problems for many years now [Williams, 1992; Sutton and Barto, 1998]
- Circa 2013: resurgence of interest in RL applied to deep learning, game-playing [Mnih et al., 2013]
- But the interest in applying RL to modern LMs is an even newer phenomenon [Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022]. **Why?**
  - RL w/ LMs has commonly been viewed as very hard to get right (still is!)
  - Newer advances in RL algorithms that work for large neural models, including language models (e.g. PPO; [Schulman et al., 2017])

AlphaGo

Optimizing for human preferences

- How do we actually change our LM parameters θ to maximize this?

$$\mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})]$$

- Let's try doing gradient ascent!

$$\theta_{t+1} \coloneqq \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s}\sim p_{\theta_t}(s)}[R(\hat{s})]$$

How do we estimate this expectation??

What if our reward function is non-differentiable??

- **Policy gradient** methods in RL (e.g., [Williams, 1992]) give us tools for estimating and optimizing this objective.

A very brief introdution to Policy Gradient

- We want to obtain

(defn. of expectation) (linearity of gradient)

$$\nabla_\theta \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})] = \nabla_\theta \sum_s R(s) p_\theta(s) = \sum_s R(s) \nabla_\theta p_\theta(s)$$

- Here we'll use a very handy trick known as the **log-derivative trick**. Let's try taking the gradient of $\log p_\theta(s)$

$$\nabla_\theta \log p_\theta(s) = \frac{1}{p_\theta(s)} \nabla_\theta p_\theta(s) \qquad \Longrightarrow \qquad \nabla_\theta p_\theta(s) = p_\theta(s)\ \nabla_\theta \log p_\theta(s)$$

(chain rule)

- Plug back in:

This is an expectation of this

$$\sum_s R(s) \nabla_\theta p_\theta(s) = \sum_s p_\theta(s) R(s) \nabla_\theta \log p_\theta(s)$$

$$= \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s}) \nabla_\theta \log p_\theta(\hat{s})]$$

A very brief introdution to Policy Gradient

- Now we have put the gradient "inside" the expectation, we can approximate this objective with Monte Carlo samples:

$$\nabla_\theta \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})] = \mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})\,\nabla_\theta \log p_\theta(\hat{s})] \approx \frac{1}{m}\sum_{i=1}^{m} R(s_i)\,\nabla_\theta \log p_\theta(s_i)$$

This is why it's called **"reinforcement learning"**: we **reinforce** good actions, increasing the chance they happen again.

- Giving us the update rule:

$$\theta_{t+1} := \theta_t + \alpha \frac{1}{m}\sum_{i=1}^{m} R(s_i)\,\nabla_{\theta_t} \log p_{\theta_t}(s_i)$$

If R is +++

Take gradient steps to maximize $p_\theta(s_i)$

If R is ---

Take steps to minimize $p_\theta(s_i)$

This is **heavily simplified**! There is a *lot* more needed to do RL w/ LMs. **Can you see any problems with this objective?**

RLHF: Putting it all together [Christiano et al., 2017; Stiennon et al., 2020]

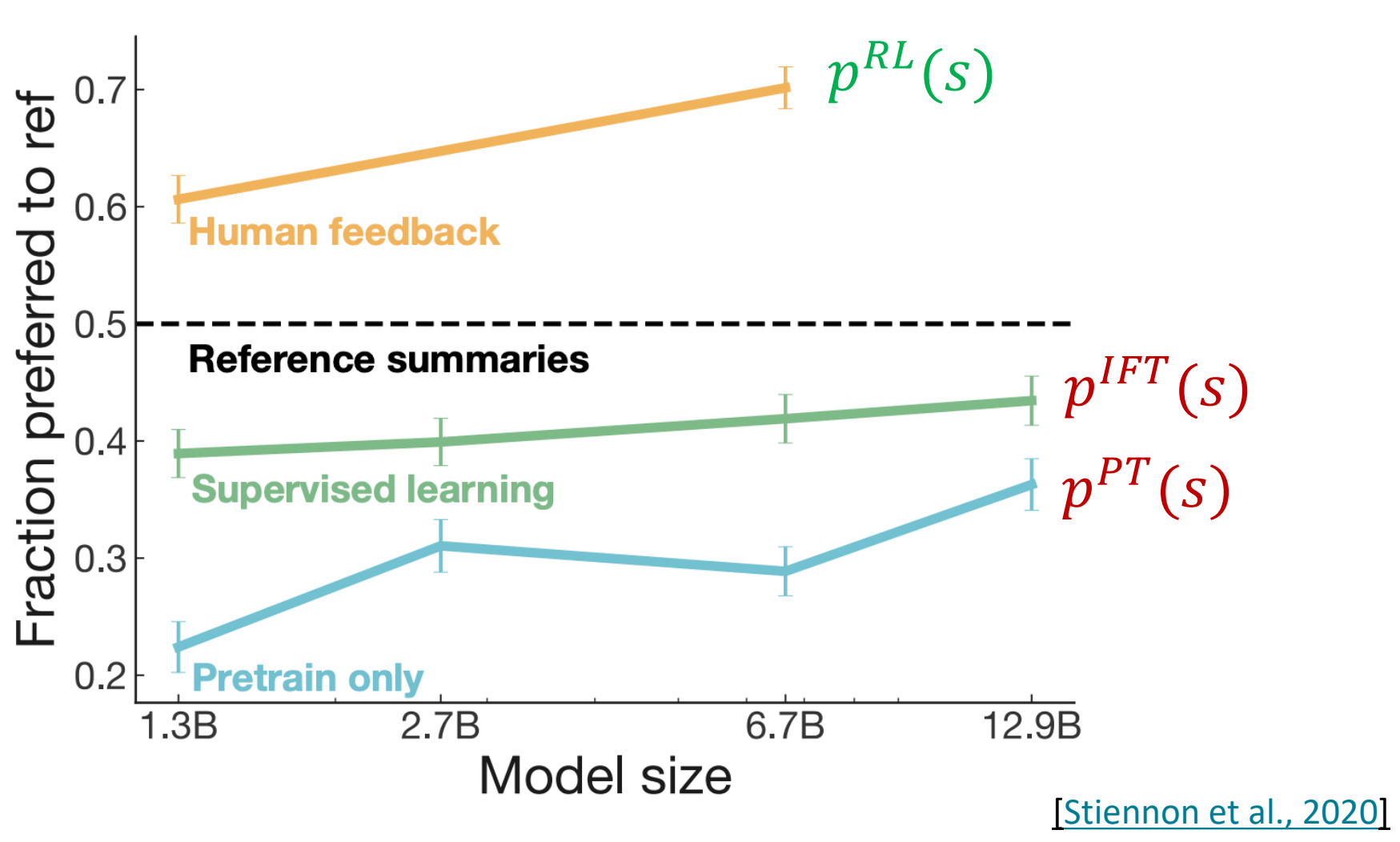
- Finally, we have everything we need:
  - A pretrained (possibly instruction-finetuned) LM $p^{\text{PT}}(s)$
  - A reward model $RM(s)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
  - A method for optimizing LM parameters towards an arbitrary reward function: *policy gradient*
- Now to do RLHF:
  - Initialize a copy of the model $p_\theta^{\text{RL}}(s)$, with parameters θ we would like to optimize
  - Optimize the following reward with RL:

$$R(s) = RM_\phi(s) - \underbrace{\beta \log \left(\frac{p_\theta^{RL}(s)}{p^{PT}(s)} \right)}$$

Pay a price when $p_\theta^{RL}(s) > p^{PT}(s)$

This is a penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler** (**KL**) divergence between $p_\theta^{RL}(s)$ and $p^{PT}(s)$.

RLHF improves over pretraining and finetuning



[Stiennon et al., 2020]

ChatGPT: Instruction Finetuning + RLHF for dialog agents [https://openai.com/blog/chatgpt/]

ChatGPT: Optimizing Language Models for Dialogue

Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge…

Methods

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

(RLHF!)

Reinforcement Learning for Quantitative Trading

FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance, Liu et al. Deep RL Workshop, NeurIPS 2020.

https://github.com/AI4Finance-Foundation/FinRL

Why RL for Trading?

1. Modern Portfolio Theory (MPT) performs not well in out-of-sample data, sensitive to outliers and only based on stock returns.
2. Goal of stock trading: maximize returns.
3. DRL solves optimization problems by maximizing the expected total reward defined as future returns, without human labels

Trading Markov Decision Process

- Trading agent is modeled as a Markov Decision Process (MDP)
- Note that this Markov process might not be stationary or static
- Components:
  - **State**
    - $\boldsymbol{s} = [\boldsymbol{p}, \boldsymbol{h}, \boldsymbol{f}, b]$, $\boldsymbol{p}$: stock prices, $\boldsymbol{f}$ **:** features**,** $\boldsymbol{h}$: stock shares, b: remaining balance
  - **Action**
    - Three actions: **a** ∈ {−1, 0, 1}, where −1, 0, 1 represent selling, holding, and buying one stock.
    - Multiple action space **a** ∈ {−k, ..., −1, 0, 1, ..., k}, where k denotes the number of shares.
    - An action can be carried upon multiple shares. For example, "Buy 10 shares of AAPL" or "Sell 10 shares of AAPL" are 10 or −10, respectively. Resulting in $(2k+1)^d$ actions for d stocks.
  - **Reward**
    - $r(s,a,s')$: the direct reward of acting a at state s and arriving at the new state s', e.g. the change of the portfolio value when action a is taken at state s and arriving at new state s', i.e., r(s, a, s') = v' − v, where v' and v represent the portfolio values at state s' and s, respectively'.
  - Q-value function
    - $Q_\pi\ (s, a)$: the expected reward of acting a at state s following policy π

State Space

- State Space
  - **Balance**: available amount of money left in the account currently
  - **Price**: current adjusted close price of each stock
  - **Shares**: shares owned of each stock
  - **ADX**: *Average Directional Index*, is a trend strength indicator.
  - **MACD**: *Moving Average Convergence Divergence,* is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price. The MACD is calculated by subtracting the 26-period exponential moving average (EMA) from the 12-period EMA.
  - **RSI**: *Relative Strength Index*, is classified as a momentum oscillator, measuring the velocity and magnitude of directional price movements
  - **CCI**: *Commodity Channel Index,* is a momentum-based oscillator used to help determine when an investment vehicle is reaching a condition of being overbought or oversold.
  - One could use language models such as LSTM to extract more features.

Action space

- **Action**
  - Three actions: $a \in \{-1, 0, 1\}$, where $-1, 0, 1$ represent selling, holding, and buying one stock.
  - Multiple action space $a \in \{-k, \ldots, -1, 0, 1, \ldots, k\}$, where k denotes the number of shares one can buy or sell.
  - An action can be carried upon multiple stocks. Therefore the size of the enire action space is $(2k+1)^d$ where d is the number of stocks.
  - For example, "Buy 10 shares of AAPL" or "Sell 10 shares of AAPL" are $a=10$ or $a=-10$, respectively.

Reward function

- **Reward**
  - $r(s,a,s')$: the direct reward of acting a at state s and arriving at the new state s'
  - For example, the change of the portfolio value when action a is taken at state s and arriving at new state s', i.e., r(s, a, s') = v' − v, where v' and v represent the portfolio values at state s' and s, respectively'
  - Transaction cost is usually involved
  - One can also use *Sharpe ratio* as reward,

The Formula for Sharpe Ratio Is

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

where:

R_p = return of portfolio

R_f = risk-free rate

σ_p = standard deviation of the portfolio's excess return

Constraints

- **Market liquidity:**
  - Assume that stock market will not be affected by our reinforcement trading agent
- **Nonnegative balance:**
  - the allowed actions should not result in a negative balance.
- **Transaction cost:**
  - transaction costs are incurred for each trade.
- **Risk-aversion for market crash:**
  - employ the financial **turbulence index** that measures extreme asset price movements.

Learning Algorithms

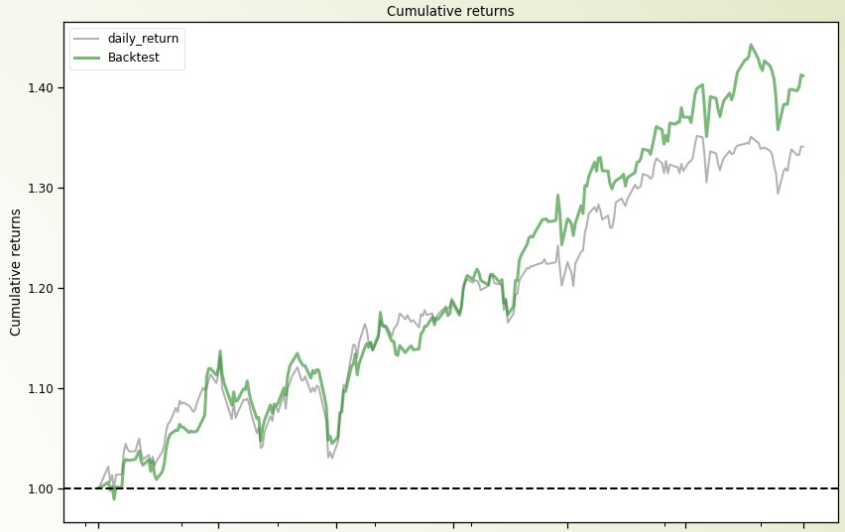
- Critic-only approach
  - Q-learning, DQN, etc
- Actor-only approach
  - Policy Gradient
- Actor-critic approach
  - A2C
  - PPO
  - DDPG
  - **SAC**

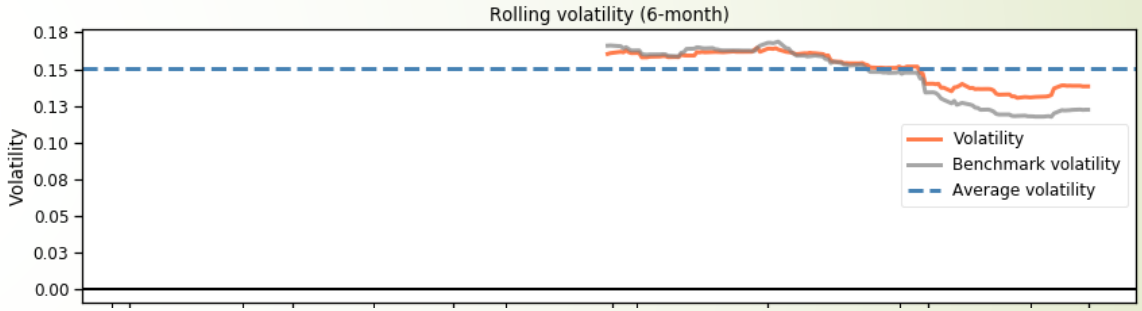
Data

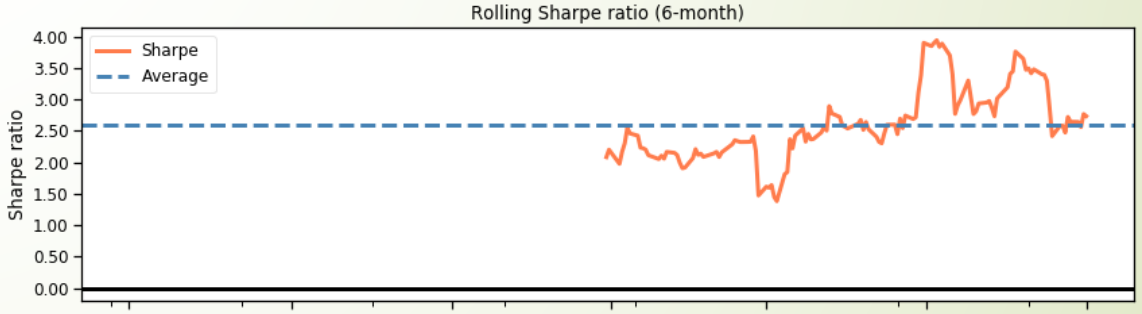
- Dow 30 constituents:
  - ['AXP', 'AMGN', 'AAPL', 'BA', 'CAT', 'CSCO', 'CVX', 'GS', 'HD', 'HON', 'IBM', 'INTC', 'JNJ', 'KO', 'JPM', 'MCD', 'MMM', 'MRK', 'MSFT', 'NKE', 'PG', 'TRV', 'UNH', 'CRM', 'VZ', 'V', 'WBA', 'WMT', 'DIS', 'DOW']
- Training
  - Daily OHLC prices and features from '2009-01-01' to '2020-07-01'
  - N = 83897
- BackTest trading
  - Daily OHLC prices and features from '2020-07-01' to '2021-07-06'
  - N = 7337
  - Baseline: Dow Jones Index (DJI)

A successful SAC agent

- SAC:
  - Annual return 0.409532
  - Cumulative returns 0.411453
  - Annual volatility 0.149417
  - Sharpe ratio 2.382402
- Baseline: DJI
  - Annual return 0.335107
  - Cumulative returns 0.336639
  - Annual volatility 0.145596
  - Sharpe ratio 2.066650







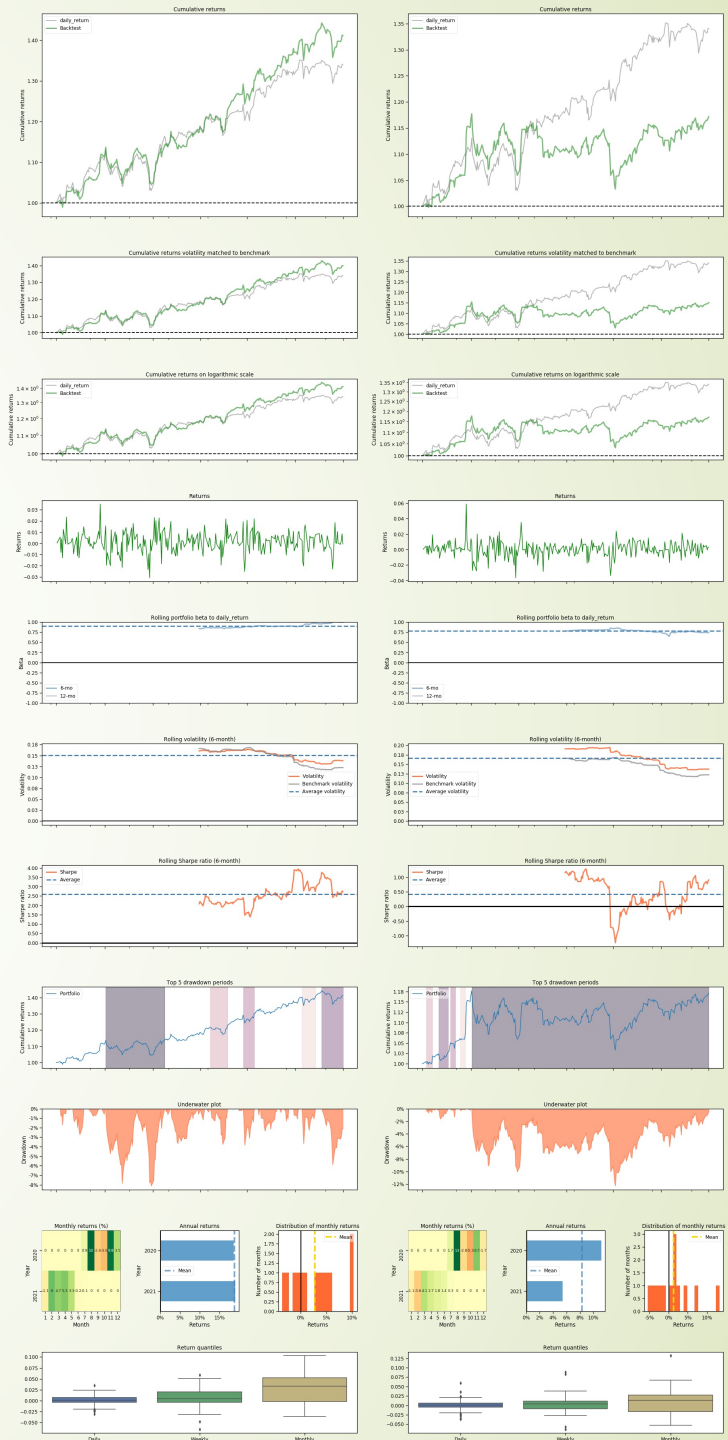
RL may be highly instable: two SAC runs

Good

- Results:
  - Annual return 0.409532
  - Cumulative returns 0.411453
  - Annual volatility 0.149417
  - Sharpe ratio 2.382402

Bad

- Results
  - Annual return 0.250596
  - Cumulative returns 0.251707
  - Annual volatility 0.148737
  - Sharpe ratio 1.584268



Case: Hierarchical Reinforced Trader (HRT)
(Zhao & Welsch, arXiv: 2410.14927)

- *Hierarchical Reinforced Trader (HRT): A Bi-Level Approach for Optimizing Stock Selection and Execution*, by Zhao and Welsch, **https://arxiv.org/abs/2410.14927**
- **High Level Controller (HLC):** determine the *subset of stocks* to buy, sell, or hold, executing stock selection
- **Low Level Controller (LLC):** optimize the trade volumes for the selected stocks, thereby determining the optimal number of shares to transact

Hierarchical Reinforced Trader (HRT): A Bi-Level Approach for Optimizing Stock Selection and Execution

Zijie Zhao
Massachusetts Institute of Technology
Cambridge, MA, USA
zijiezha@mit.edu

Roy E. Welsch
Massachusetts Institute of Technology
Cambridge, MA, USA
rwelsch@mit.edu

ABSTRACT

Leveraging Deep Reinforcement Learning (DRL) in automated stock trading has shown promising results, yet its application faces significant challenges, including the curse of dimensionality, inertia in trading actions, and insufficient portfolio diversification. Addressing these challenges, we introduce the **H**ierarchical **R**einforced **T**rader (**HRT**), a novel trading strategy employing a bi-level Hierarchical Reinforcement Learning framework. The HRT integrates a Proximal Policy Optimization (PPO)-based High-Level Controller (HLC) for strategic stock selection with a Deep Deterministic Policy Gradient (DDPG)-based Low-Level Controller (LLC) tasked with optimizing trade executions to enhance portfolio value. In our empirical analysis, comparing the HRT agent with standalone DRL models and the S&P 500 benchmark during both bullish and bearish market conditions, we achieve a positive and higher Sharpe ratio. This advancement not only underscores the efficacy of incorporating hierarchical structures into DRL strategies but also mitigates the aforementioned challenges, paving the way for designing more profitable and robust trading algorithms in complex markets.

CCS CONCEPTS

• **Computing methodologies** → **Reinforcement learning**.

KEYWORDS

Deep Reinforcement Learning, Markov Decision Process, Automated Stock Trading, Hierarchical Reinforcement Learning

1 INTRODUCTION

Profitable automated stock trading strategies are pivotal for investment companies and hedge funds. A classical method is Harry Markowitz's Modern Portfolio Theory (MPT) [12], which determines the optimal portfolio allocation by calculating the expected returns and the covariance matrix of stock prices. This optimization aims to either maximize returns for a given risk level or minimize risk for a specified return range. However, implementing MPT can be complex, especially when portfolio managers wish to dynamically adjust decisions at each time step and consider additional factors. An alternative approach models the stock trading problem as a Markov Decision Process (MDP) [1], solved using dynamic programming. Nevertheless, this model's scalability is constrained by the expansive state spaces inherent in real stock markets.

Recent research has turned to Deep Reinforcement Learning (DRL) methods for stock trading [4, 22]. DRL overcomes scalability issues by using deep neural networks to approximate complex functions, solving MDPs without the limitations of traditional models. Liu, Xiao-Yang, et al. [9] formalize the stock trading problem as an MDP and employ Deep Deterministic Policy Gradient (DDPG) [7] to discover optimal trading strategies that yield higher cumulative returns and Sharpe ratios in the volatile stock market. Subsequent research integrates the strengths of DDPG, Proximal Policy Optimization (PPO) [18], and Advantage Actor Critic (A2C) [14] into an ensemble strategy [22], adapting robustly to varying market conditions. Despite these advancements, several challenges persist in applying DRL to stock trading:

- **Curse of Dimensionality**: The computational complexity, sample inefficiency, and potential training instability escalate as the number of stocks increases, expanding the dimensionality of data and the state and action spaces exponentially. For instance, if the action for a single stock is defined as $a \in \{-k, \ldots, -1, 0, 1, \ldots, k\}$, representing sell, hold, and buy actions, the action space becomes $(2\times k+1)^N$, where N is the number of market stocks. This complexity has limited the validation of current research to a small asset scale, ranging from Dow Jones 30 constituent stocks to only tens of assets.
- **Inertia or Momentum Effect**: DRL agents tend to repeat a previous action (buy, sell, or hold) based on the reward received, without necessarily considering the currently most profitable action. If an agent receives a large reward for a particular action (buy, sell, or hold), it may exploit this action in subsequent steps. Even though DDPG introduces action exploration through the addition of noise to the actions selected by its deterministic policy, we still observe crowded or clustered trading operations in **Figure 1** under the example of Dow Jones 30 constituent stocks portfolio.
- **Insufficient Diversification**: Diversification, a core principle of finance aimed at risk mitigation, is compromised when DRL agents focus repeatedly on a narrow selection of stocks. This behavior, evidenced in **Figure 1**, increases exposure to sector-specific risks, making the portfolio more susceptible to adverse developments within those sectors.

To mitigate the three issues mentioned above and to enhance performance and deliver superior trading strategies, we introduce the **H**ierarchical **R**einforced **T**rader (**HRT**), an innovative approach to stock trading that utilizes a Hierarchical Reinforcement Learning (HRL) framework [16]. Our HRT agent is structured around two principal components, each serving distinct but complementary roles in the trading strategy: (1) **High-Level Controller (HLC)**: Positioned at the strategic apex of the hierarchy, the HLC's mandate is to determine the subset of stocks to buy, sell, or hold, effectively executing stock selection. (2) **Low-Level Controller (LLC)**: ollowing the HLC's directives, the LLC is tasked with refining these decisions by optimizing the trade volumes for the selected stocks, thereby determining the optimal number of shares to transact. By dividing the trading strategy into high-level stock selection and

HRT scheme

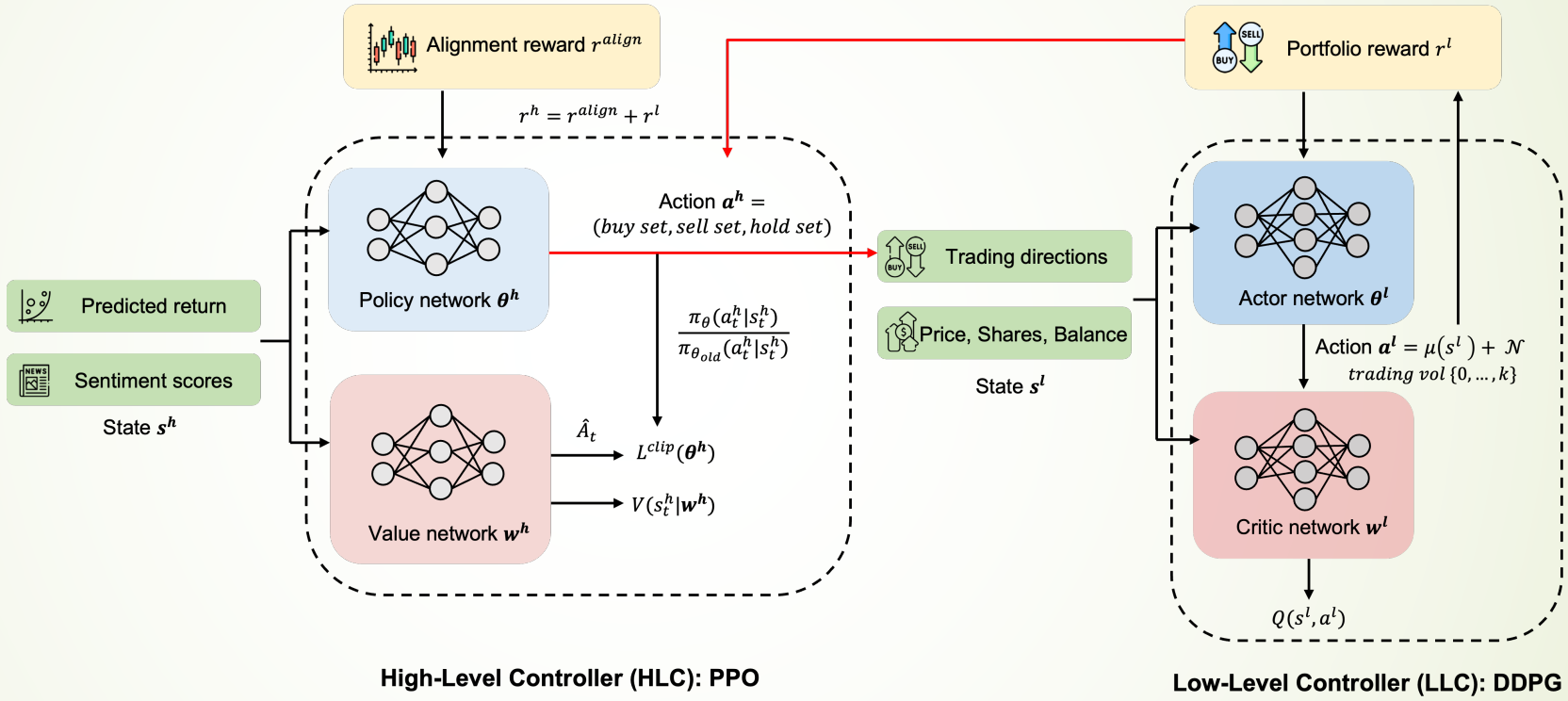
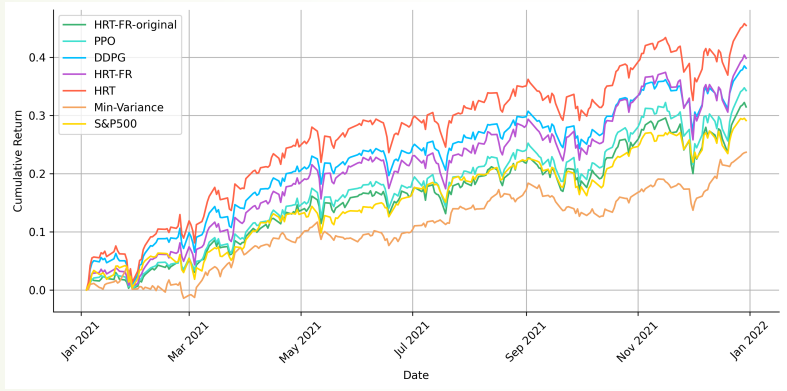


Figure 2: Overview of the Hierarchical Reinforced Trader (HRT) architecture. Interactions between the HLC and LLC are indicated by the red arrows.

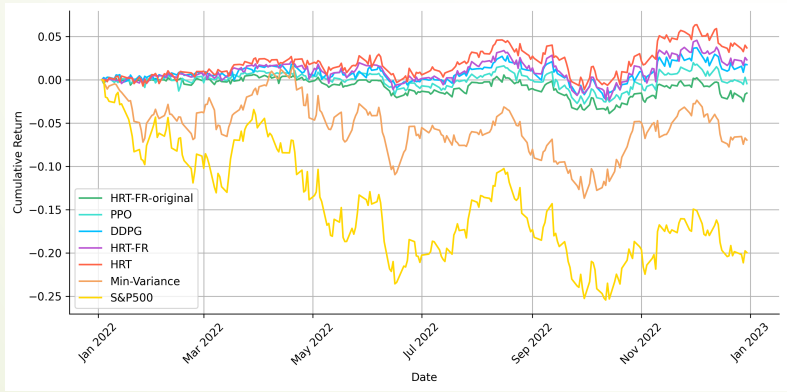
Note: Transformer Encoders and LLaMA 2 13B sentiment analysis are used.

Evaluations on 2021/2022

Cumulative Returns



(a) Year 2021 (bullish market)



(b) Year 2022 (bearish market)

Other metrics

(a) Year 2021 (bullish market)

Metric	HRT-FR-original	PPO	DDPG	HRT-FR	HRT	Min-Var	S&P500
Cumulative Return	0.3147 ± 0.010	0.3428 ± 0.009	0.3813 ± 0.008	0.3983 ± 0.007	0.4548 ± 0.008	0.2368	0.2913
Annualized Return	0.3200 ± 0.009	0.3486 ± 0.009	0.3879 ± 0.007	0.4051 ± 0.007	0.4628 ± 0.008	0.2406	0.2961
Annualized Volatility	0.1489 ± 0.010	0.1498 ± 0.008	0.1458 ± 0.005	0.1665 ± 0.009	0.1687 ± 0.009	0.0980	0.1303
Sharpe Ratio	2.1494 ± 0.156	2.3274 ± 0.138	2.6601 ± 0.103	2.4336 ± 0.138	2.7440 ± 0.154	2.4549	2.2736
Max Drawdown	-0.0738 ± 0.018	-0.0808 ± 0.010	-0.0651 ± 0.013	-0.0845 ± 0.010	-0.0755 ± 0.016	-0.0516	-0.0521

(b) Year 2022 (bearish market)

Metric	HRT-FR-original	PPO	DDPG	HRT-FR	HRT	Min-Var	S&P500
Cumulative Return	-0.0154 ± 0.008	-0.0045 ± 0.004	0.0174 ± 0.005	0.0229 ± 0.005	0.0368 ± 0.004	-0.0696	-0.1995
Annualized Return	-0.0156 ± 0.007	-0.0045 ± 0.003	0.0176 ± 0.005	0.0232 ± 0.005	0.0372 ± 0.005	-0.0704	-0.2016
Annualized Volatility	0.0586 ± 0.008	0.0646 ± 0.007	0.0790 ± 0.008	0.0893 ± 0.005	0.0901 ± 0.005	0.1513	0.2416
Sharpe Ratio	-0.2664 ± 0.050	-0.0701 ± 0.037	0.2224 ± 0.059	0.2594 ± 0.045	0.4132 ± 0.048	-0.4654	-0.8344
Max Drawdown	-0.0448 ± 0.009	-0.0431 ± 0.008	-0.0484 ± 0.007	-0.0554 ± 0.008	-0.0548 ± 0.009	-0.1507	-0.2543

Summary

- Model-free reinforcement learning trading
- RL agent is unstable:
  - The reward is highly noisy
  - The environment in stock prices is not stationary
  - RL itself might not be stable
  - Perhaps consider multiple agents

Optimized Execution, Market Microstructure and Reinforcement Learning

[Y. Nevmyvaka. Y. Feng, MK; ICML 2006]
[MK, Y. Nevmyvaka; In "High Frequency Trading", O'Hara et al. eds, Risk Books 2013]

Michael Kearns, University of Pennsylvania, ICML 2014, Beijing

A Brief Field Guide to Wall Street

- "Buy Side": Attempt to outperform market via proprietary research
  - Includes hedge funds, mutual funds, statistical arbitrage, HFT, prop trading groups
  - May or may not be quantitative and automated
  - Have investors but not clients
  - Take and hold positions → risk
  - Generation of "alpha" still more art than science
- "Sell Side": Provide brokerage and execution services
  - Includes bank and independent brokerages, exchanges
  - Almost entirely quantitative and automated
  - Clients are the buy side
  - Do not hold risk; paid via fees/commissions/etc.
- In reality, alpha and execution are blurred
  - Especially at shorter holding periods (e.g. HFT)

A Canonical Trading Problem

- Goal (buy side to sell side): Sell V shares in T time steps; maximize revenue
- Strategy Evaluation Metric Benchmarks:
  - Volume Weighted Average Price (VWAP)
  - Time Weighted Average Price (TWAP)
  - Implementation Shortfall (midpoint of bid-ask spread at beginning)
- Natural to view as a problem of *state-based control (RL)*
  - State variables: inventory V and time remaining T (discretized)
  - Features capturing market activity?

Market Microstructure

refresh | island home | disclaimer | help

GET STOCK

MSFT

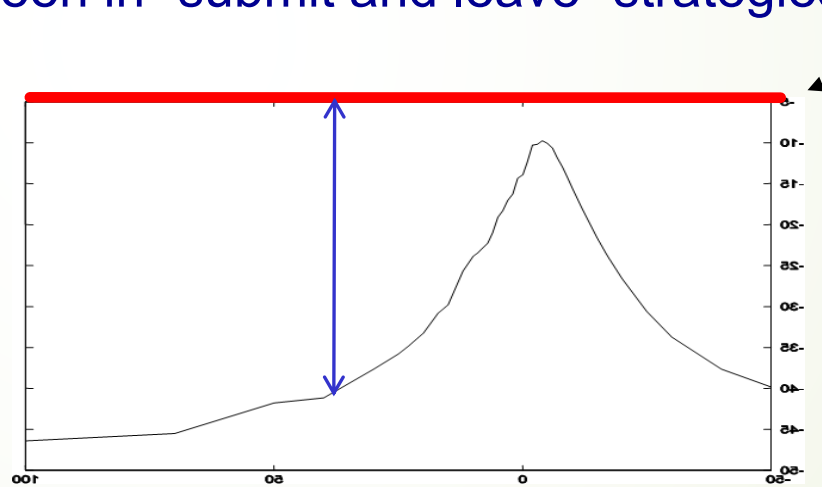
MSFT go

Symbol Search

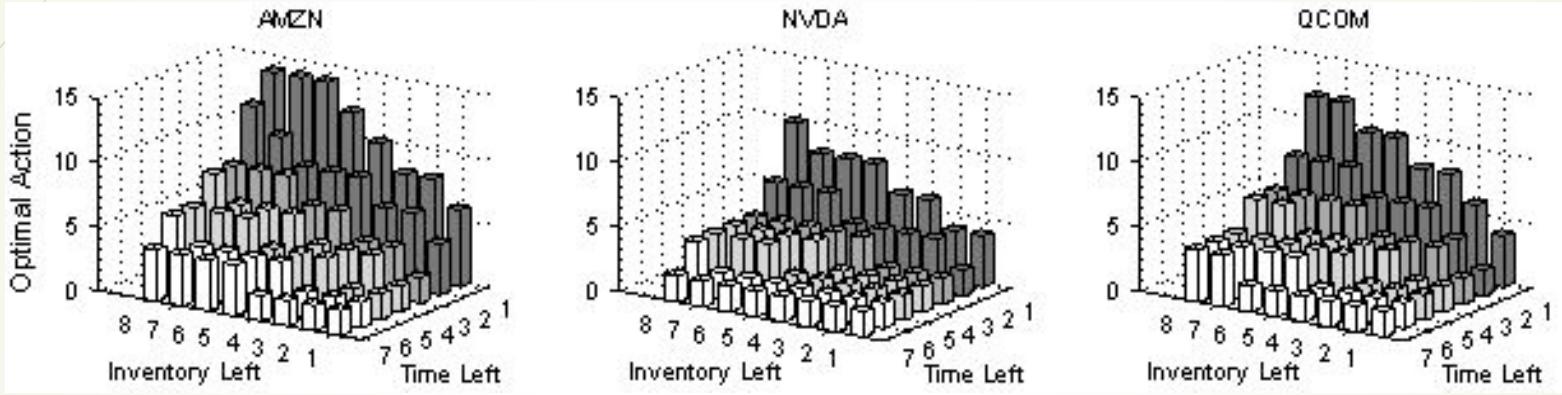
LAST MATCH		TODAY'S ACTIVITY	
Price	23.7790	Orders	1,630
Time	9:01:55.614	Volume	44,839

BUY ORDERS		SELL ORDERS	
SHARES	PRICE	SHARES	PRICE
1,000	23.7600	100	23.7800
3,087	23.7500	800	23.7990
200	23.7500	500	23.8000
100	23.7400	1,720	23.8070
1,720	23.7280	900	23.8190
2,000	23.7200	200	23.8500
1,000	23.7000	1,000	23.8500
100	23.7000	1,000	23.8500
100	23.7000	1,000	23.8600
800	23.6970	200	24.0000
500	23.6500	500	24.0000
3,000	23.6500	1,000	24.0300
4,300	23.6500	200	24.0300
2,000	23.6500	1,100	24.0400
200	23.6200	500	24.0500
(195 more)		(219 more)	

- Continuous double auction with limit orders: buy orders decreasing; sell orders increasing
- Volatile and dynamic; sub-millisecond time scale
- Cancellations, revisions, partial executions
- How do individual orders (micro) influence aggregate market behavior (macro)?
- Tradeoff between *immediacy* and *price*
- Seen in “submit and leave” strategies:



Policies Learned: Time and Volume Remaining



- Experimental framework
  - Full historical order book reconstruction and simulation
  - Learn optimal policy on 1 year training; test on following 6 months
  - Pitfalls: directional drift, “counterfactual” market impact
- Overall shape is consistent and sensible
  - Become more aggressive (spread crossing) as time runs out or inventory is too large
  - Learning optimizes this qualitative schedule

Additional Improvement From Order Book Features

Bid Volume	-0.06%	Ask Volume	-0.28%
Bid-Ask Volume Misbalance	0.13%	**Bid-Ask Spread**	**7.97%**
Price Level	0.26%	**Immediate Market Order Cost**	**4.26%**
Signed Transaction Volume	**2.81%**	Price Volatility	-0.55%
Spread Volatility	1.89%	Signed Incoming Volume	0.59%
Spread + Immediate Cost	**8.69%**	**Spread+ImmCost+Signed Vol**	**12.85%**

Some Idealized Trading Scenarios and Risks

- Assume all the transactions cross the bid/ask spread at approximate midpoint (median) price
  - Example: V={1,0,-1} (long/nothing/short), T=1 min
- *Return* maximization with no-regret sequential (online) strategies:
  - Compete with best single strategy in hindsight
  - Unfortunately methods work poorly in practice
- Could ask for no-regret to best strategy in *risk-adjusted metrics:*
  - Sharpe Ratio: μ(returns)/σ(returns)
  - Mean-Variance: μ(returns) - σ(returns)
- Yet strong negative results in risk-adjusted metrics:
  - No-regret provably impossible
  - 1 + ε lower bound on competitive ratio
- Intuition: Volatility terms σ introduce additional costs that one has to pay
- Loss design should incorporate risk measurements, or internalize risks in strategies

Online Tutorials

- A GitHub repo for *deep reinforcement learning strategies and environments for quantitative trading*
  - https://github.com/Ceruleanacg/Personae/blob/master/README.md
  - This is a good start for the application of deep reinforcement learning in algorithmic trading
  - Can you **reproduce** the results there?

Thank you!