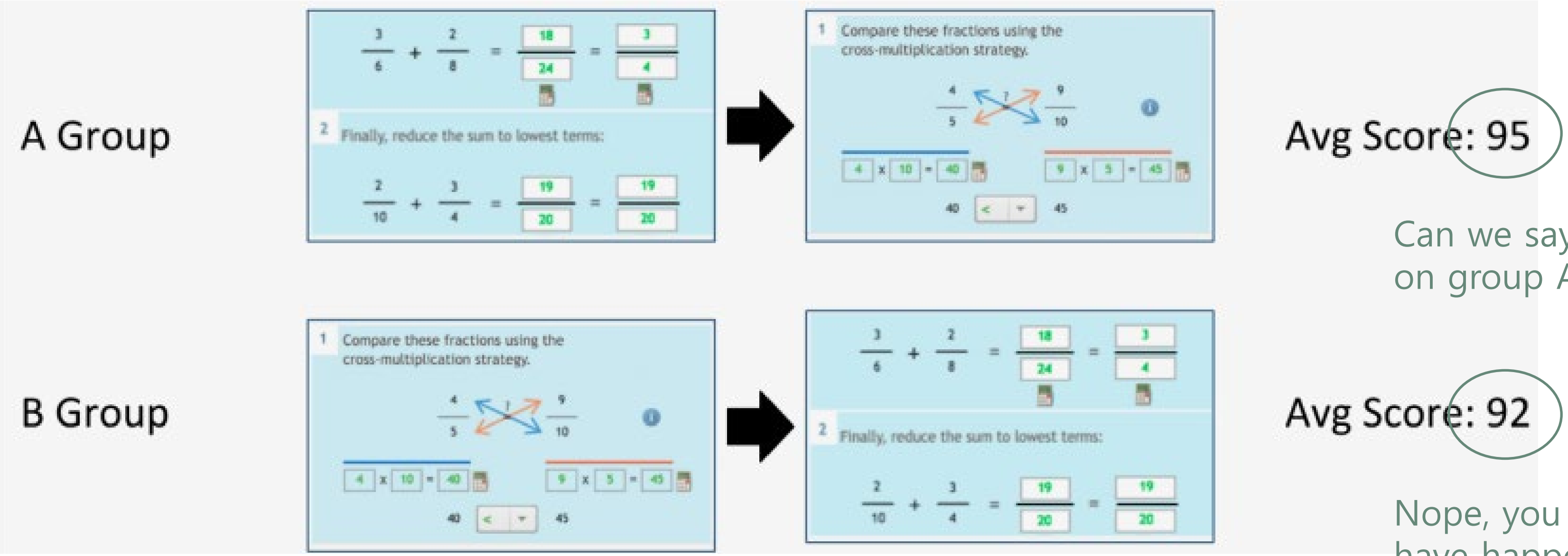
Lecture 15 : Batch RL

Why

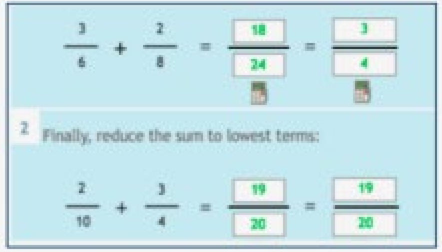


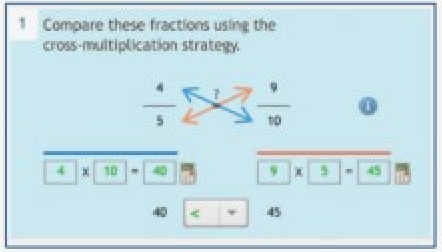
Can we say that method applied on group A is good enough?

Nope, you don't know what would have happened if group B went through same sequence of interventions as group A

Why

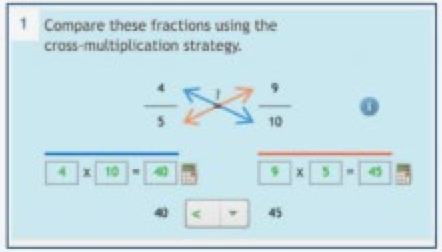
A Group

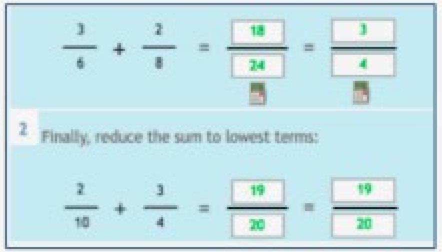




Avg Score: 95

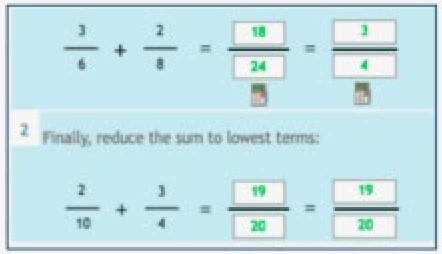
B Group

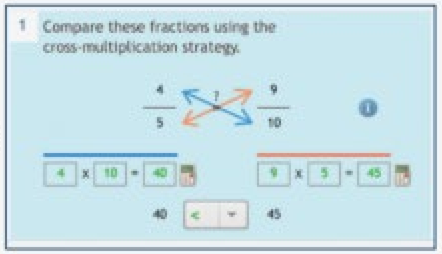




Avg Score: 92

B Group





???

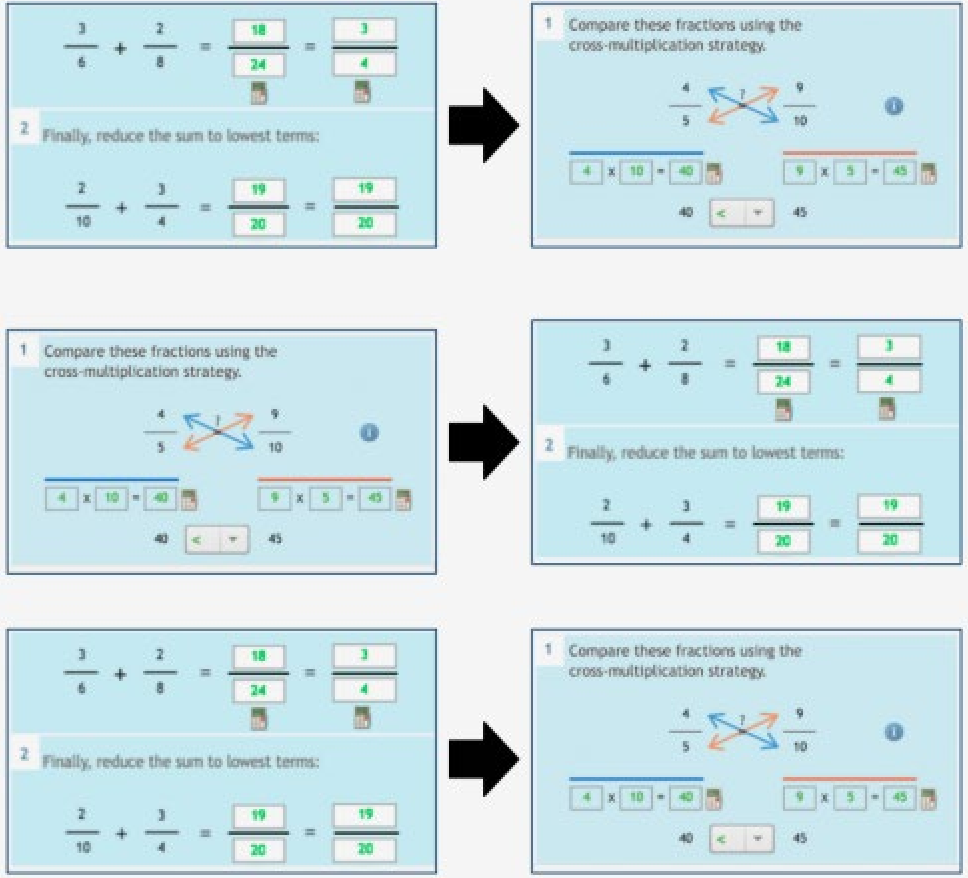
Test the first method by trying it and comparing it to what group B previously had done

Why

A Group

B Group

B Group



Avg Score: 95

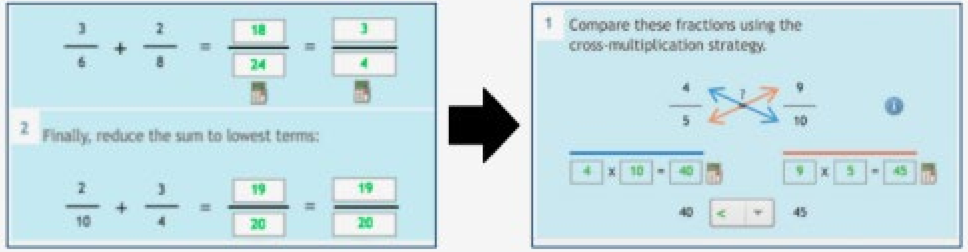
Avg Score: 92

???

So is this good enough?

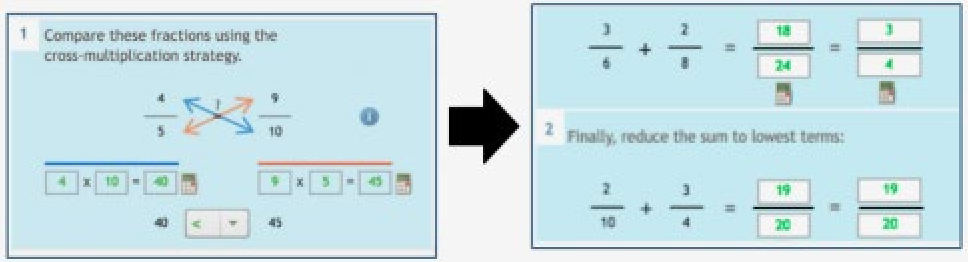
Why

A Group



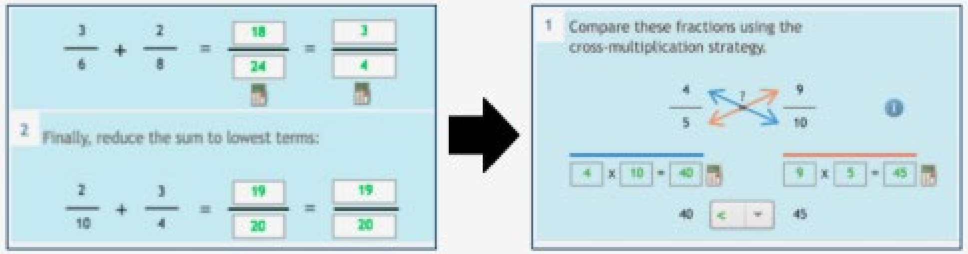
Avg Score: 95

B Group



Avg Score: 92

B Group



???

So is this good enough?

Why

Generalization

Fair, but can improve

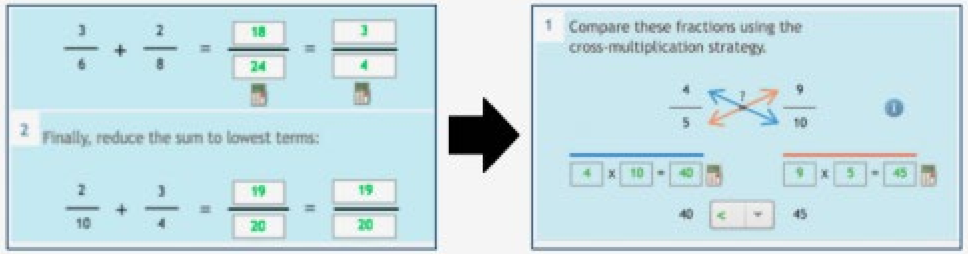
Q1. check on the meaning

In this case, generalization would mean making sure all candidate methods have been used to all possible groups to get rid of that factor that certain methods return unusual results when applied on certain groups

So, we want to reach towards a point where we know that all candidate methods can be generalized: know what averaged return they will give when applied to almost all of the possible groups to be tested on.

Improvement would include testing group A in the sequence that group B originally has gone through.

B Group



???

Why

Generalization

Connects…

The idea of using the data that already has been collected and making the most out of it to make decisions on moving forward

Especially when making bad decisions can be costly or dangerous

Similarly if obtaining more data itself is costly or not possible

Q2 Would this mean: obtaining the most generalized result from given data?

Why

Generalization

Connects...

The idea of using the data that already has been collected and making the most out of it to make decisions on moving forward

Especially when making bad decisions can be costly or dangerous

Similarly if obtaining more data itself is costly or not possible

→ Safe batch reinforcement learning

Safe Batch Reinforcement Learning

Some probability that the new policy will not be worse than the old policy

$$\Pr(V^{\mathcal{A}(\mathcal{D})} \geq V^{\pi_b}) \geq 1 - \delta$$

Easier to calc

$$\Pr(V^{\mathcal{A}(\mathcal{D})} \geq V_{min}) \geq 1 - \delta$$

Safe batch reinforcement learning

Safe Batch Reinforcement Learning

Off-policy Policy Evaluation (OPE)

Use data to make estimate of its Value function

High-confidence Off-policy Policy Evaluation (HCOPE)

Add the idea of confidence on the lower bound of value function

Safe Policy Improvement (SPI)

Safe batch reinforcement learning algorithm = which would mean finding improved policies

Safe Batch Reinforcement Learning

Historical Data, D
Proposed Policy, π_e

→ Estimate of V^{π_e} → Importance Sampling (IS)

Historical Data, D
Proposed Policy, π_e
Probability, $1-\delta$

→ $1-\delta$ confidence lower bound on $J(\pi_e)$

Historical Data, D
Probability, $1-\delta$

→ New policy π, or `No Solution Found`

Importance Sampling

$$IS(D) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}\right)\left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

Importance Sampling (IS)

Importance Sampling

$$IS(D) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}\right)\left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

$$\prod \frac{p(a_j \mid s_j)^{\pi_e}}{p(a_j \mid s_j)^{\pi_b}} \qquad G(h_j)$$

Importance Sampling

$$IS(D) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}\right)\left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

Q3.

n = number of batches (epochs?)

L = number of timesteps within batch(epoch?)

Importance Sampling

$$IS(D) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}\right)\left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

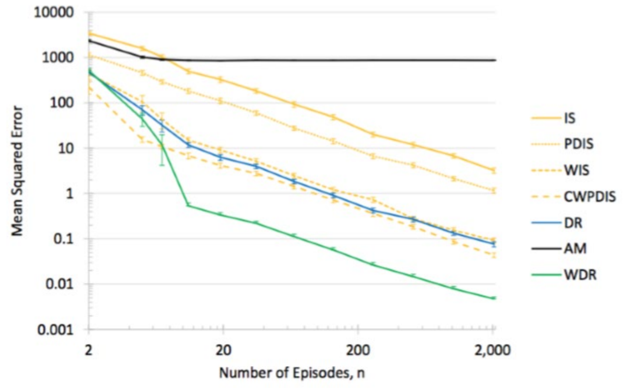
$$PSID(D) = \sum_{t=1}^{L}\gamma^t \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{\tau=1}^{t}\frac{\pi_e(a_\tau \mid s_\tau)}{\pi_b(a_\tau \mid s_\tau)}\right) R_t^i$$

$$WIS(D) = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

$$DR(\pi_e \mid D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i (R_t^i - \hat{q}^{\pi_e}(S_t^i, A_t^i)) + \gamma^t \rho_{t-1}^i \hat{v}^{\pi_e}(S_t^i),$$

WDR

MAGIC



Importance Sampling

$$IS(D) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{t=1}^{L}\frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}\right)\left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

$$PSID(D) = \sum_{t=1}^{L}\gamma^t \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{\tau=1}^{t}\frac{\pi_e(a_\tau \mid s_\tau)}{\pi_b(a_\tau \mid s_\tau)}\right) R_t^i$$

temporal

$$WIS(D) = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \left(\sum_{t=1}^{L}\gamma^t R_t^i\right)$$

weighted

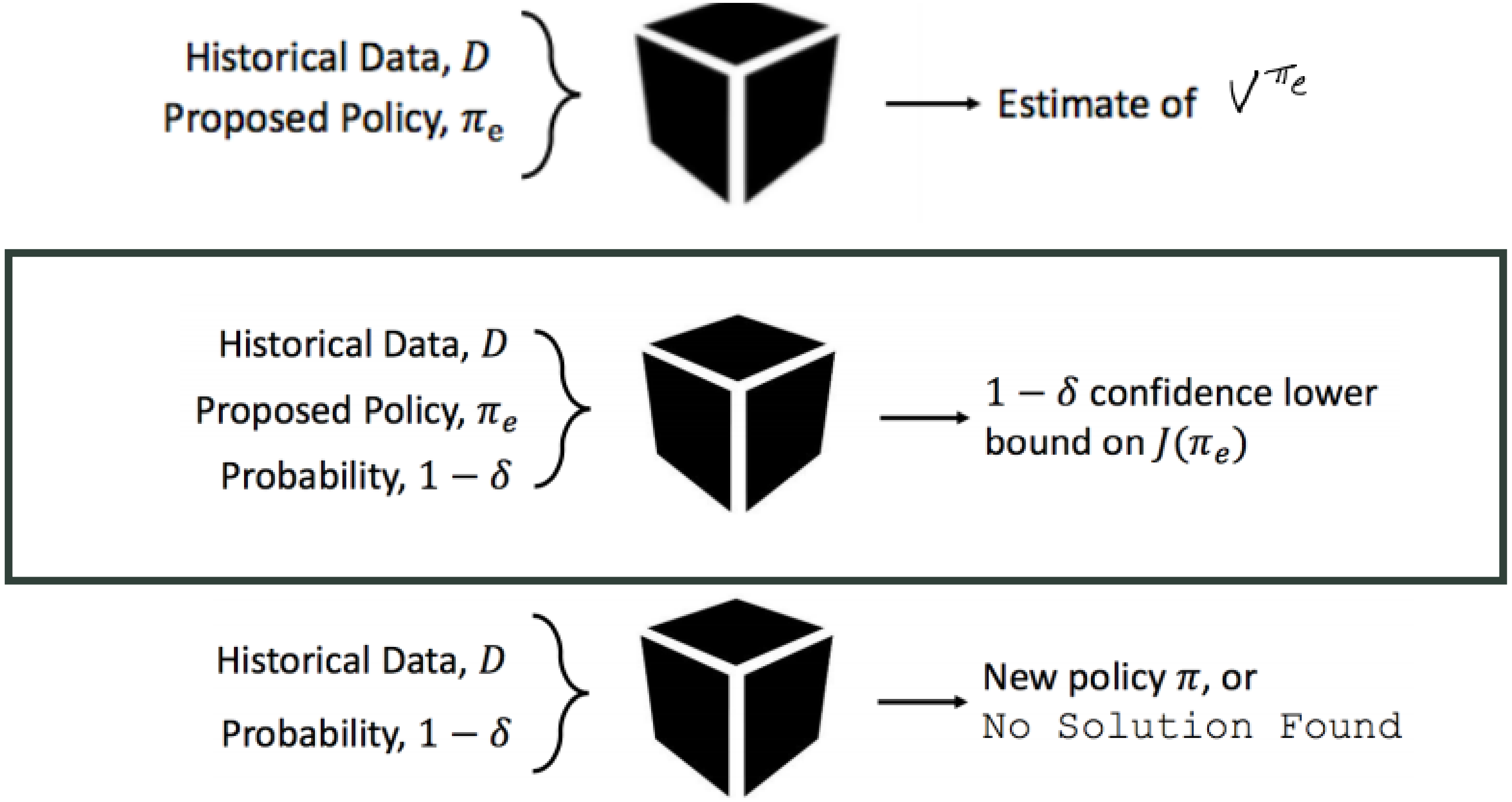
$$DR(\pi_e \mid D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i (R_t^i - \hat{q}^{\pi_e}(S_t^i, A_t^i)) + \gamma^t \rho_{t-1}^i \hat{v}^{\pi_e}(S_t^i),$$

approximated model + IS

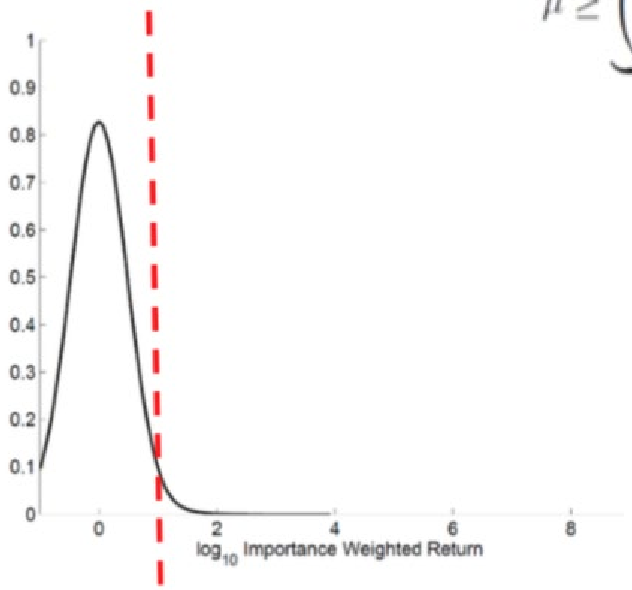
WDR weighted DR

MAGIC

High-confidence off-policy policy evaluation (HCOPE)



High-confidence off-policy policy evaluation (HCOPE)



$$\mu \geq \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i}}_{\text{empirical mean}} - \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \frac{7n\ln(2/\delta)}{3(n-1)}}_{\text{term that goes to zero as } 1/n \text{ as } n \to \infty} - \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^{n} \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2}}_{\text{term that goes to zero as } 1/\sqrt{n} \text{ as } n \to \infty}.$$

1. Use some of the data to cutoff / tune the confidence interval
2. Compute lower bound (value function)

lecture 15

Frozen Lake

Used tensorflow and gym

Used e-greedy or random noise for just frozen lake (not slippery) deterministic

Apply
learning rate
for slow
learning

Used e-greedy or random noise for slippery and windy frozen lake stochastic

Used q-network for slippery and windy frozen lake