Exploration-Exploitation Tradeoff

Part II

UCB Exploration via Q-Ensembles (Chen et al.)

Unifying Count-based Exploration and Intrinsic Motivation (G. Bellemare et al.)

April 4th, 2019

Presenter: Yilun Wu

Deep Reinforcement Learning Seminar (Spring 19')

ETH Zürich

The Multi-Armed Bandit Problem

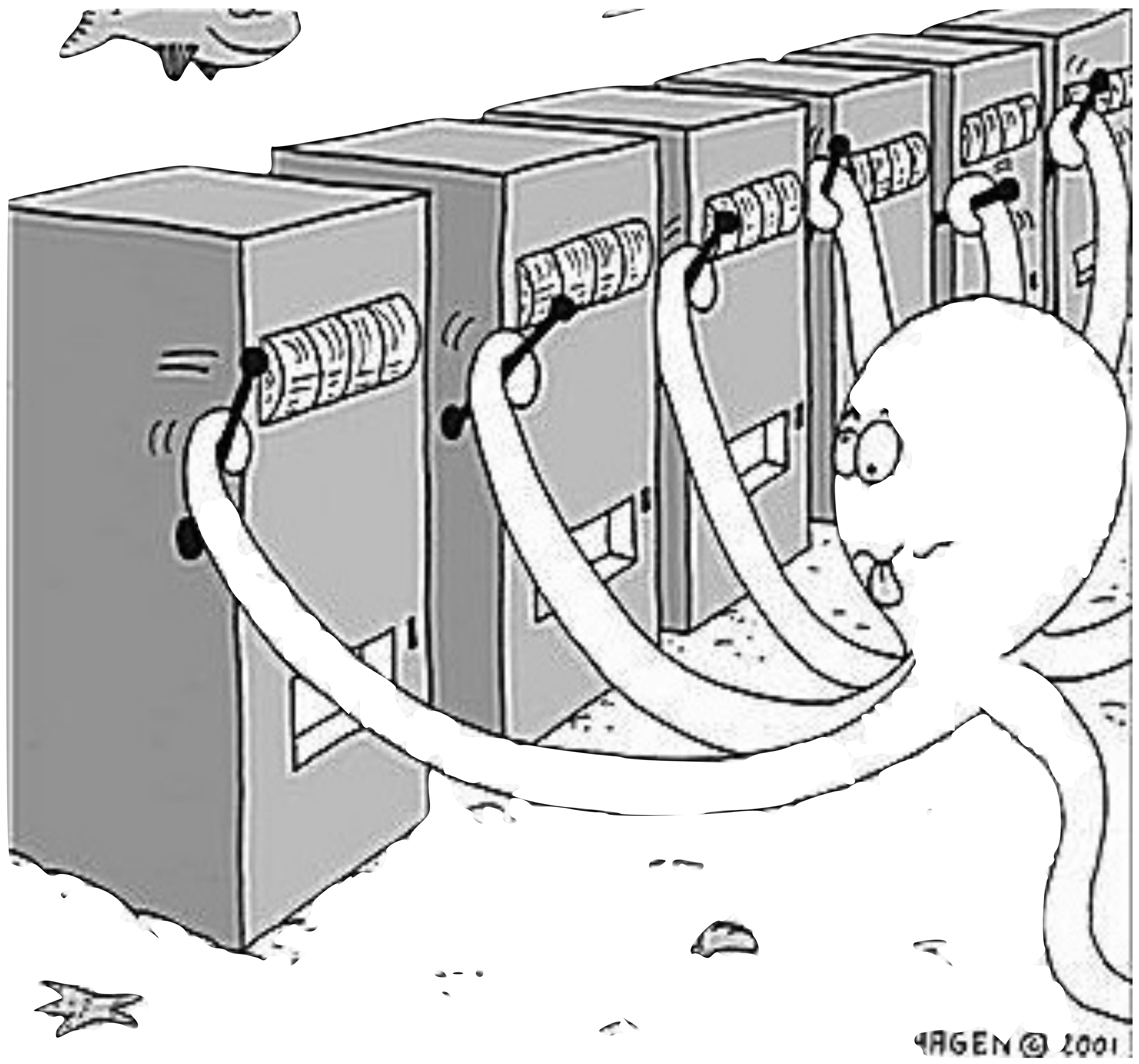
Setup: A – **Action Space,** R – **Reward**

$$\mathscr{R} = \mathbb{P}[R = r \mid A = a]$$

Task: Get maximum reward after a given set of trials

Or minimize regret:

$$L_t = \mathbb{E}\Big[\sum_{\tau=1}^{t} v^* - q(A_\tau)\Big]$$ **, where** $q(a) = \mathbb{E}[R \mid A = a]$



The Multi-Armed Bandit Problem

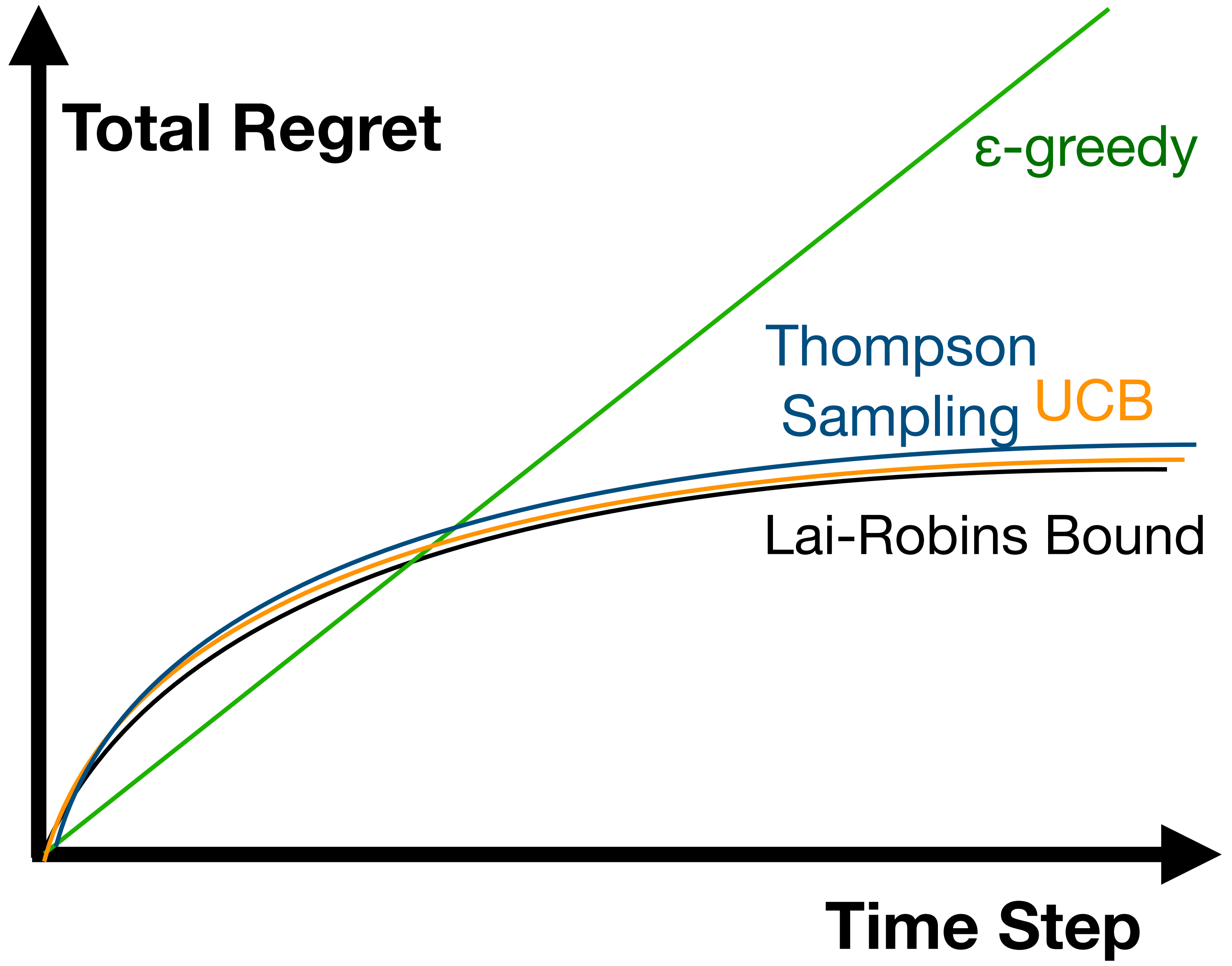
$$\mathcal{R} = \mathbb{P}[R = r \mid A = a]$$

$$L_t = \mathbb{E}\left[\sum_{\tau=1}^{t} v^* - q(A_\tau)\right]$$, **where** $q(a) = \mathbb{E}[R \mid A = a]$

Fundamental Lower Bound (Lai and Robbins [1985]):

$$\lim_{t\to\infty} L_t \geq \log t \sum_a \frac{v^* - q(a)}{KL(R^a, R^{a^*})}$$

- Exploration Strategy
  - Random Exploration (e.g. epsilon-greedy)
  - Optimism in the face of uncertainty (e.g. UCB)
  - Posterior Sampling (e.g. Thompson Sampling)



UCB (Upper Confidence Bound)

Hoeffding's Inequality:

Let X_1, X_2, …, X_t be i.i.d. r.v. in [0,1], and $\bar{X}_t = \frac{1}{t}\sum_{\tau=1}^{t} X_\tau$ be the empirical mean, then

$$\mathbb{P}\left[E[X] > \bar{X}_t + u\right] \leq \exp(-2tu^2)$$

Apply it to the bandit setting:

$$\mathbb{P}\left[q(a) > \hat{q}(a) + U_t(a)\right] \leq exp(-2N_t(a)U_t(a)^2) = P$$

$$U_t(a) = \sqrt{\frac{-\log P}{2N_t(a)}}$$

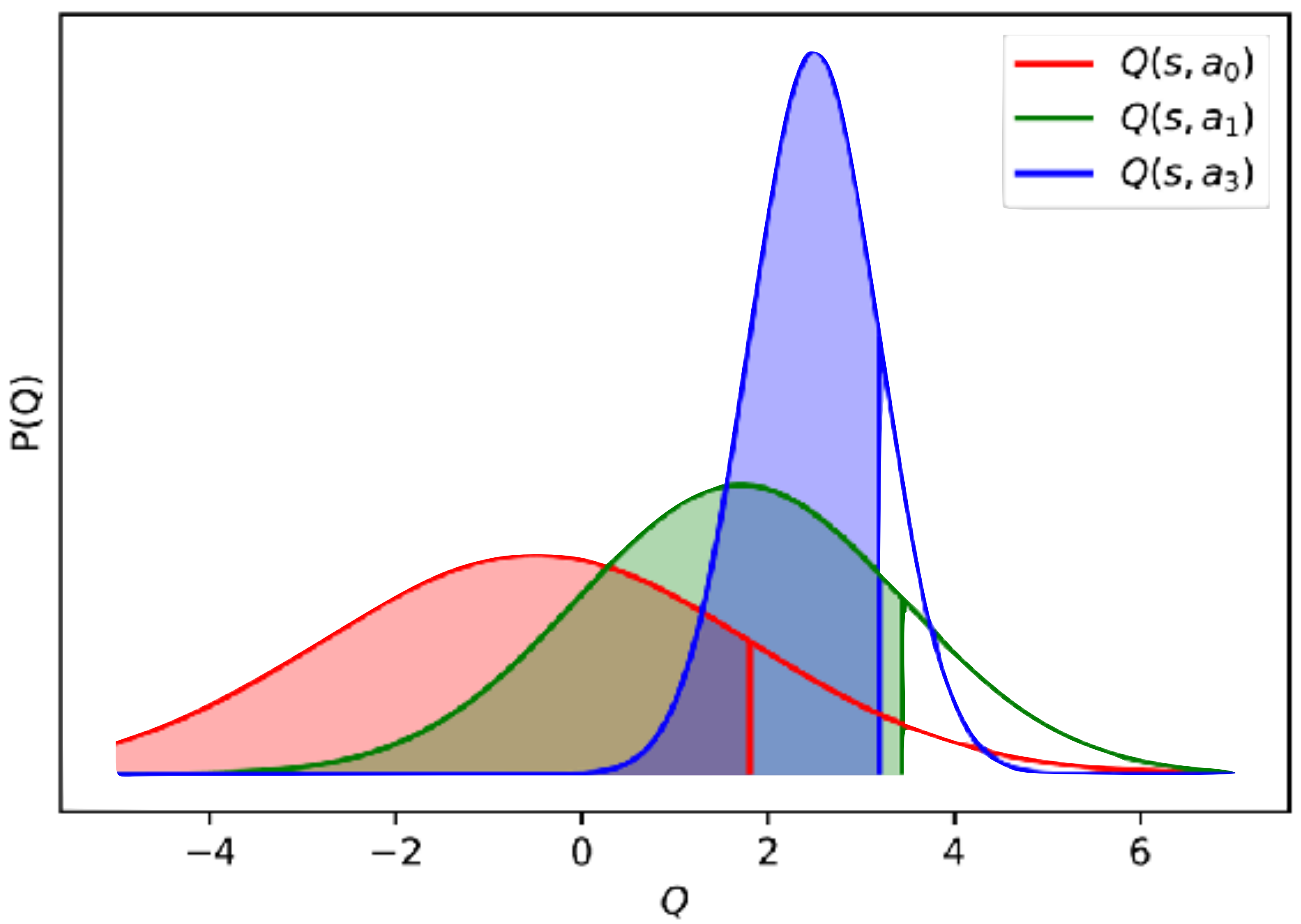
UCB for Multi-Armed Bandit

Apply it to the bandit setting:

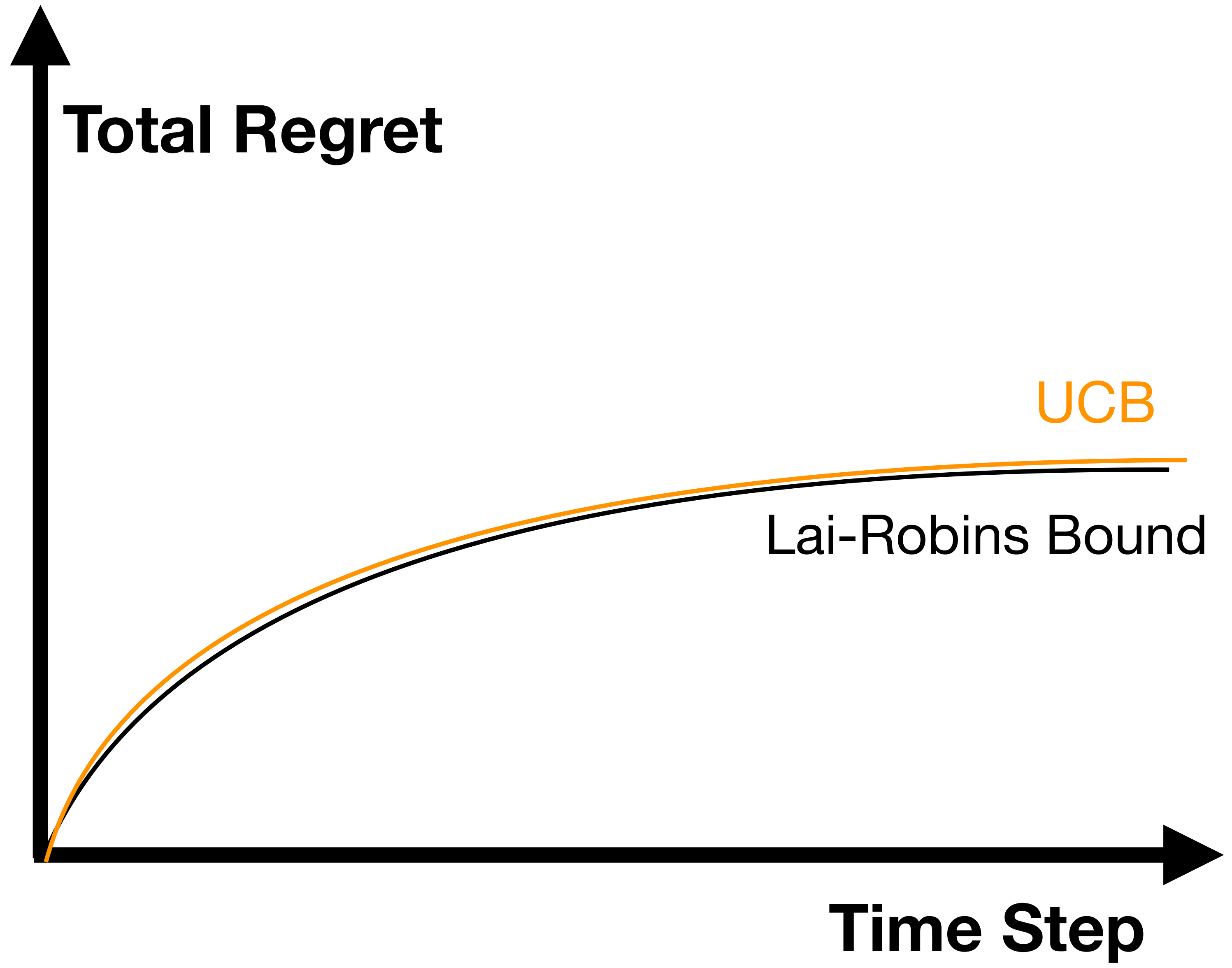
$$\mathbb{P}\left[q(a) > \hat{q}(a) + U_t(a)\right] \leq exp(-2N_t(a)U_t(a)^2) = P$$

$$U_t(a) = \sqrt{\frac{-\log P}{2N_t(a)}}$$

UCB Algorithm for Optimal Regret Bound (Bandit):



$$A_t = \underset{a \in \mathcal{A}}{\arg\max}\, \hat{q}_t(a) + \sqrt{\frac{-\log P}{2N_t(a)}}$$



UCB for Multi-Armed Bandit (1-step MDP)

UCB Algorithm for Optimal Regret Bound (Bandit):

$$A_t = \arg\max_{a \in \mathcal{A}} \hat{q}_t(a) + \sqrt{\frac{-\log P}{2N_t(a)}}$$

UCB for MDP

UCB Algorithm for Optimal Regret Bound (Bandit):

$$A_t = \arg\max_{a \in \mathcal{A}} \hat{q}_t(a) + \sqrt{\frac{-\log P}{2\boxed{N_t(a)}}}$$

MBIE-EB (Model-based Interval Estimation with **Exploration Bonuses**)

(Strehl and Littman, 2008)

$$V(x) = \max_{a \in \mathcal{A}} \left[\hat{R}(x,a) + \gamma \mathbb{E}_{\hat{P}}[V(x')] + \frac{\beta}{\sqrt{\boxed{N(x,a)}}} \right]$$

UCB for Large MDP

MBIE-EB (Model-based Interval Estimation with **Exploration Bonuses**)

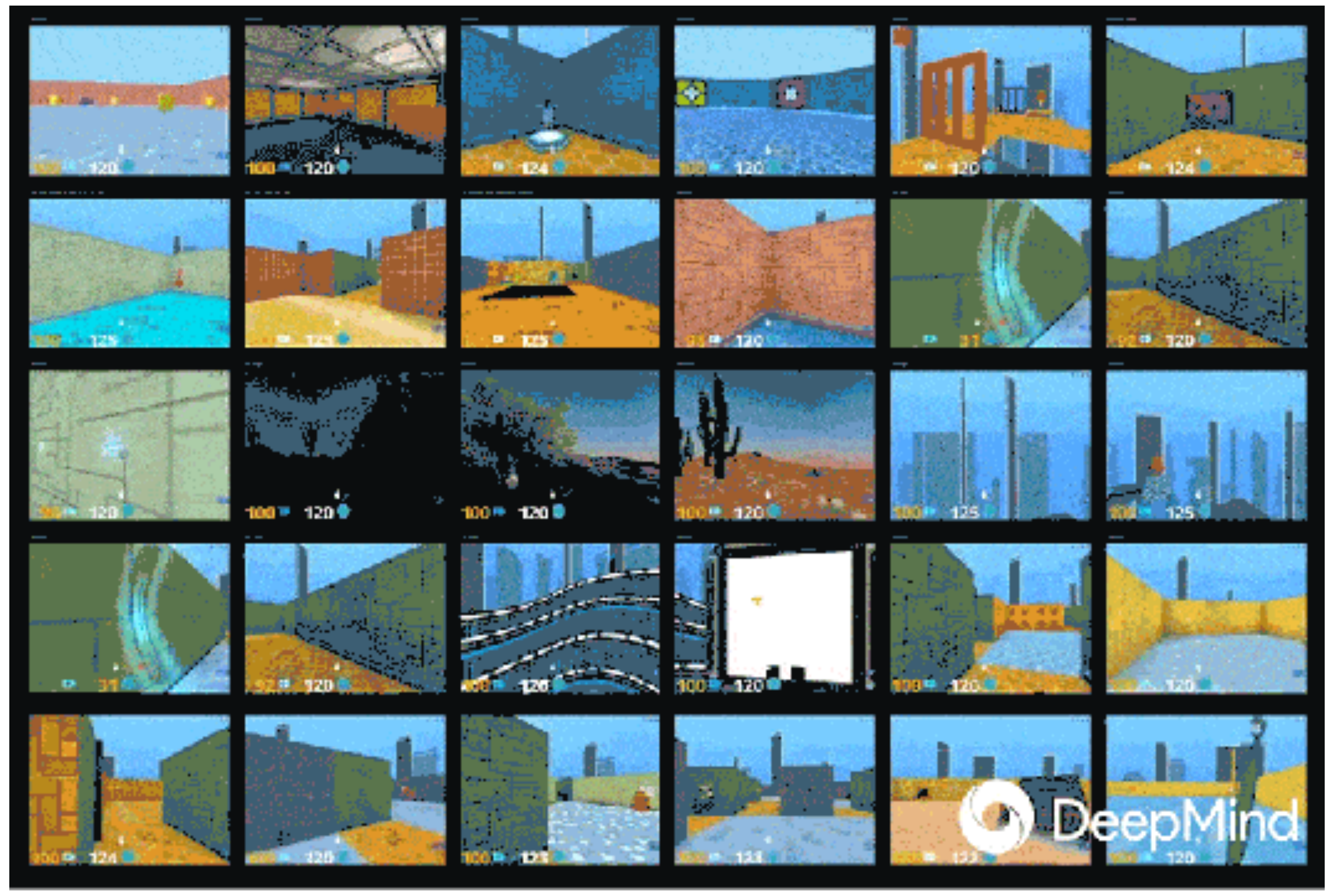
(Strehl and Littman, 2008)

$$V(x) = \max_{a \in \mathscr{A}} \left[\hat{R}(x, a) + \gamma \mathbb{E}_{\hat{P}}[V(x')] + \frac{\beta}{\sqrt{N(x, a)}} \right]$$

For MDPs with huge state space, count will be zero for most states.

Thus, we need a generalized state visit count - **pseudo-counts**.

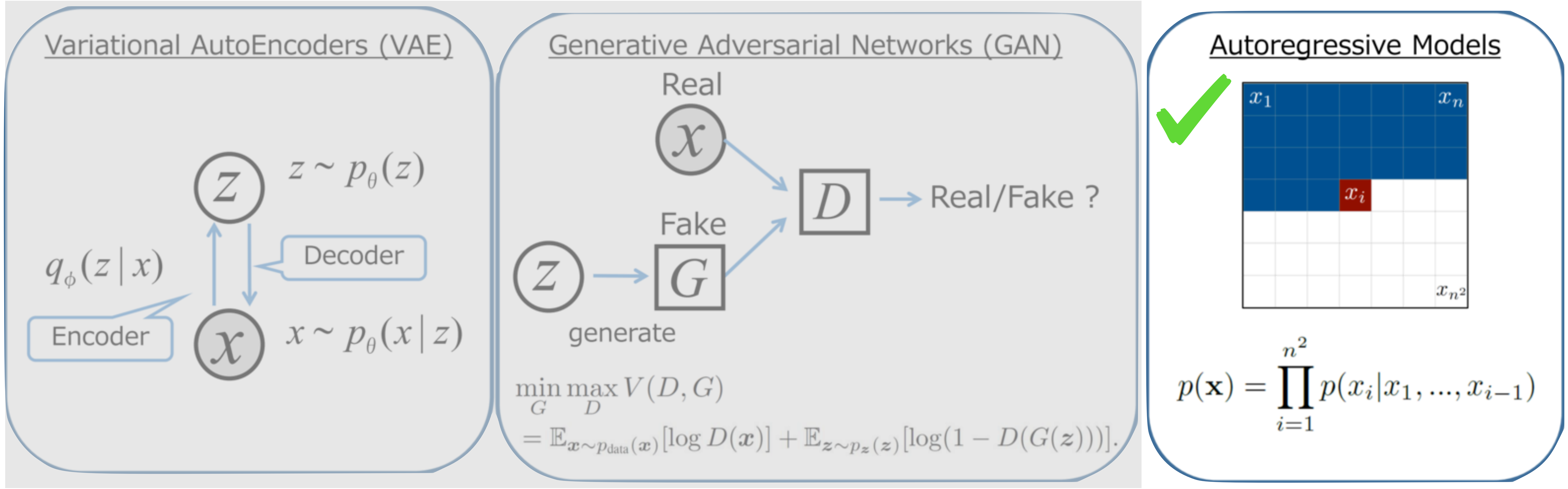
(Bellemare et al., 2016)



Pseudo-Counts derived from density models

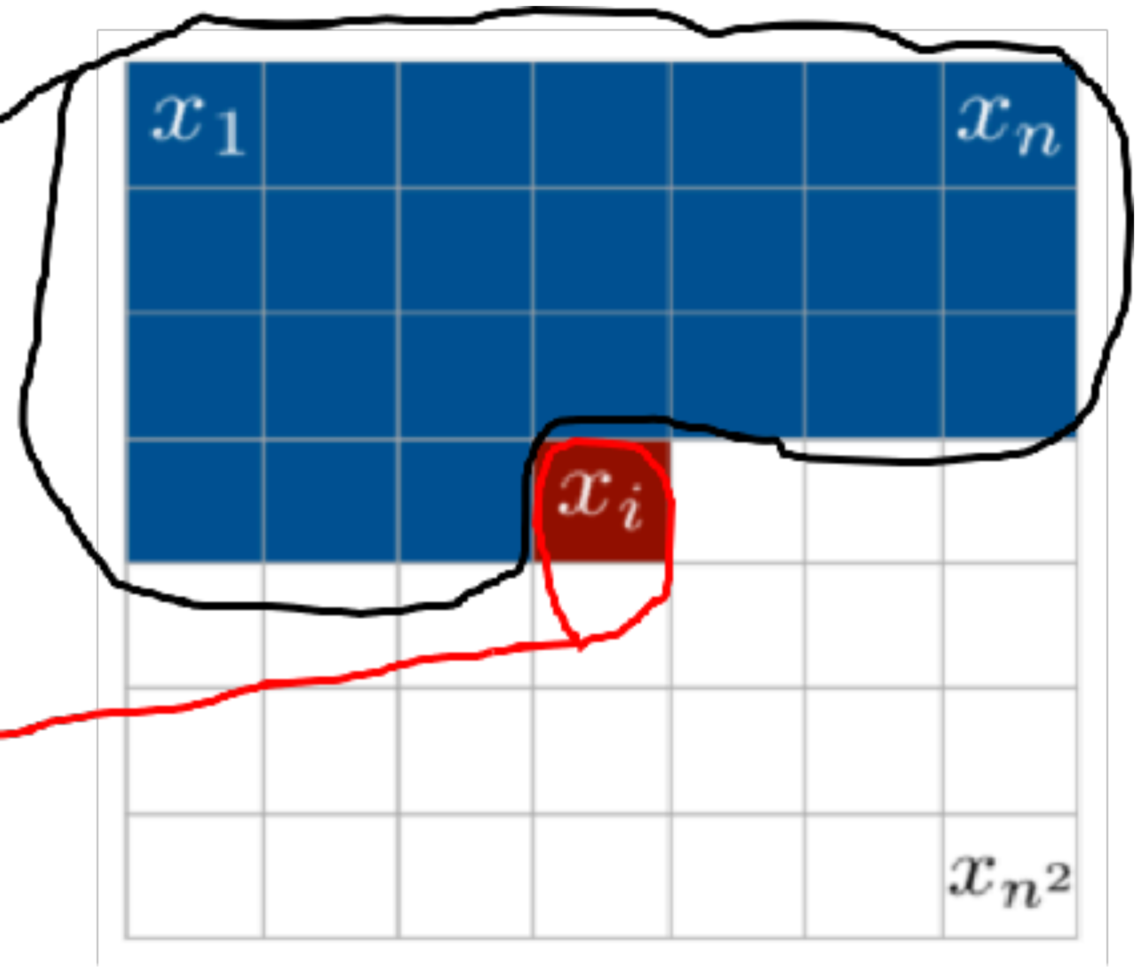
Density Model: A model which gives the **distribution** of states which assumes states are **independently** distributed.

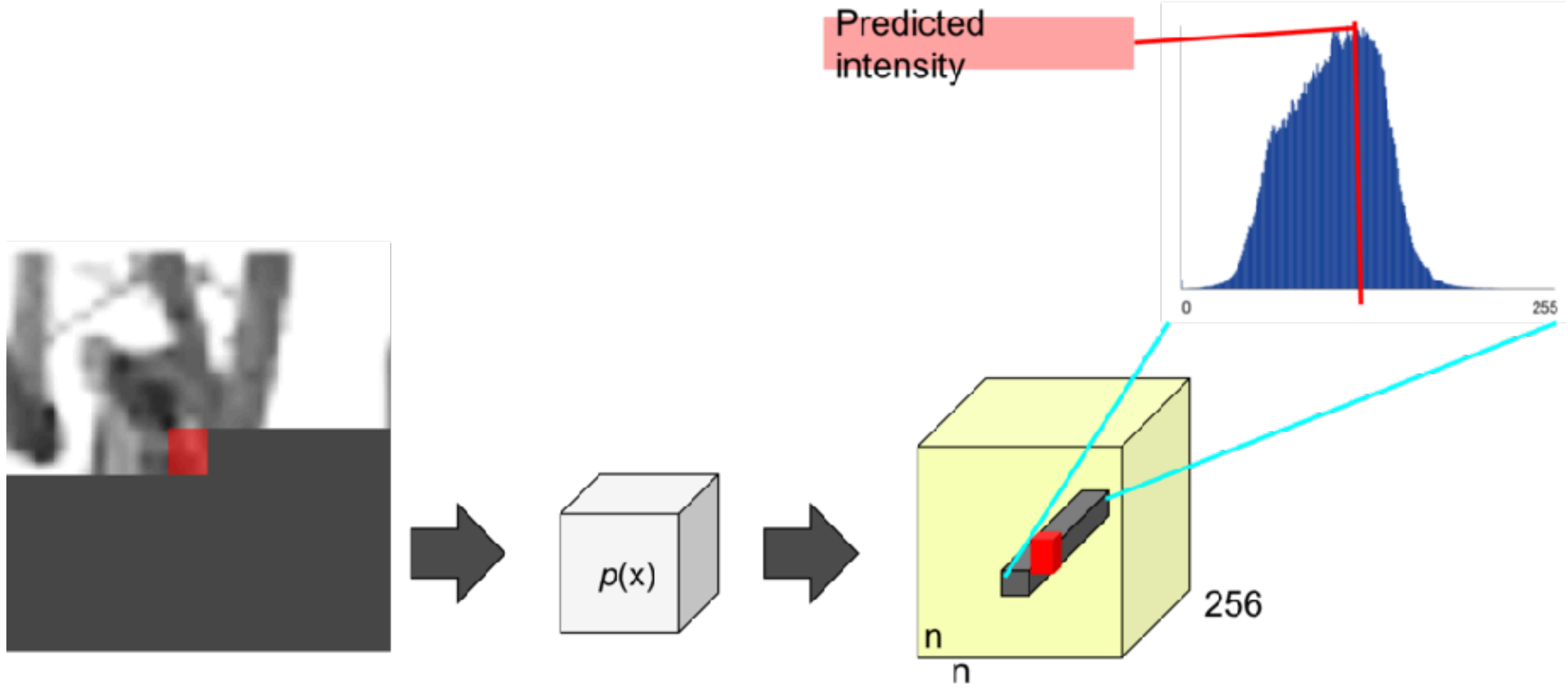
Density model is a kind of generative model which explicitly gives the likelihood/similarity of data (distribution of data) given the training dataset.



Density model Example: PixelCNN

$$p(x_i|x_1, \ldots, x_{i-1})$$





From Density Model to Pseudo-Counts

Density Model: $\rho_n(x) := \rho(x \mid x_{1:n})$

- probability of state x given all the experience so far.

Recoding Probability: $\rho'_n(x) := \rho(x \mid x_{1:n}, x)$

- probability of state x given all the experience so far and **hypothetically observe state x at the next step**.

Define Pseudo-count $\hat{N}_n(x)$ and Pseudo-count Total $\hat{n}$

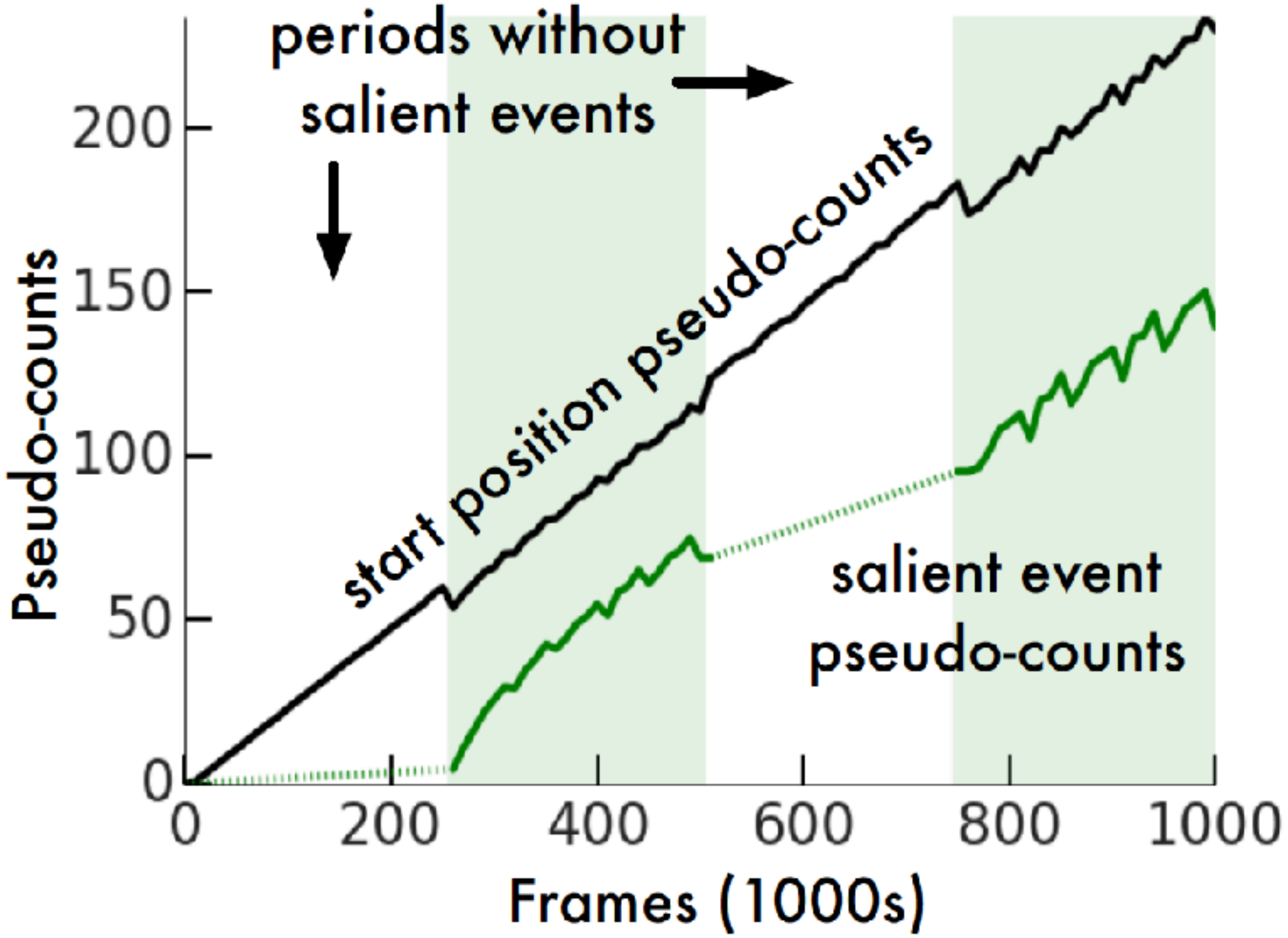
such that $\rho_n(x) = \dfrac{\hat{N}_n(x)}{\hat{n}}, \rho'_n(x) = \dfrac{\hat{N}_n(x) + 1}{\hat{n} + 1}$

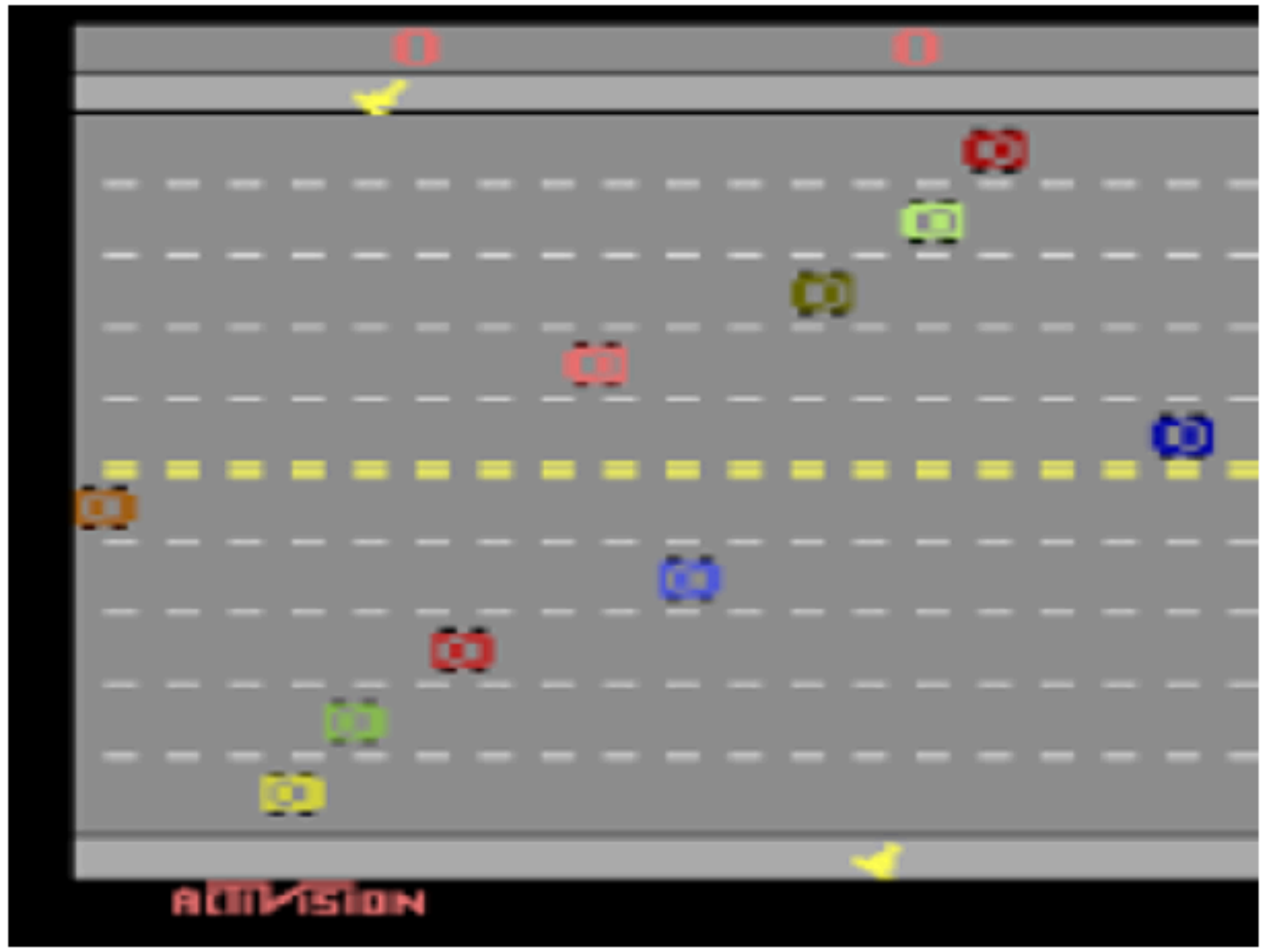
Can be solved via $\hat{N}_n(x) = \dfrac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)} = \hat{n}\rho_n(x)$

Does it work?

Freeway Game

Policy: 250,000 frames of wait + 250,000 frames of UP action





Intrinsic Motivation Exploration

- Forget about rewards (external motivation)
- The goal of learning is to **gather information**
- Information is gathered if the uncertainty of **a quantity of interest** (reward, transition probability, optimal policy, etc.) decreases
- This decrease of uncertainty can also be viewed as **large difference between prior and posterior distribution** *(surprise)*

Intrinsic Motivation Exploration

In the context of **modelling state density**, consider a weighted density model from a class of density models $\mathscr{M}$

$$\xi_n(x) := \int_{\rho \in \mathscr{M}} w_n(\rho)\rho(x \mid x_{1:n})d\rho$$

Update the weight through bayesian filtering: $$w_n(\rho, x) := \frac{w_n(\rho)\rho(x \mid x_{1:n})}{\xi(x)}$$

Measure the Information Gain through **distance between prior and posterior** (KL-Divergence): $$\mathbf{IG}_n(x) := \mathbf{KL}\Big(w_n(\rho, x) \,||\, w_n(\rho)\Big)$$

Connection between Counts and Intrinsic Motivation

Measure the Information Gain through **distance between prior and posterior** (KL-Divergence):

$$\textbf{IG}_n(x) := \textbf{KL}\big(w_n(\rho, x)\,||\,w_n(\rho)\big)$$

Use PG (Prediction Gain) as an approximate to IG:

$$\textbf{PG}_n(x) := \log \rho'_n(x) - \log \rho_n(x)$$

Recall that:

$$\hat{N}_n(x) = \frac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)} = \hat{n}\rho_n(x)$$

PG is related to pseudo-count in that:
With equality when $\rho'_n(x) \to 0$

$$\hat{N}_n(x) \approx \left(e^{\mathrm{PG}_n(x)} - 1\right)^{-1}$$

Connection between Counts and Intrinsic Motivation

Measure the Information Gain through **distance between prior and posterior** (KL-Divergence):

$$\textbf{IG}_n(x) := \textbf{KL}\big(w_n(\rho, x)\,||\,w_n(\rho)\big)$$

Use PG (Prediction Gain) as an approximate to IG:

$$\textbf{PG}_n(x) := \log \rho'_n(x) - \log \rho_n(x)$$

Recall that:

$$\hat{N}_n(x) = \frac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)} = \hat{n}\rho_n(x)$$

PG is related to pseudo-count in that:
With equality when $\rho'_n(x) \to 0$

$$\hat{N}_n(x) \approx \left(e^{\mathrm{PG}_n(x)} - 1\right)^{-1}$$

Connection between Counts and Intrinsic Motivation

$$\textbf{IG}_n(x) := \textbf{KL}\big(w_n(\rho, x) \,||\, w_n(\rho)\big)$$

$$\textbf{PG}_n(x) := \log \rho'_n(x) - \log \rho_n$$

$$\hat{N}_n(x) = \frac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)} = \hat{n}\rho_n(x)$$

$$\hat{N}_n(x) \approx \left(e^{\mathrm{PG}_n(x)} - 1\right)^{-1}$$

Furthermore,

$$\mathrm{IG}_n(x) \leq \mathrm{PG}_n(x) \leq \hat{N}_n(x)^{-1} \leq \hat{N}_n(x)^{-1/2}$$

Connection between Counts and Intrinsic Motivation

Furthermore,

$$\mathrm{IG}_n(x) \leq \mathrm{PG}_n(x) \leq \hat{N}_n(x)^{-1} \leq \hat{N}_n(x)^{-1/2}$$

MBIE-EB (Model-based Interval Estimation with **Exploration Bonuses**)

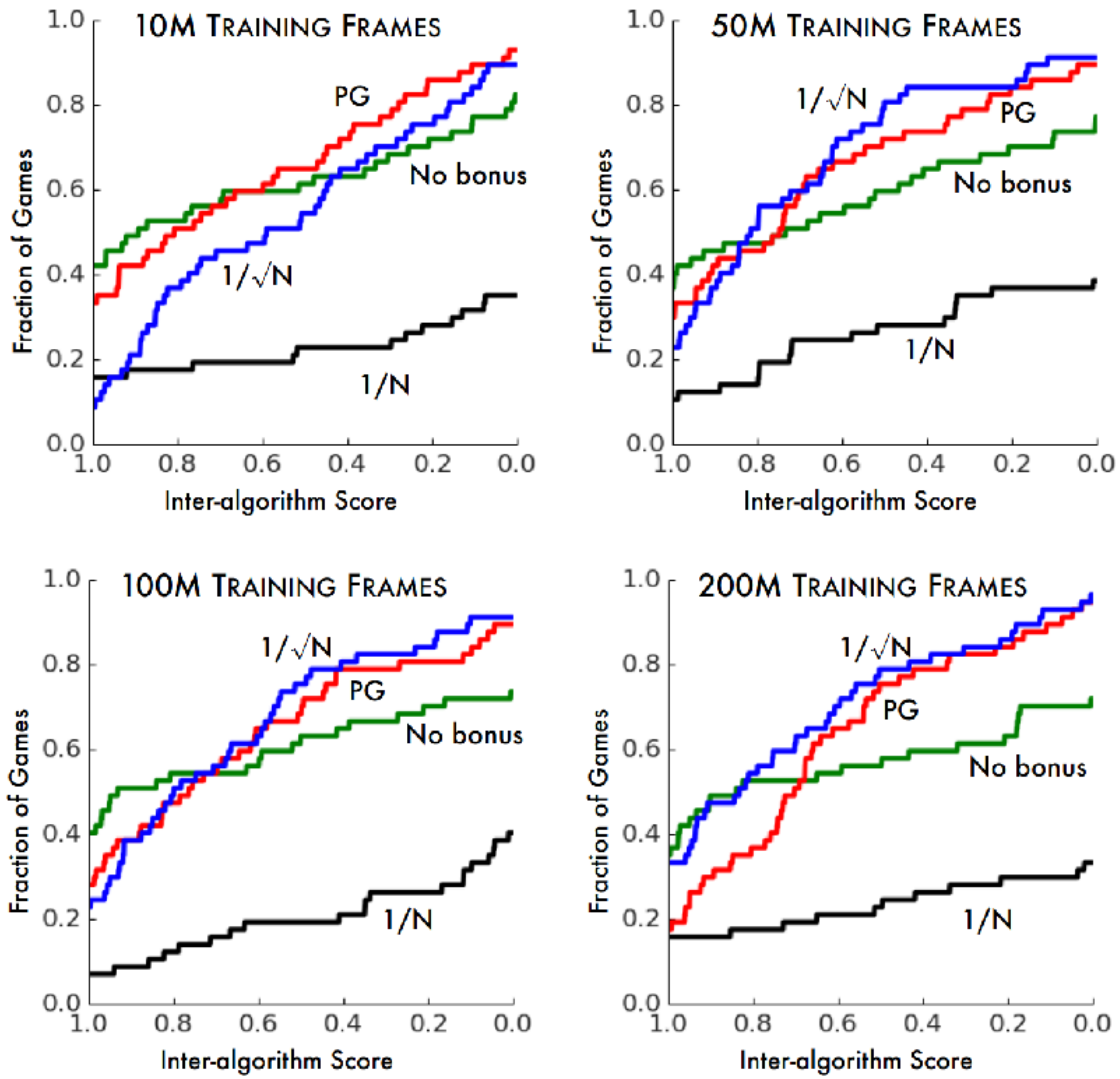
(Strehl and Littman, 2008)

$$V(x) = \max_{a \in \mathscr{A}} \left[\hat{R}(x, a) + \gamma \mathbb{E}_{\hat{P}}[V(x')] + \frac{\beta}{\sqrt{N(x, a)}} \right]$$

$$V(x) = \max_{a \in \mathscr{A}} \left[\hat{R}(x, a) + \gamma \mathbb{E}_{\hat{P}}[V(x')] + \beta IG_n(x) \right]$$

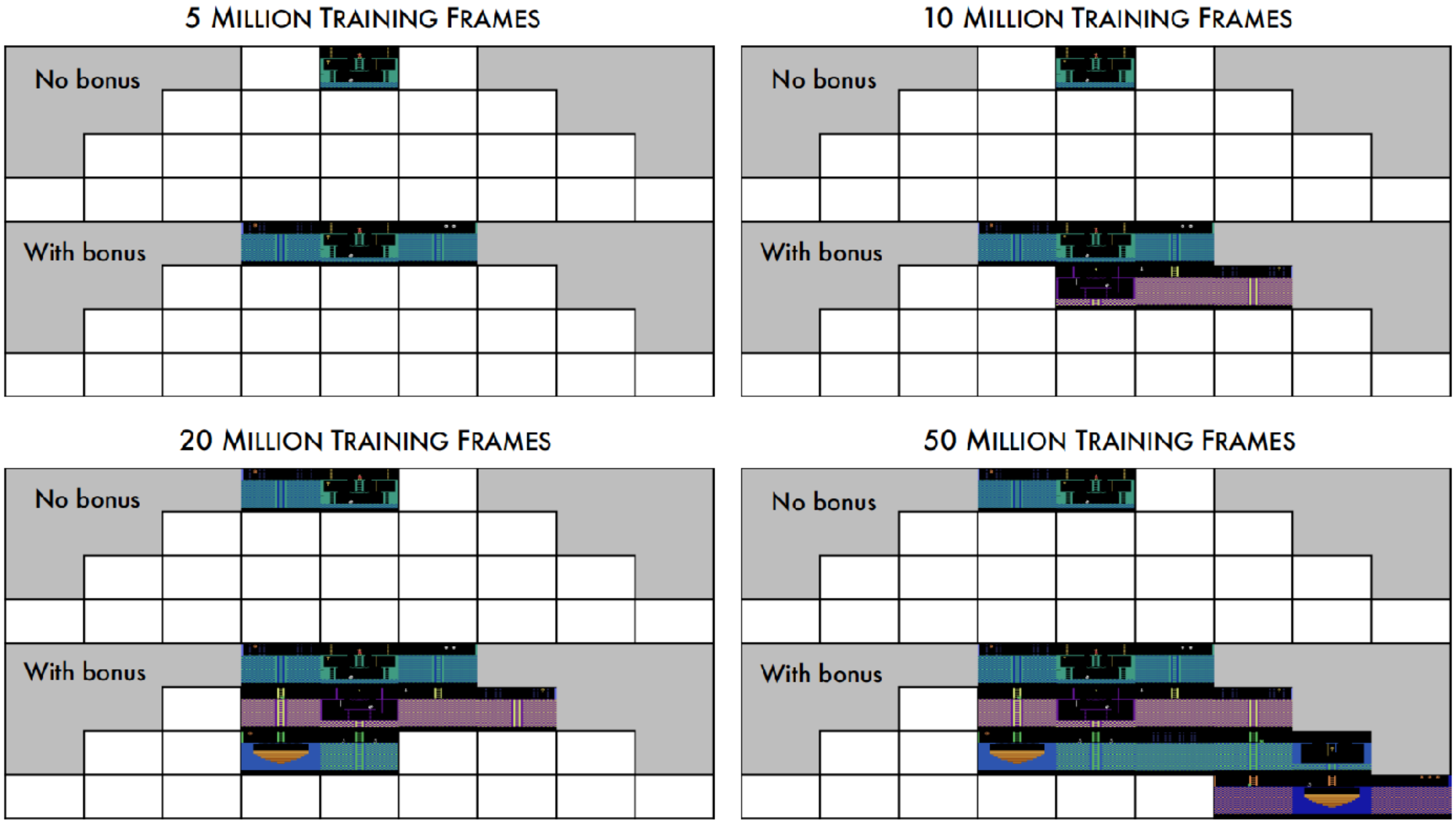
Results with different exploration bonus

$$\mathrm{IG}_n(x) \leq \mathrm{PG}_n(x) \leq \hat{N}_n(x)^{-1} \leq \hat{N}_n(x)^{-1/2}$$



Tested on 60 Atari games
A3C with bonus
Algorithm achieve a normalized X score on Y fraction of the games

Exploration in Montezuma's Revenge



The Multi-Armed Bandit Problem

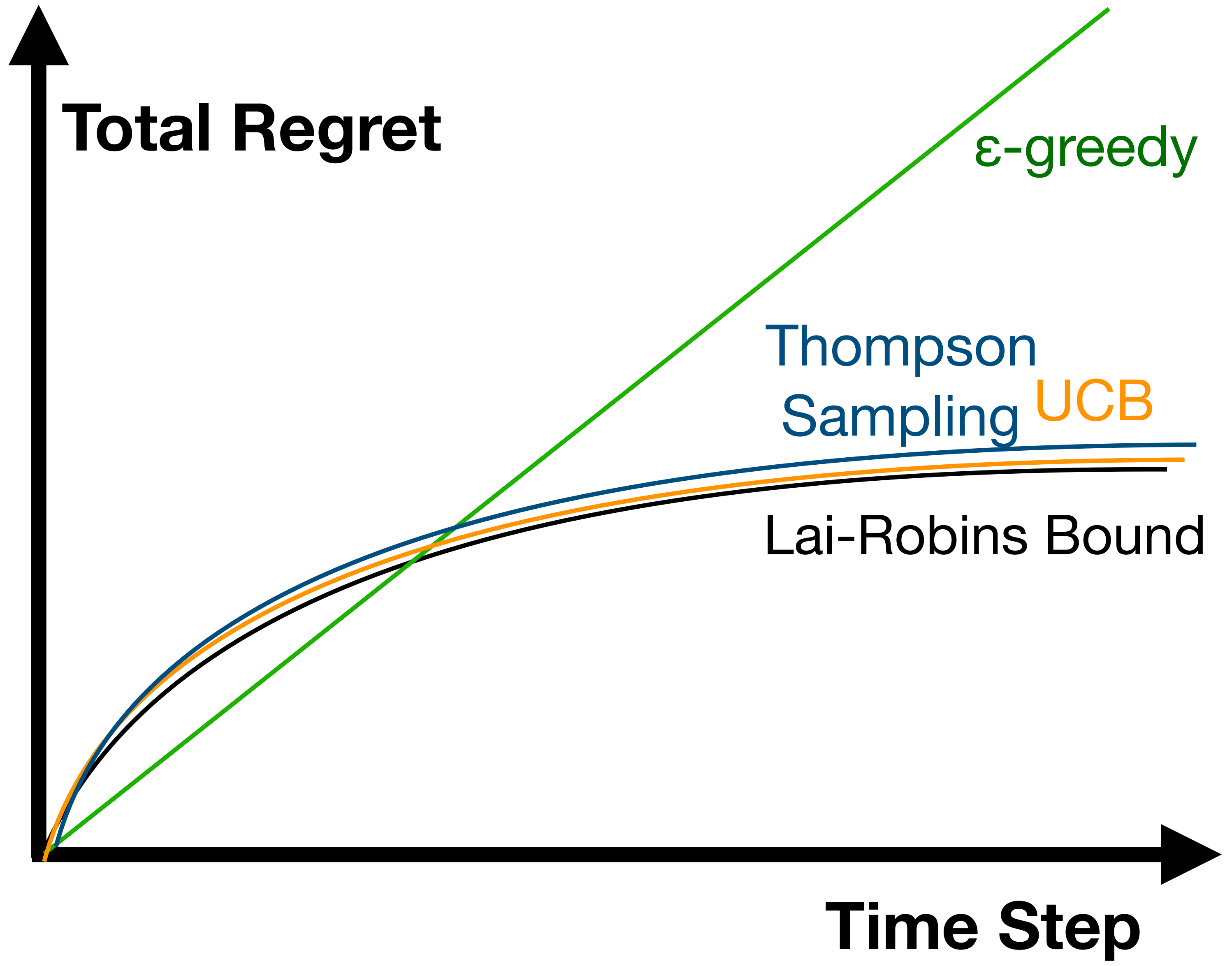
$$\mathcal{R} = \mathbb{P}[R = r \mid A = a]$$

$$L_t = \mathbb{E}\left[\sum_{\tau=1}^{t} v^* - q(A_\tau)\right]$$ **, where** $q(a) = \mathbb{E}[R \mid A = a]$

Fundamental Lower Bound (Lai and Robbins [1985]):

$$\lim_{t \to \infty} L_t \geq \log t \sum_a \frac{v^* - q(a)}{KL(R^a, R^{a^*})}$$

- Exploration Strategy
  - Random Exploration (e.g. epsilon-greedy)
  - Optimism in the face of uncertainty (e.g. UCB)
  - Posterior Sampling (e.g. Thompson Sampling)



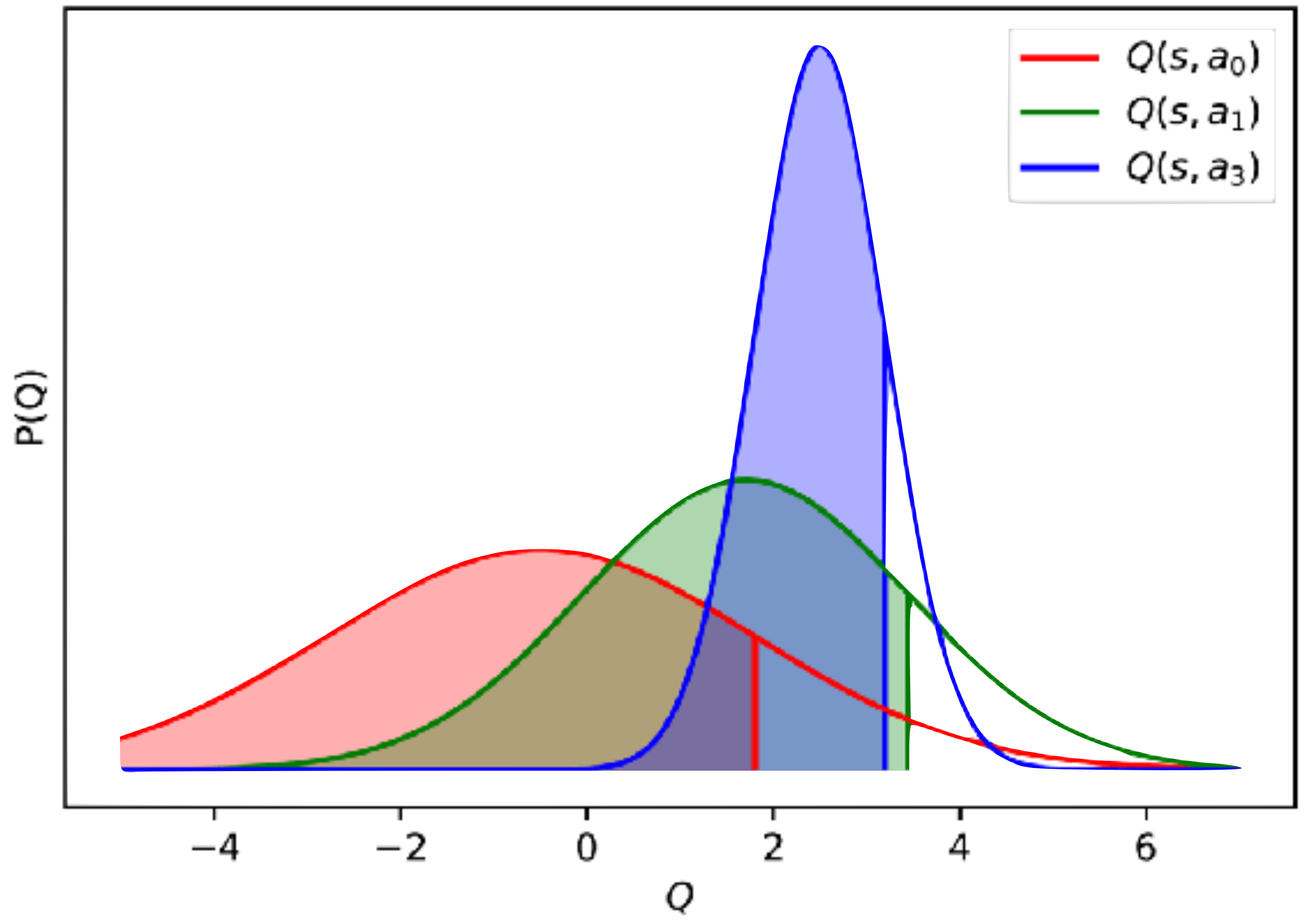
Thompson Sampling

Maintain belief over rewards: $q_1, q_2, \ldots, q_n \sim \hat{p}(q_1, q_2, \ldots, q_n)$

Sampling and act greedily: $A_t = \underset{a \in \mathcal{A}}{\arg\max}\, q(a)$

What is the **equivalent** of rewards from bandit in MDP?

Q-values!



How to maintain a distribution of Q-values?

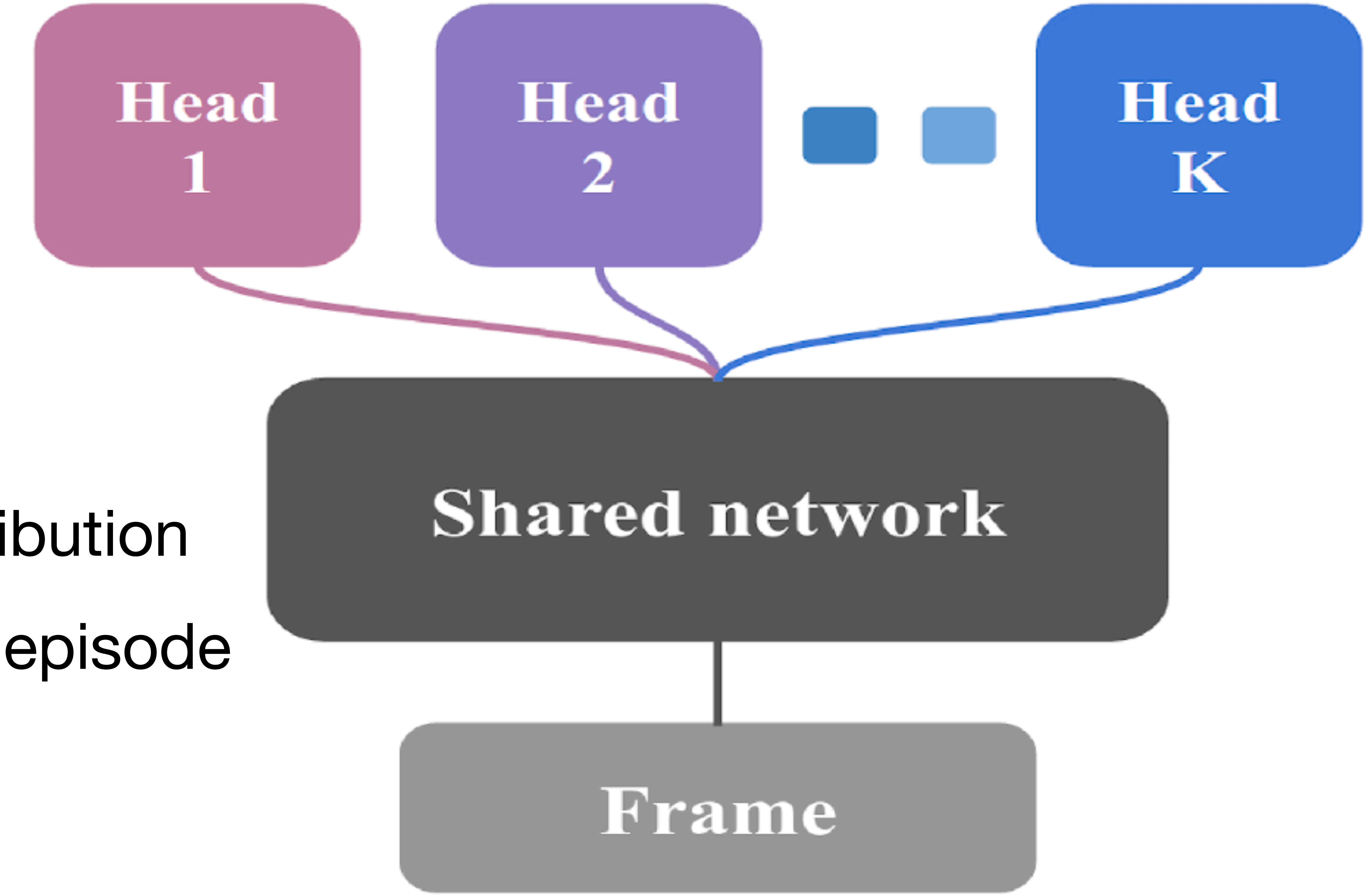
- **PSRL** (Posterior Sampling for Reinforcement Learning, Osband et al.)
    - **Sample MDP** from belief distribution
    - Solve for **optimal policy** of the sampled MDP
    - Use observed transition and reward to update MDP belief

- **Q-ensembles** neural network (Bootstrapped DQN, Osband et al.)
    - **Sample Q function** from belief distribution
    - Act **greedily** according to Q for one episode
    - Update belief of Q

Bootstrapped DQN

Training many independent NNs is costly

Solution: Share most layers



- Q-ensembles neural network
  - **Sample Q function** from belief distribution
  - Act **greedily** according to Q for one episode
  - Update belief of Q

UCB Exploration using Q-Ensembles

Add UCB into Q-ensembles:

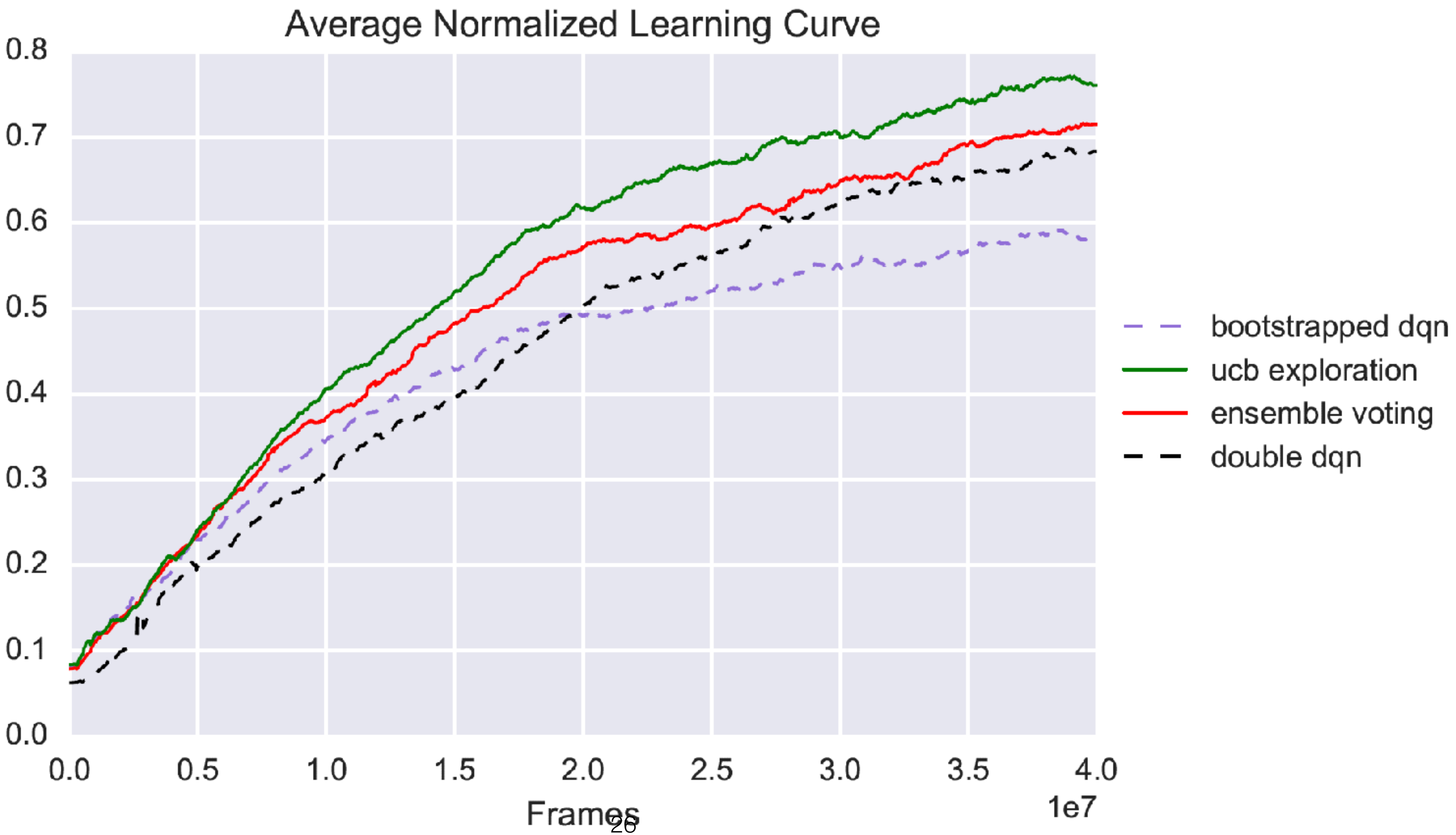
Instead of sampling Q and act greedily on the sampled Q,

Select action according to **UCB of empirical mean reward**

Algorithm 2 UCB Exploration with Q-Ensembles

Input: Value function networks Q with K outputs $\{Q_k\}_{k=1}^{K}$. Hyperparameter λ.
Let B be a replay buffer storing experience for training.
for each episode **do**
    Obtain initial state from environment s_0
    for step $t = 1, \dots$ until end of episode **do**
        Pick an action according to $a_t \in \operatorname{argmax}_a \left\{ \tilde{\mu}(s_t, a) + \lambda \cdot \tilde{\sigma}(s_t, a) \right\}$
        Receive state s_{t+1} and reward r_t from environment, having taken action a_t
        Add (s_t, a_t, r_t, s_{t+1}) to replay buffer B
        At learning interval, sample random minibatch and update $\{Q_k\}$ according to (12)
    end for
end for

Q-ensemble Results



Benchmark on Atari Games

Maximal Mean Reward in 100 consecutive episodes
Evaluated on 48 Games

UCB-Exploration achieved the highest score in **30/48** games

	Bootstrapped DQN	Double DQN	Ensemble Voting	UCB-Exploration
Alien	1445.1	2059.7	2282.8	**2817.6**
Amidar	430.58	667.5	**683.72**	663.8
Assault	2519.06	2820.61	3213.58	**3702.76**
Asterix	3829.0	7639.5	**8740.0**	8732.0
Asteroids	1009.5	1002.3	**1149.3**	1007.8
Atlantis	1314058.0	1982677.0	1786305.0	**2016145.0**
Bank Heist	795.1	789.9	869.4	**906.9**
Battle Zone	26230.0	24880.0	**27430.0**	26770.0
Beam Rider	8006.58	7743.74	7991.9	**9188.26**
Bowling	28.62	30.92	32.92	**38.06**
Boxing	85.91	94.07	94.47	**98.08**
Breakout	400.22	**467.45**	426.78	411.31
Centipede	5328.77	5177.51	6153.28	**6237.18**
Chopper Command	2153.0	3260.0	3544.0	**3677.0**
Crazy Climber	110926.0	124456.0	126677.0	**127754.0**
Demon Attack	9811.45	23562.55	30004.4	**59861.9**
Double Dunk	-10.82	-14.58	-11.94	**-4.08**
Enduro	1314.31	1439.59	1999.88	**2752.55**
Fishing Derby	21.89	23.69	**30.02**	29.71
Freeway	33.57	32.93	33.92	**33.96**
Frostbite	1284.8	529.2	1196.0	**1903.0**
Gopher	7652.2	12030.0	10993.2	**12910.8**
Gravitar	227.5	279.5	**371.5**	318.0
Ice Hockey	-4.62	-4.63	**-1.73**	-4.71
Jamesbond	594.5	594.0	602.0	**710.0**
Kangaroo	8186.0	7787.0	8174.0	**14196.0**
Krull	8537.52	8517.91	8669.17	**9171.61**
Kung Fu Master	24153.0	**32896.0**	30988.0	31291.0
Montezuma Revenge	2.0	4.0	1.0	4.0
Ms Pacman	2508.7	2498.1	3039.7	**3425.4**
Name This Game	8212.4	**9806.9**	9255.1	9570.5
Pitfall	-5.99	-7.57	-3.37	**-1.47**
Pong	21.0	20.67	21.0	20.95
Private Eye	1815.19	788.63	**1845.28**	1252.01
Qbert	10557.25	6529.5	12036.5	**14198.25**
Riverraid	11528.0	11834.7	12785.8	**15622.2**
Road Runner	52489.0	49039.0	**54768.0**	53596.0
Robotank	21.03	29.8	31.83	**41.04**
Seaquest	9320.7	18056.4	20458.6	**24001.6**
Space Invaders	1549.9	1917.5	1890.8	**2626.55**
Star Gunner	20115.0	**52283.0**	41684.0	47367.0
Tennis	-15.11	-14.04	-11.63	**-7.8**
Time Pilot	5088.0	5548.0	6153.0	**6490.0**
Tutankham	167.47	**223.43**	208.61	200.76
Up N Down	9049.1	11815.3	19528.3	**19827.3**
Venture	**115.0**	96.0	78.0	67.0
Video Pinball	364600.85	**374686.89**	343380.29	372564.11
Wizard Of Wor	2860.0	3877.0	5451.0	**5873.0**
Zaxxon	592.0	**8903.0**	3901.0	3695.0
Times best	1	7	9	30

Comparison to A3C+[1]

Maximal Mean Reward in 100 consecutive episodes
Evaluated on 48 Games

UCB-Exploration trained with **40 million** frames
A3C+ trained with **200 million** frames

UCB-Exploration achieved the highest score in **28/48** games
A3C+ achieved the highest score in **10/48** games

Why Q-Ensembles achieve better performance?

	Ensemble Voting	UCB-Exploration	A3C+
Alien	2282.8	**2817.6**	1848.33
Amidar	683.72	663.8	**964.77**
Assault	3213.58	**3702.76**	2607.28
Asterix	**8740.0**	8732.0	7262.77
Asteroids	1149.3	1007.8	**2257.92**
Atlantis	1786305.0	**2016145.0**	1733528.71
Bank Heist	869.4	906.9	**991.96**
Battle Zone	**27430.0**	26770.0	7428.99
Beam Rider	7991.9	**9188.26**	5992.08
Bowling	32.92	38.06	**68.72**
Boxing	94.47	**98.08**	13.82
Breakout	**426.78**	411.31	323.21
Centipede	6153.28	**6237.18**	5338.24
Chopper Command	3544.0	3677.0	**5388.22**
Crazy Climber	126677.0	**127754.0**	104083.51
Demon Attack	30004.4	**59861.9**	19589.95
Double Dunk	-11.94	**-4.08**	-8.88
Enduro	1999.88	**2752.55**	749.11
Fishing Derby	**30.02**	29.71	29.46
Freeway	33.92	**33.96**	27.33
Frostbite	1196.0	**1903.0**	506.61
Gopher	10993.2	**12910.8**	5948.40
Gravitar	**371.5**	318.0	246.02
Ice Hockey	**-1.73**	-4.71	-7.05
Jamesbond	602.0	710.0	**1024.16**
Kangaroo	8174.0	**14196.0**	5475.73
Krull	8669.17	**9171.61**	7587.58
Kung Fu Master	30988.0	**31291.0**	26593.67
Montezuma Revenge	1.0	4.0	**142.50**
Ms Pacman	3039.7	**3425.4**	2380.58
Name This Game	9255.1	**9570.5**	6427.51
Pitfall	-3.37	**-1.47**	-155.97
Pong	**21.0**	20.95	17.33
Private Eye	**1845.28**	1252.01	100.0
Qbert	12036.5	14198.25	**15804.72**
Riverraid	12785.8	**15622.2**	10331.56
Road Runner	**54768.0**	53596.0	49029.74
Robotank	31.83	**41.04**	6.68
Seaquest	20458.6	**24001.6**	2274.06
Space Invaders	1890.8	**2626.55**	1466.01
Star Gunner	41684.0	47367.0	**52466.84**
Tennis	-11.63	**-7.8**	-20.49
Time Pilot	6153.0	**6490.0**	3816.38
Tutankham	**208.61**	200.76	132.67
Up N Down	19528.3	**19827.3**	8705.64
Venture	**78.0**	67.0	0.00
Video Pinball	343380.29	**372564.11**	35515.92
Wizard Of Wor	5451.0	**5873.0**	3657.65
Zaxxon	3901.0	3695.0	**7956.05**
Times Best	10	28	10

[1] A3C+ - A3C (Asynchronous Advantage Actor-Critic) with pseudo-count based reward

Further Readings

- Optimal Exploration for small MDP
  - MBIE-EB (Strehl, Littman)
- Density Model
  - Skip Context Tree Switching (Bellemare, et al.)
  - Count-Based Exploration with Neural Density Models (Ostrovski et al.)
  - EX2: Exploration with Exemplar Models for Deep Reinforcement Learning (Fu et al.)
- Q-Ensmeble methods
  - Deep Exploration via Bootstrapped DQN (Osband et al.)
  - Posterior sampling for reinforcement learning: worst-case regret bounds (Agrawal et al.)
- Information Gain based Exploration
  - VIME: Variational Information Maximizing Exploration (Houthooft et al.)

For questions: Yilun Wu <wuyil@student.ethz.ch>

End of Presentation

Questions?