Exploration Methods

Presenter: Kyungjae Lee

Contents

Contents

- Decision Making Problem under Uncertainty
- **Follow the Regularized Leader**
- Follow the Perturbed Leader

Decision Making under Uncertainty

- Find **the best decision** based on **imperfect observations** with **unknown outcomes**

- Problem Formulation
  - Multi-Armed Bandit / Contextual Bandit / Reinforcement Learning

(Stochastic) Multi-Armed Bandit

$$a_t \longrightarrow R_t$$

- Set of K actions, unknown rewards

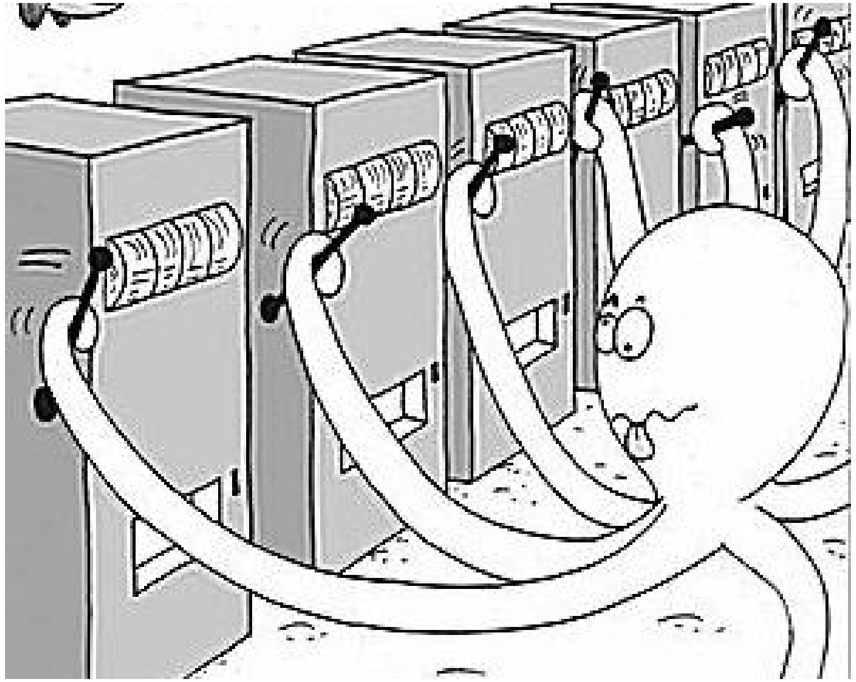
  $$\mathcal{A} = \{a_1, \cdots, a_K\} \quad r_a \in [0, 1]$$

- Given information
  - Noisy reward

  $$R_{t,a} = r_a + \epsilon_{t,a}$$

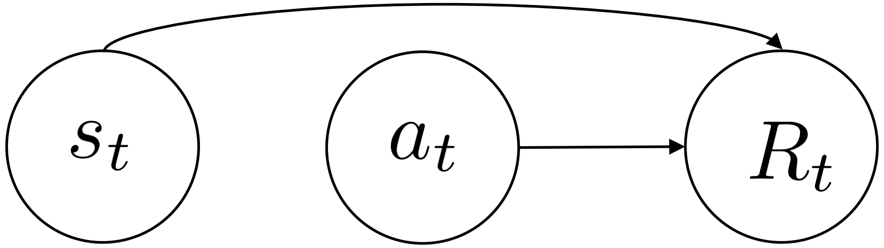
- Find the best action
  - Maximum true reward

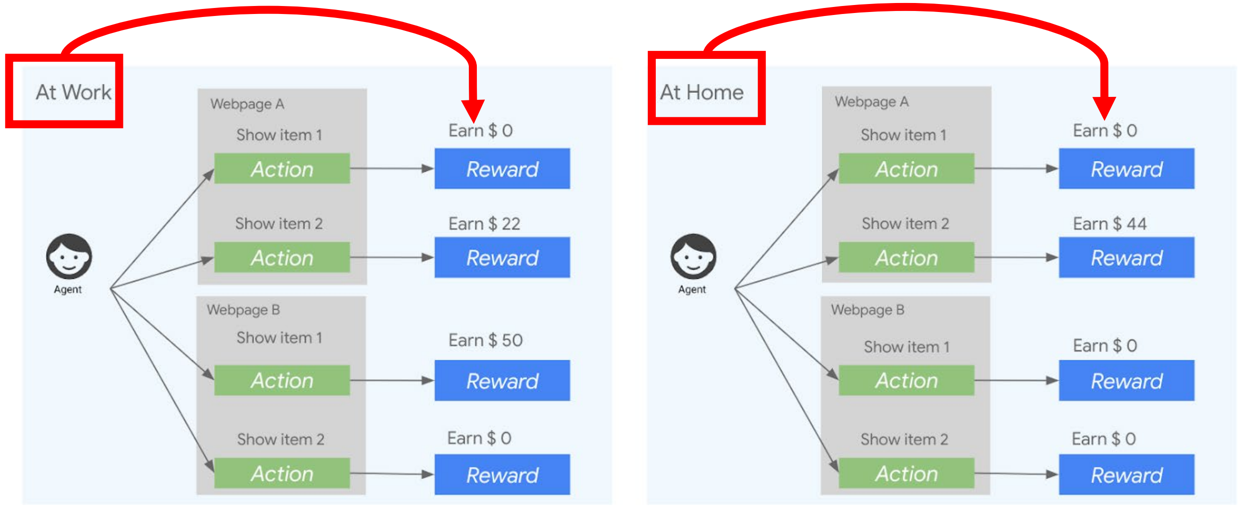
  $$a^\star := \arg\max_a r_a$$



(Stochastic) Contextual Bandit

- Given information
  - **Context (state)**
    $s_t \in \mathcal{S}$
  - Contexts are i.i.d.
  - Noisy reward
    $R_t = r_{a_t}(s_t) + \epsilon_t$
- Given s, find the best action
  $a^\star := \arg\max_a r_a(s)$





Reinforcement Learning

- States are dependent
  - **Markov property**

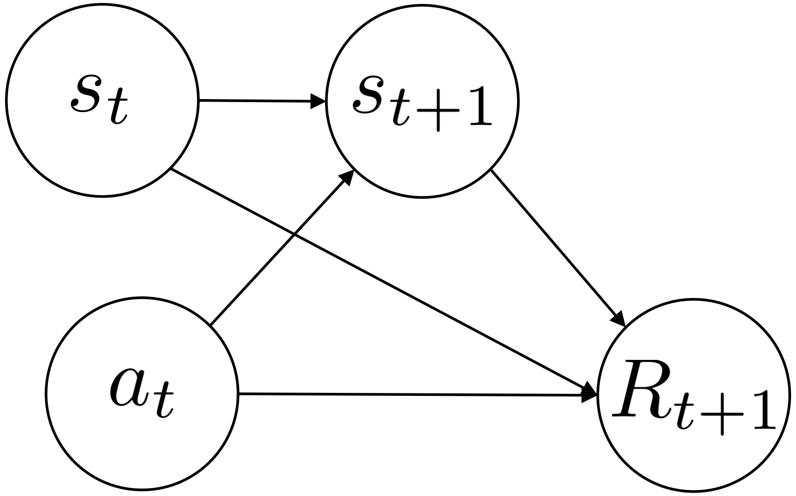
$$P(S_{t+1}|S_t) = P(S_{t+1}|S_t, \cdots, S_0)$$

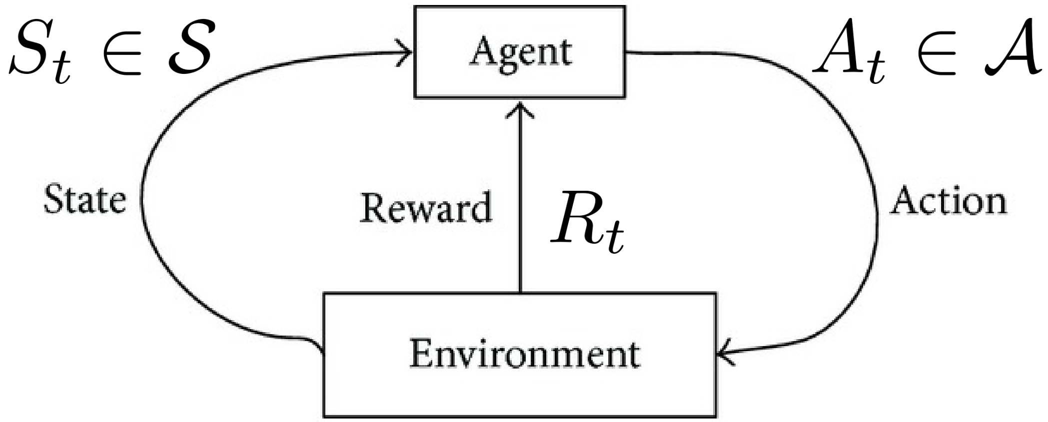
  - Prediction from previous states

$$R_{t+1} = r(s_t, a_t, s_{t+1}) + \epsilon_{t+1}$$

- Return: sum of rewards

$$E\left[\sum_{t=0}^{T} R_{t+1}\right]$$

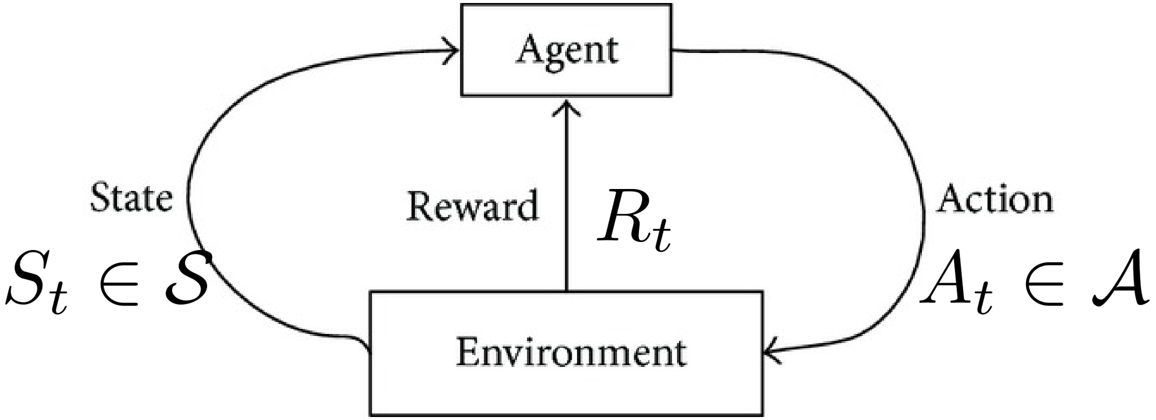


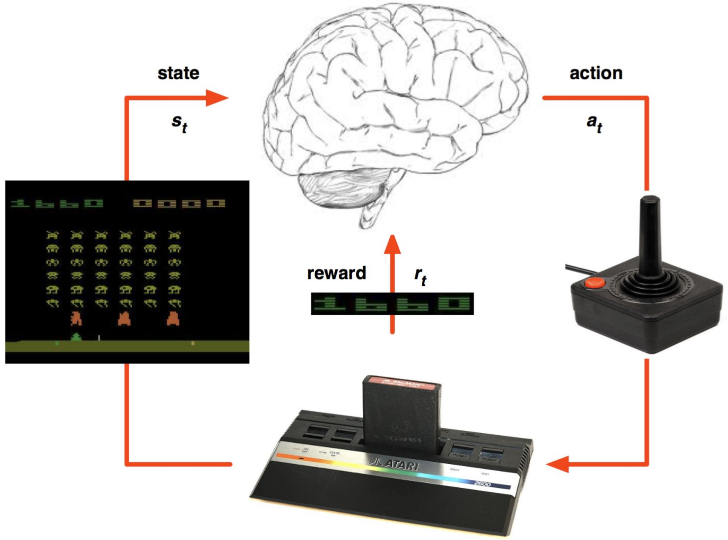


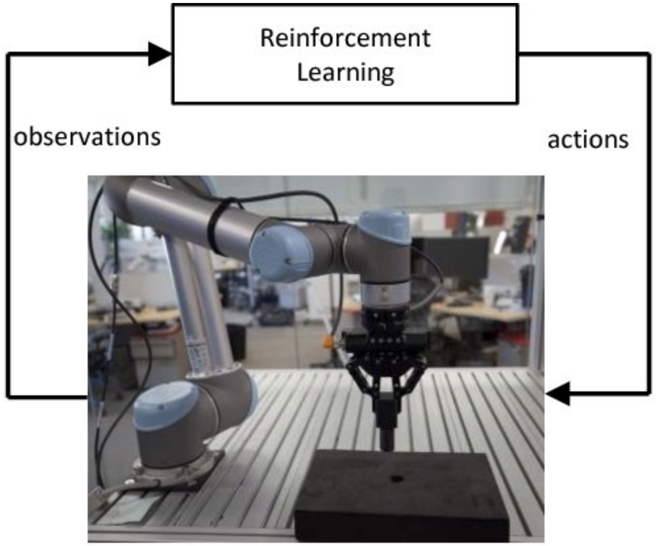
Reinforcement Learning

- Given s, find the best action

$$a^\star := \arg\max E\left[\sum_{t=0}^{T} R_{t+1} \middle| s, a\right]$$

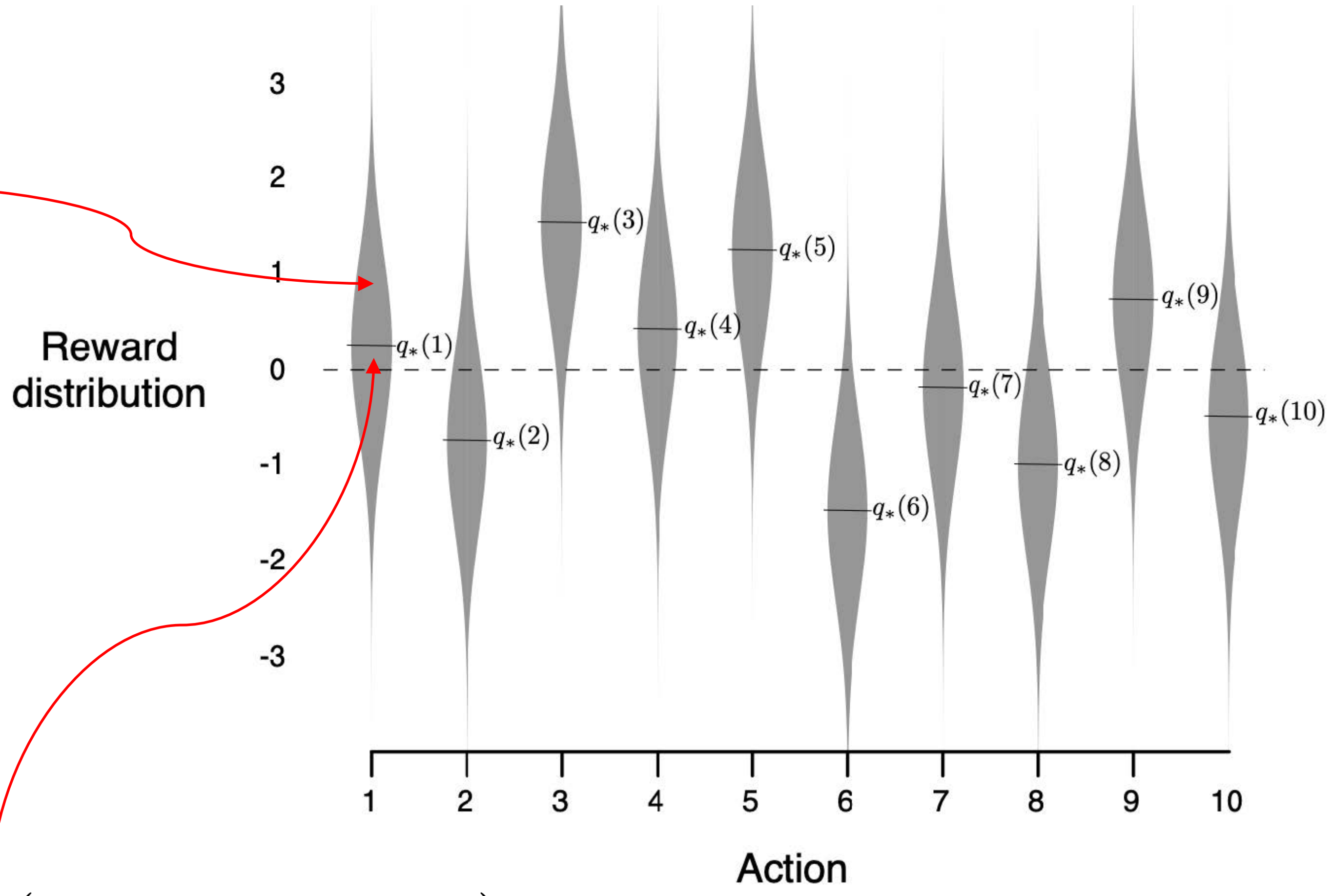






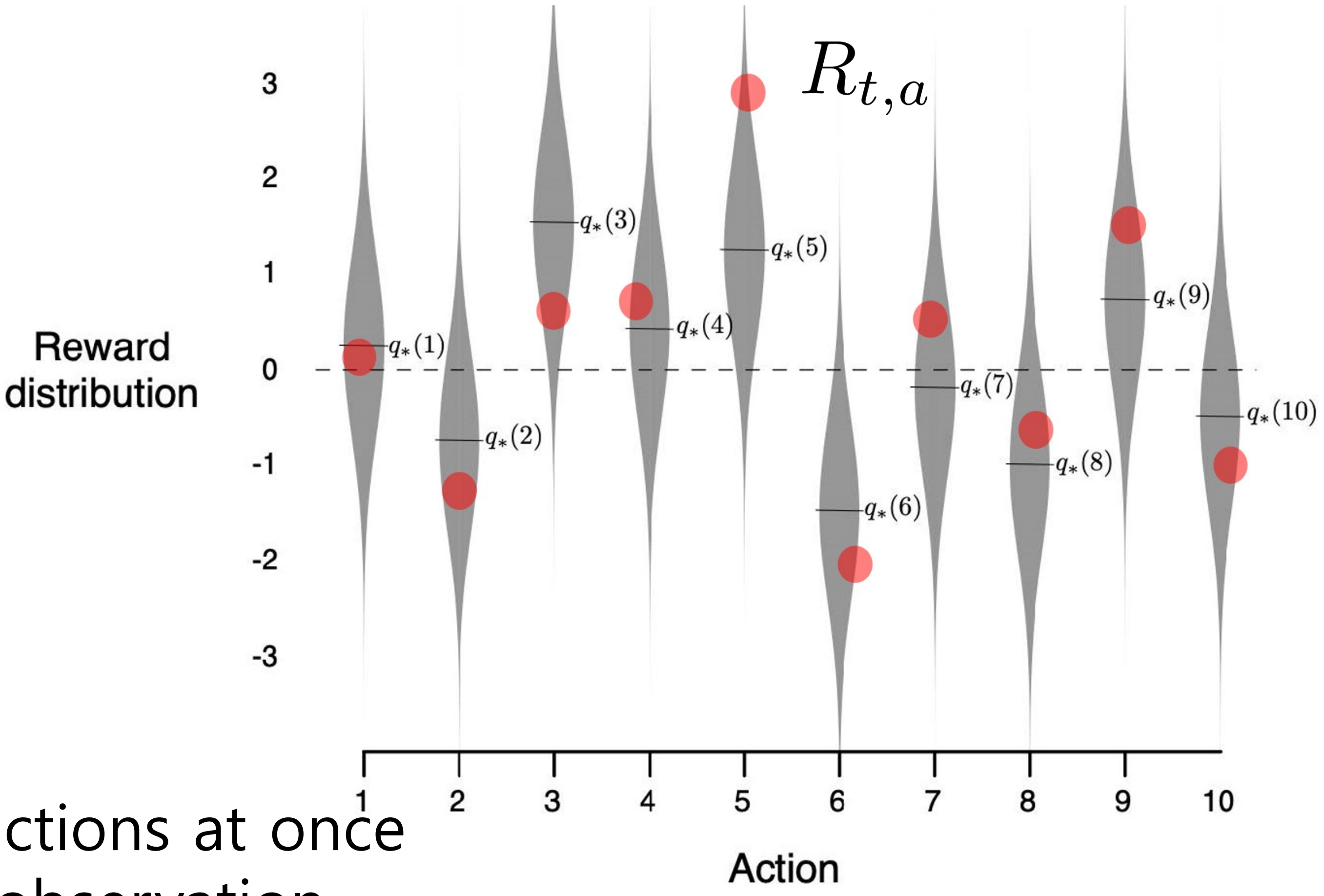
Example: 10-Armed Bandit

$\epsilon_{t,a}$



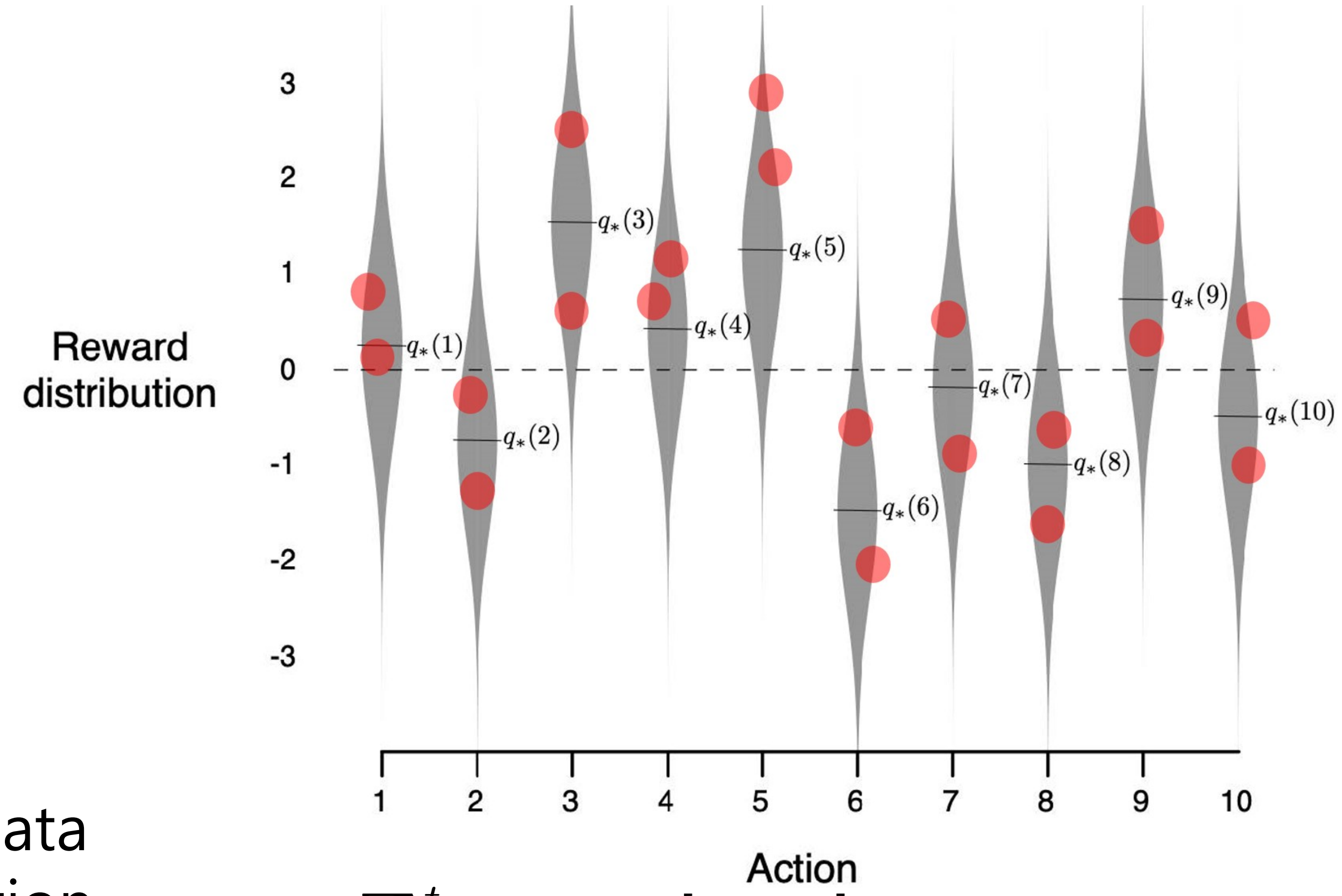
r_a:True Reward (Mean Reward)

Example: 10-Armed Bandit



- Choose every actions at once
- Red dot: noisy observation

Example: 10-Armed Bandit



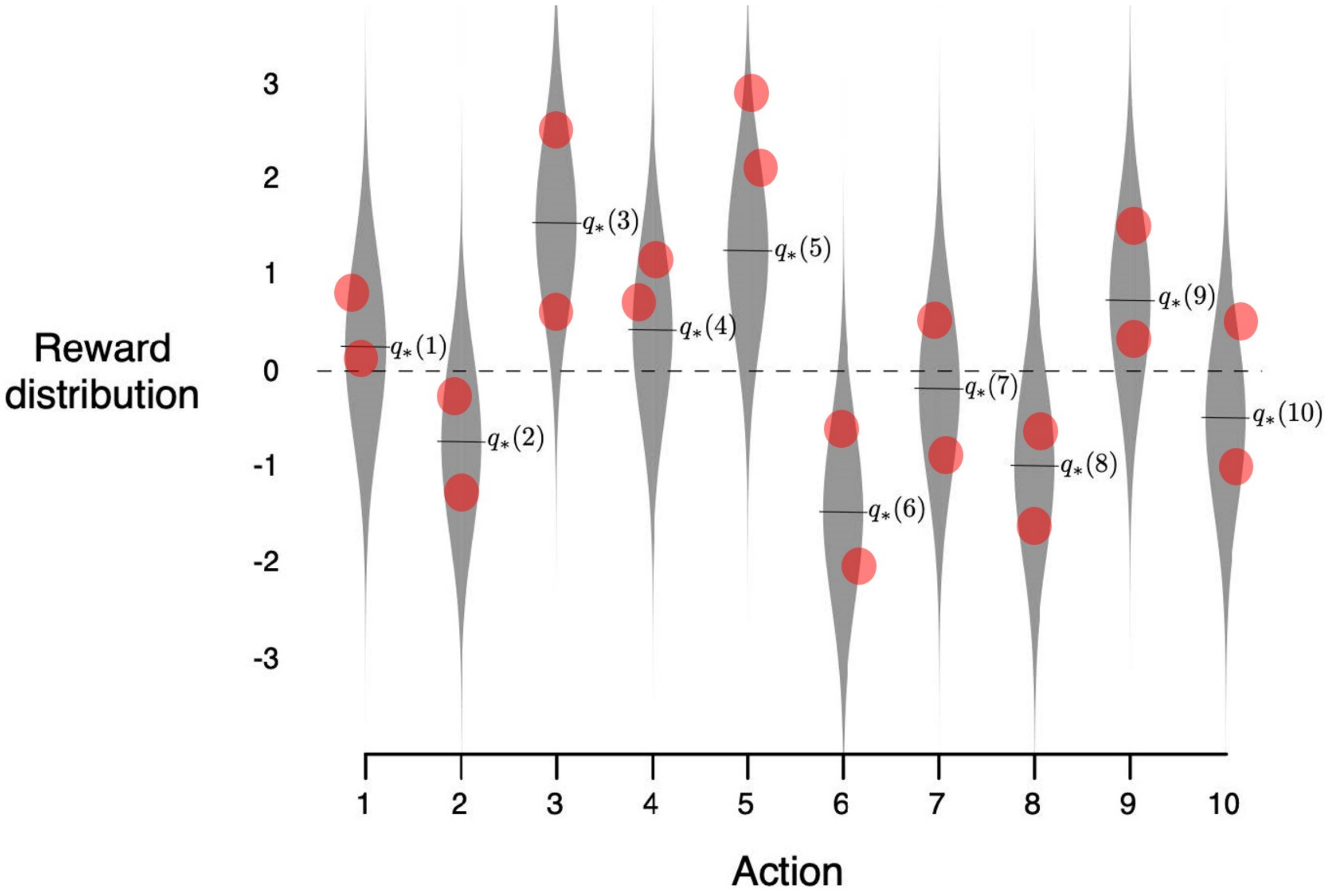
- Collect more data
- Precise estimation

$$\hat{r}_{t,a} := \frac{\sum_{s=1}^{t} R_{s,a} I[a_s = a]}{n_t(a)}$$

Exploration vs. Exploitation

- The meaning of decision (choosing an action)
  - Choose the most uncertain action [Exploration]
    - Pros: more accurate estimation of rewards
    - Cons: loos too much rewards
  - Choose the best action based on collected information [Exploitation]
    - Pros: maximizing rewards
    - Cons: the best action may be sub-optimal due to the noisy rewards

Exploration vs. Exploitation



Precise estimation vs. Maximizing rewards

Exploration vs. Exploitation

- Restaurant Selection
    - Exploitation Go to your favorite restaurant
    - Exploration Try a new restaurant
- Online Banner Advertisements
    - Exploitation Show the most successful advert
    - Exploration Show a different advert
- Oil Drilling
    - Exploitation Drill at the best-known location
    - Exploration Drill at a new location
- Game Playing
    - Exploitation Play the move you believe is best
    - Exploration Play an experimental move

Efficiency of Exploration

- How to measure the efficiency of exploration?
- Regret

$$l_t = \max_{a'} r_{a'} - E_{a_t}[r_{a_t}]$$

$$r^\star := \max_{a'} r_{a'}$$

- Cumulative Regret

$$\mathcal{L}_T = \sum_{t=1}^{T} l_t = T \cdot r^\star - \sum_{t=1}^{T} E_{a_t}[r_{a_t}]$$

- Maximize cumulative reward == minimize total regret

Efficiency of Exploration

- Counting

$$N_T(a) := \sum_{t=1}^{T} I[a_t = a]$$

- Gap

$$\Delta_a := r^\star - r_a$$

  - Note $\Delta_{a^\star} = 0$

$$\begin{aligned}
\mathcal{L}_T &= \sum_{t=1}^{T} l_t = T \cdot r^\star - \sum_{t=1}^{T} E_{a_t}\left[r_{a_t}\right] \\
&= \sum_{t=1}^{T} E_{a_t}\left[\sum_a I[a_t = a](r^\star - r_{a_t})\right] \\
&= \sum_a E[N_T(a)](r^\star - r_{a_t}) \\
&= \sum_a E[N_T(a)]\Delta_a
\end{aligned}$$

Efficiency of Exploration

- Cumulative Regret

$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a$$

- Naïve Approach
  - Fully random search

$$E[I[a_t = a]] = \frac{1}{K} \qquad E[N_T(a)] = \frac{T}{K}$$

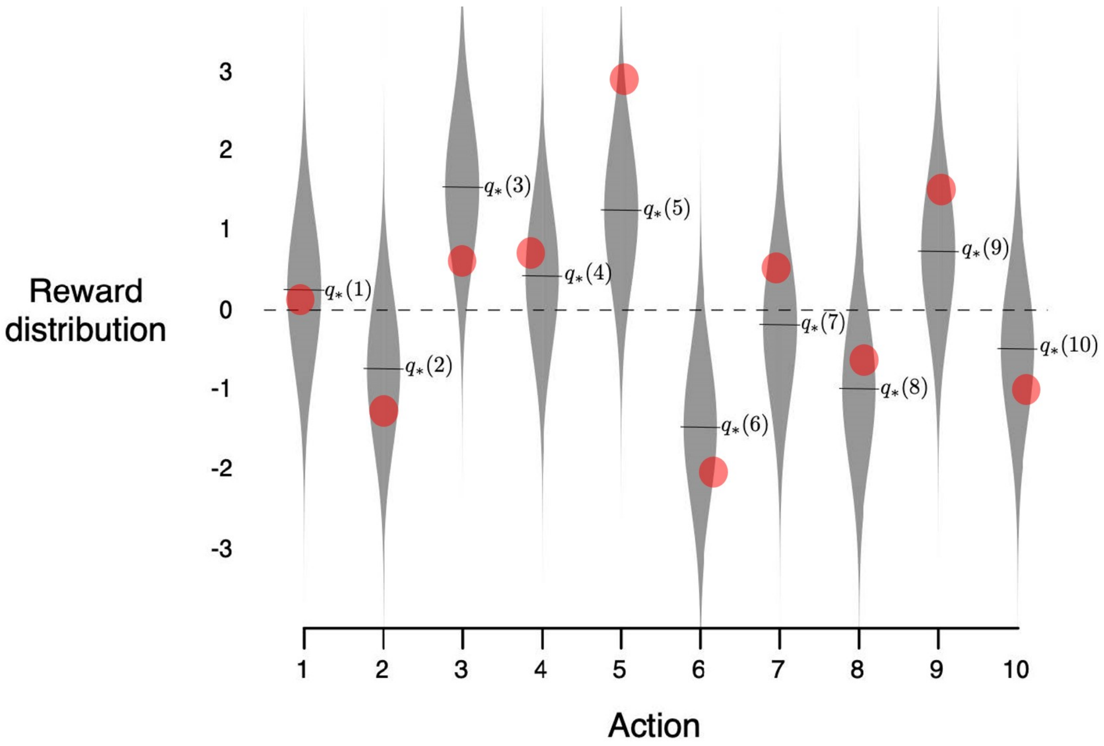
$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a = \frac{T}{K}\sum_a \Delta_a$$ Linear w.r.t. T

Efficiency of Exploration

- Cumulative Regret

$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a$$

- Naïve Approach
  - Fully random search
  - Greedy search
    - If noise is bounded
    - A greedy policy stuck with a sub-optimal action a'



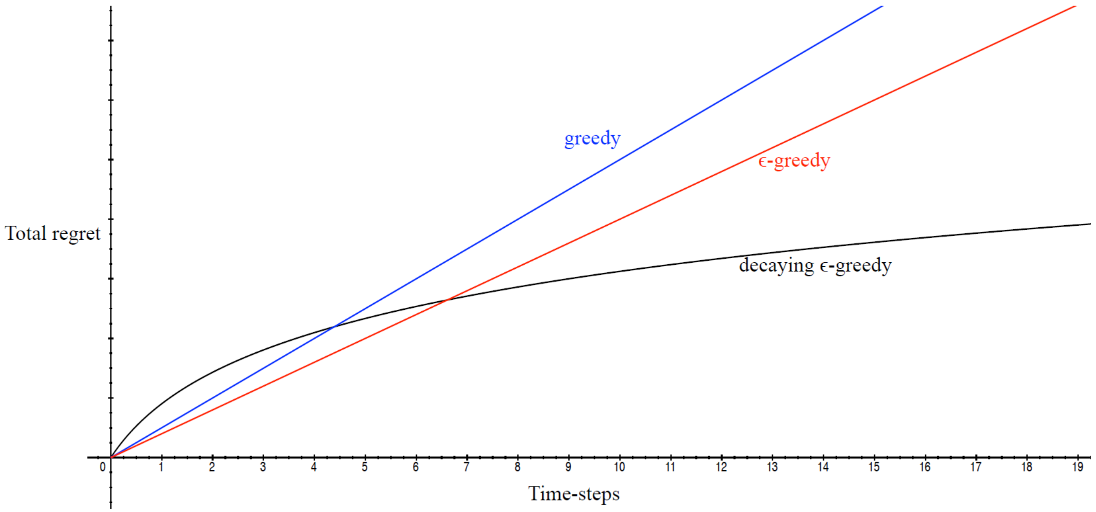
$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a = T\Delta_{a'}$$ Linear w.r.t. T

Efficiency of Exploration

- Cumulative Regret

$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a$$

- Naïve Approach
  - Fully random search
  - Greedy search
    - If noise is bounded
    - A greedy policy stuck with a sub-optimal action a'



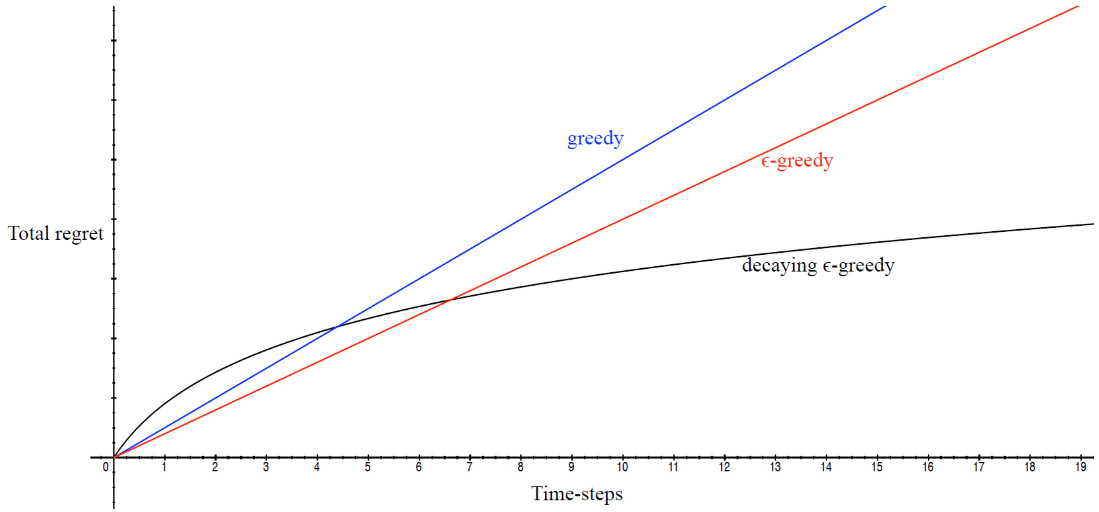
$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a = T\Delta_{a'}$$ Linear w.r.t. T

Efficiency of Exploration

- Cumulative Regret

$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a$$

- Naïve Approach
  - Fully random search
  - Greedy search
  - Eps-Greedy search



$$a_t = \begin{cases} \arg\max_a \hat{r}_a \text{ w.p. } 1-\epsilon \\ \text{Uniform}(K) \text{ w.p. } \epsilon \end{cases}$$

$$\mathcal{L}_T = \sum_a E[N_T(a)]\Delta_a \geq \frac{T}{K}\sum_a \Delta_a$$ Linear w.r.t. T

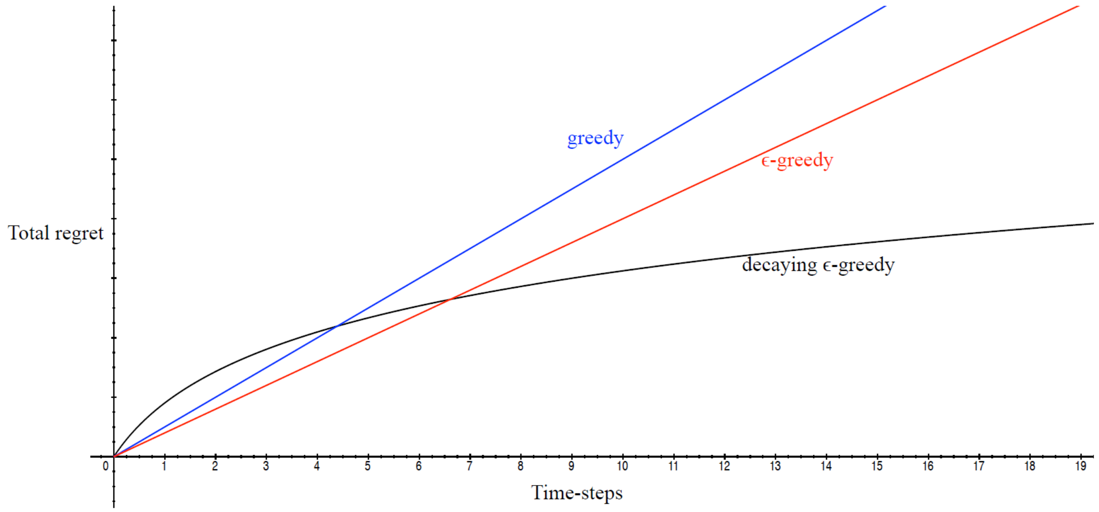
Efficiency of Exploration

- What is the best strategy
  - Sub-linear!
  - (problem-dependent) lower bound

$$\mathcal{L}_T \geq \Omega\left(\ln(T)\sum_a \frac{\Delta_a}{D_{kl}(R_a|R_{a^\star})}\right)$$

  - (problem-independent) lower bound

$$\mathcal{L}_T \geq \Omega\left(\sqrt{KT}\right)$$



Exploration Methods

- Follow-the-Regularized-Leader (FTRL)

- Follow-the-Perturbed-Leader (FTPL)

Follow-the-Regularized-Leader (FTRL)

- Regularized Policy (Categorical distribution)

$$\pi := \arg\max_{\Delta^K} \boxed{E_{a\sim\pi}[\hat{r}_a]} + \alpha \boxed{\Phi(\pi)}$$

Greedy (Exploitation) Regularization (Exploration)

- Concave regularization
  - Makes a policy a uniform distribution

$$\Phi : \Delta^K \to R$$

Follow-the-Regularized-Leader (FTRL)

- Regularized Policy (Categorical distribution)

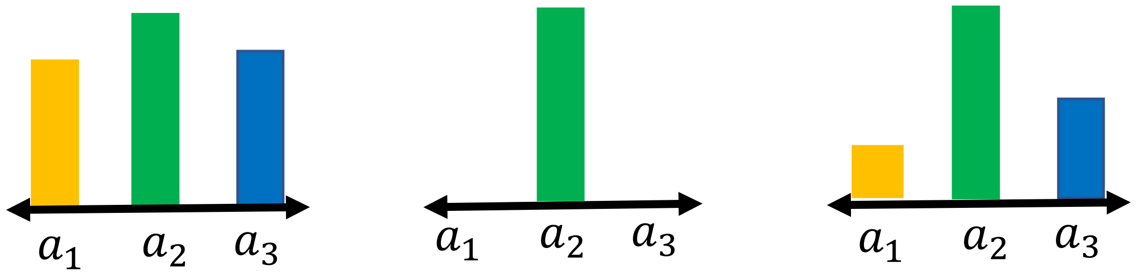
$$\pi := \arg\max_{\Delta^K} \underbrace{\boxed{E_{a\sim\pi}[\hat{r}_a]}}_{\text{Greedy (Exploitation)}} + \alpha \underbrace{\boxed{\Phi(\pi)}}_{\text{Regularization (Exploration)}}$$

  - Greedy Policy

$$\pi := \arg\max_{\Delta^K} E_{a\sim\pi}[\hat{r}_a]$$

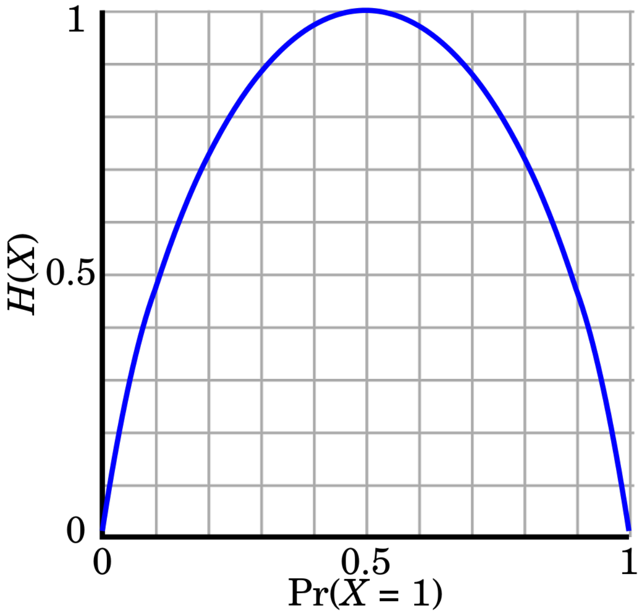
  - How to control trade-off?

$$\alpha_t = f(t)$$



Estimated Reward | Greedy Policy | Regularized Policy

Boltzmann Exploration



- Shannon-Gibbs Entropy

$$\Phi(\pi) = E_{a\sim\pi}[-\ln(\pi_a)]$$

$$\pi := \arg\max_{\Delta^K} E_{a\sim\pi}[\hat{r}_a] + \alpha E_{a\sim\pi}[-\ln(\pi_a)]$$

- Softmax distribution / Boltzmann distribution (Stochastic Bandit)

$$\pi_{t,a} = \frac{\exp(\hat{r}_a/\alpha_t)}{\sum_{a'} \exp(\hat{r}_{a'}/\alpha_t)} \quad \alpha_t^{-1} = \Theta(\ln(t))$$

$$\mathcal{L}_T \leq O\left(\frac{\ln(T)}{\min_{a\neq a^\star} \Delta_a}\right)$$

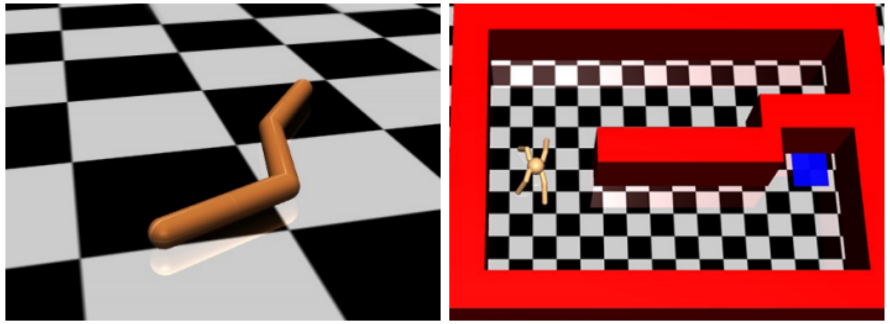
Cesa-Bianchi, Nicolò, et al. "Boltzmann exploration done right." *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017.

Soft Q-Learning

- Shannon-Gibbs Entropy in RL

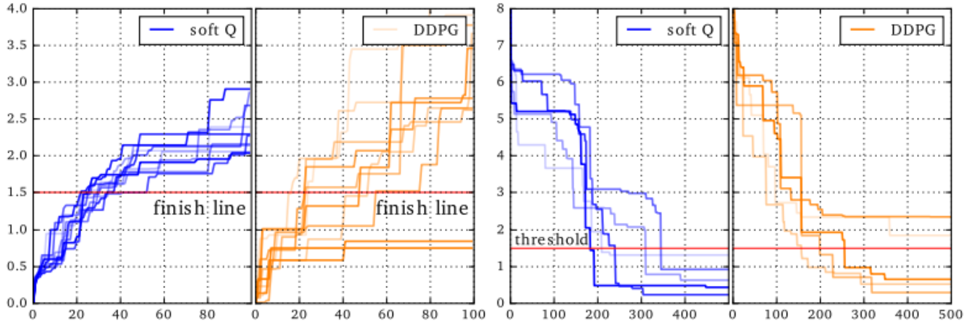
$$\max_{\pi'} E_{a\sim\pi'}\left[\hat{Q}(s,a) - \alpha\ln(\pi_a)\right]$$

- Practical benefit
  - Multi-modal exploration
  - Learning multi-modal behavior



(a) Swimming snake (b) Quadrupedal robot

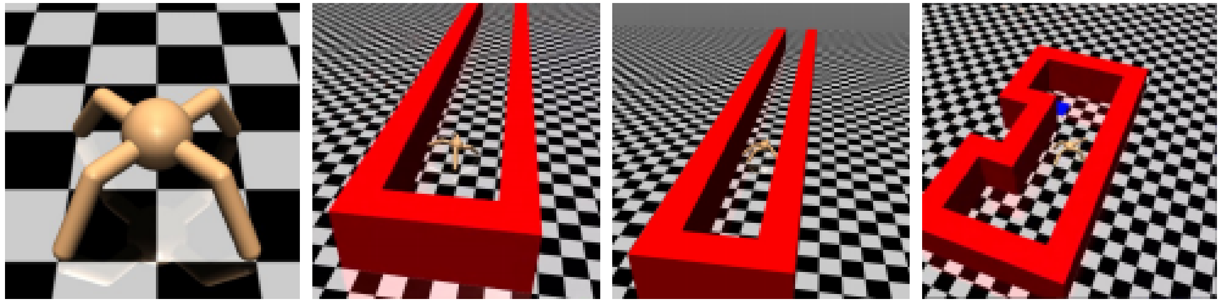
Figure 2. Simulated robots used in our experiments.



(a) Swimmer (higher is better) (b) Quadruped (lower is better)

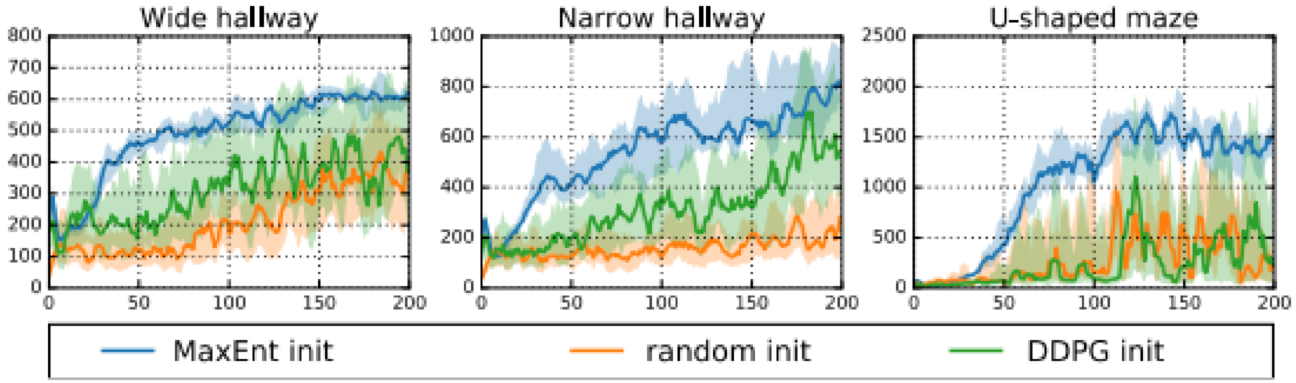
Soft Q-Learning

- Shannon-Gibbs Entropy in RL
  - a) move free direction
  - b) wide hallway
  - c) narrow hallway
  - d) U-shaped maze



(a) (b) (c) (d)

- Practical benefit
  - Transfer learning
    (provide better initialization)

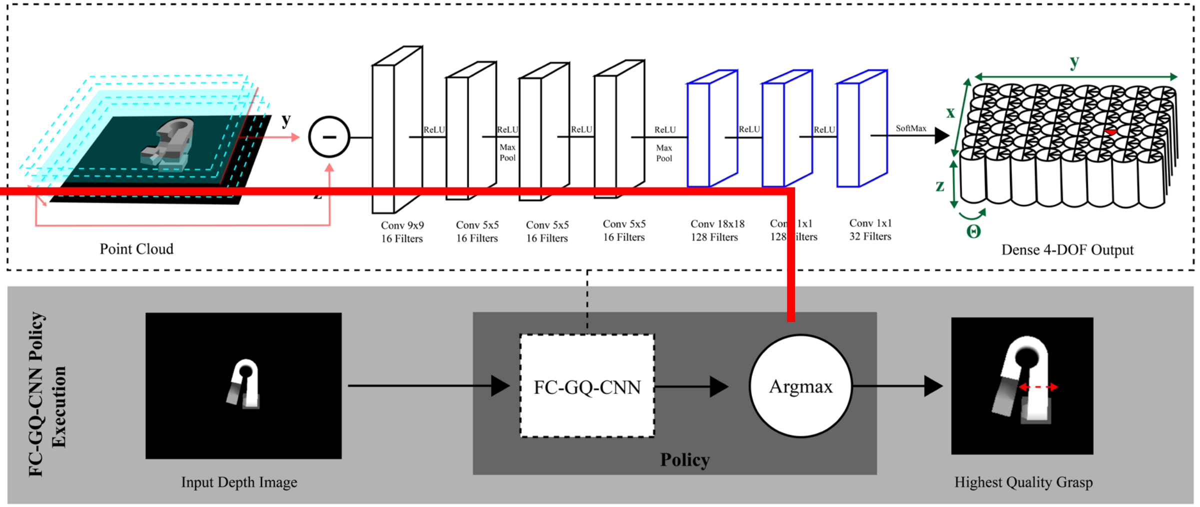


Shannon Entropy Regularized Neural Contextual Bandit

- Softmax distribution

$$\pi(a|s) = \frac{\exp(\hat{r}_a(s)/\alpha)}{\sum_{a'} \exp(\hat{r}_{a'}(s)/\alpha)}$$

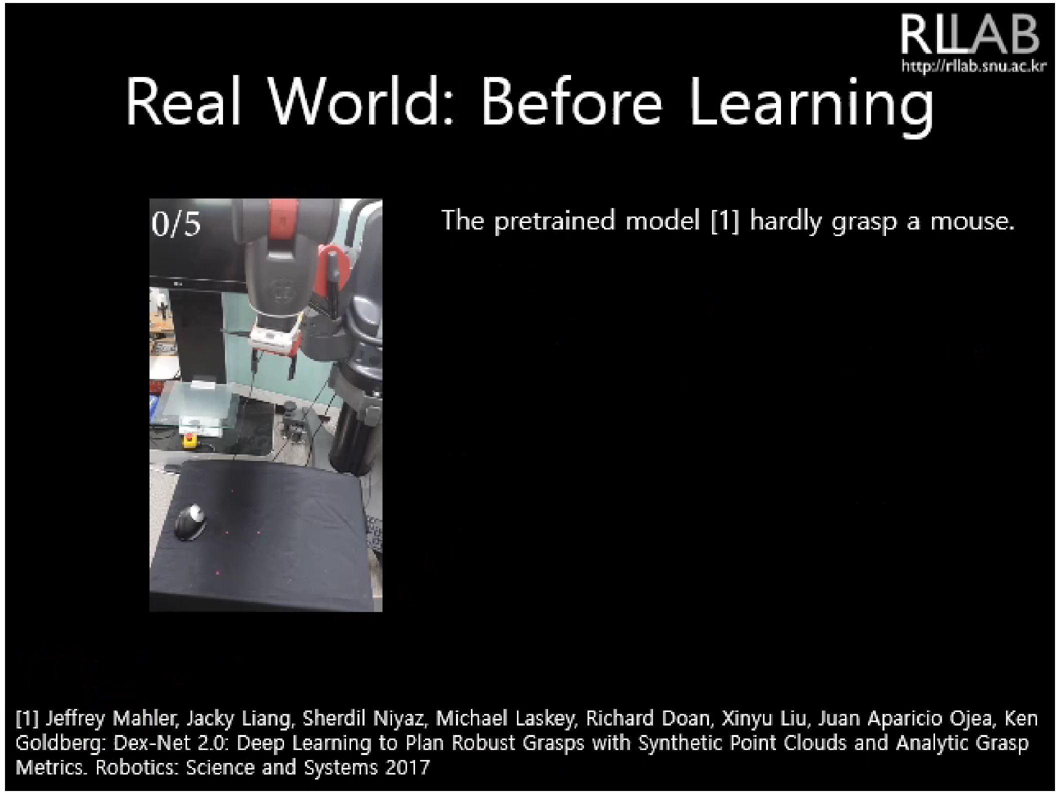
  - Context : depth image
  - Action : where to grasp (x, y, θ)

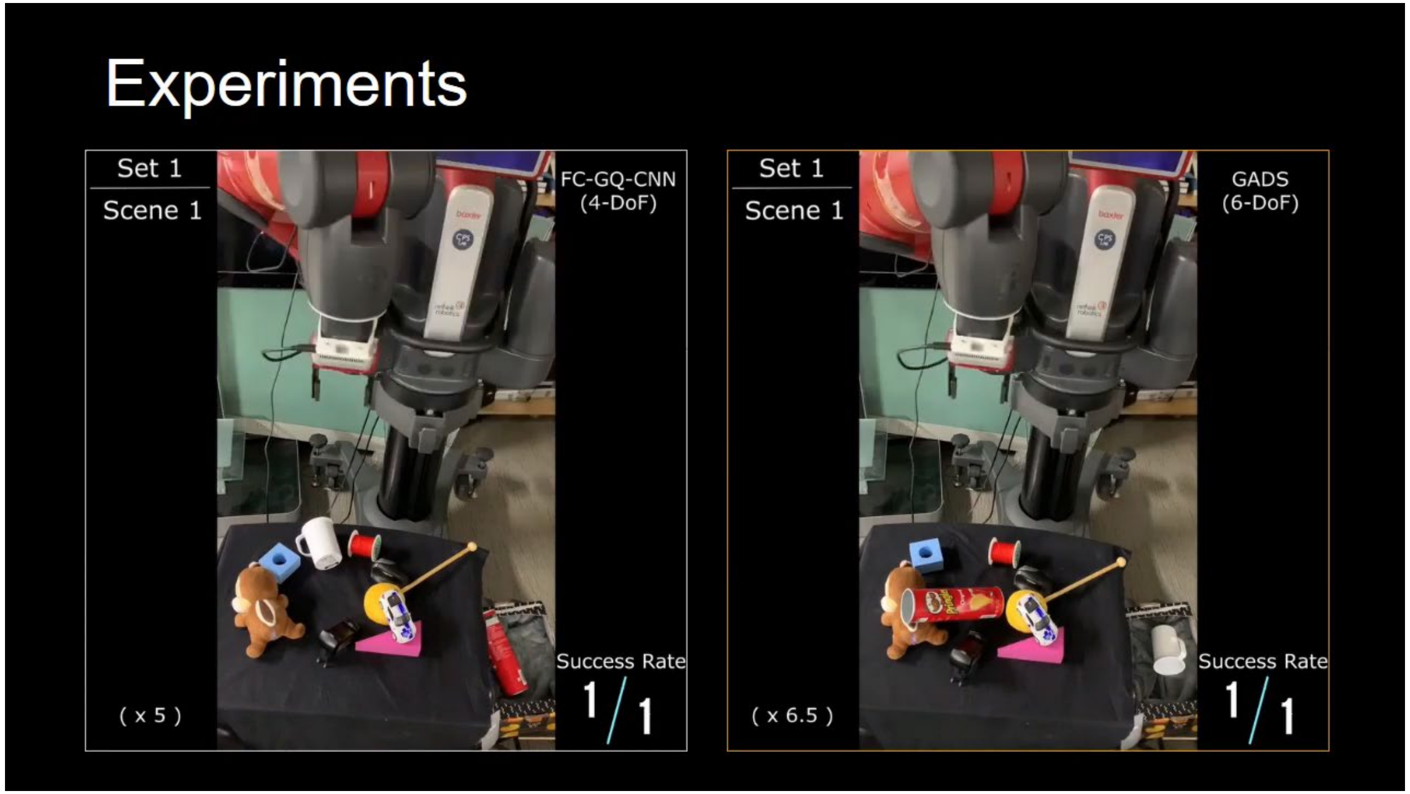


Satish, Vishal, Jeffrey Mahler, and Ken Goldberg. "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks." *IEEE Robotics and Automation Letters* 4.2 (2019): 1357-1364.

- Practical benefit
  - Searching promising actions first

Shannon Entropy Regularized Neural Contextual Bandit



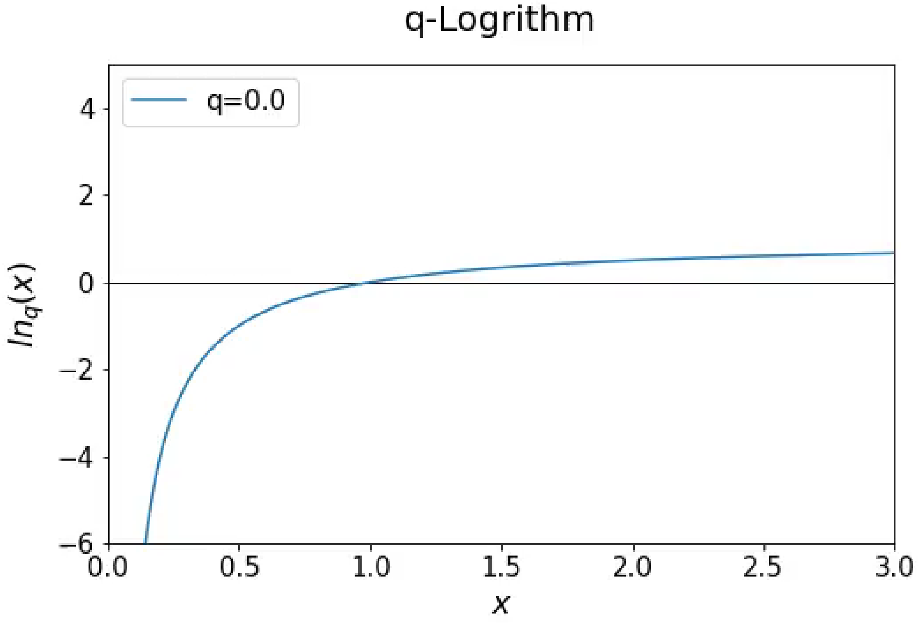


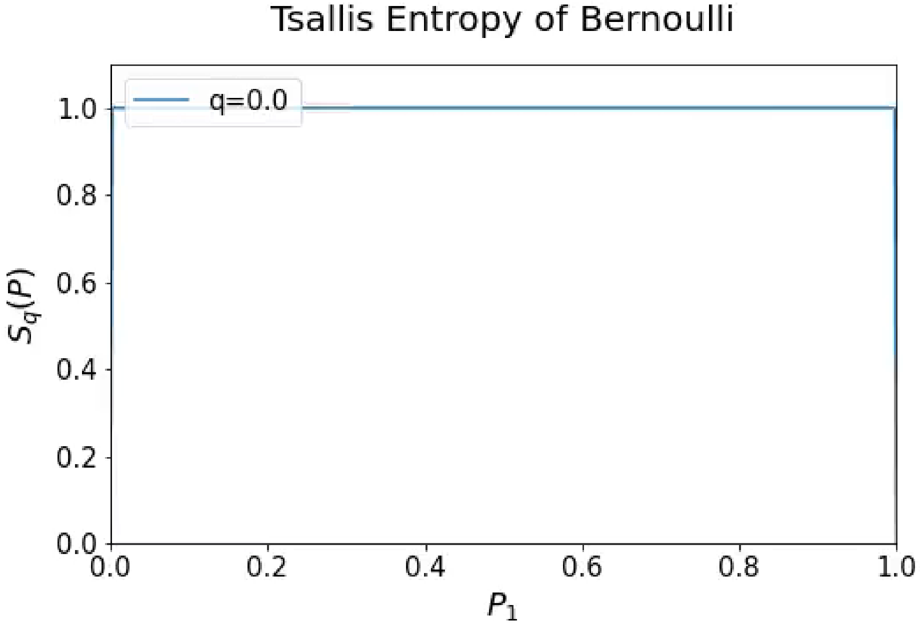
K. Lee (SNU), J. Choy (SNU), Y. Choi (SNU), H. Kee (SNU), S. Oh (SNU), “No-Regret Shannon Entropy Regularized Neural Contextual Bandit Online Learning for Robotic Grasping”, IROS, Nov. 2020
Y. Choi (SNU), H. Kee (SNU), **K. Lee (SNU)**, J. Choy (SNU), and S. Oh (SNU), ”Hierarchical 6-DoF Grasping with Approaching Direction Selection”, ICRA, May 2020.

Tsallis Entropy

$$\ln_q(x) = \frac{x^{q-1} - 1}{q-1} \qquad \Phi(\pi) = E_{a\sim\pi}\left[-\ln_q(\pi_a)\right]$$

Entropic Index (q>0) : a positive parameter controlling a type of entropy





Tsallis Entropy Reinforcement Learning

- Objective Function of Tsallis Entropy RL

$$\max_{\pi'} E_{a\sim\pi'}\left[\hat{Q}(s,a) - \alpha \ln_q(\pi_a)\right]$$

- Special Cases:
  - If $q = 1$: Shannon Gibbs entropy, Soft MDPs
  - If $q \to \infty$: $S_q \to 0$, Original MDPs without regularization

- Different entropic indices induce different optimal policies

K. Lee (SNU), S. Kim (KAIST), S. Lim (UNIST), S. Choi (SNU), M. Hong (SNU), J. Kim (SNU), Y. Park (SNU), and S. Oh (SNU), "Generalized Tsallis Entropy Reinforcement Learning and Its Application to Soft Mobile Robots" RSS, 2020.

Evaluation Task 1

x6

Follow-the-Perturbed-Leader (FTPL)

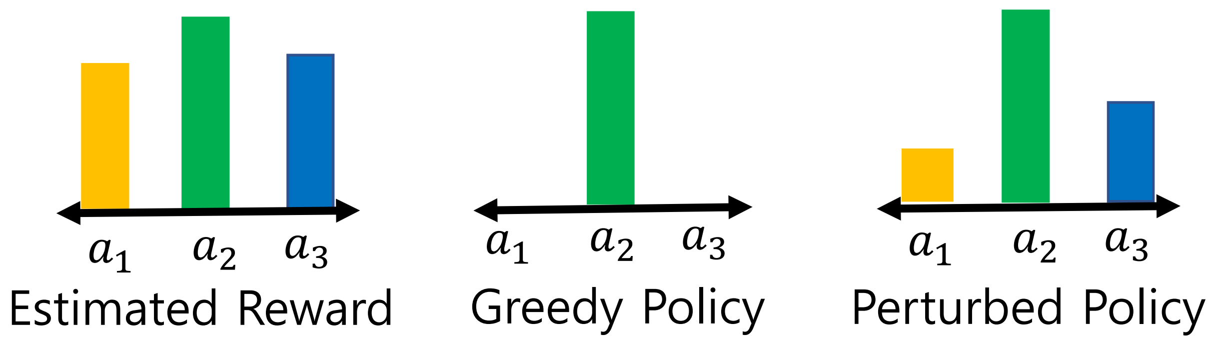
- Perturbed Policy (for stochastic bandit)

$$a_t := \arg\max_a \underbrace{\boxed{\hat{r}_a}}_{\text{Greedy (Exploitation)}} + \frac{1}{\sqrt{n_a}} \underbrace{\boxed{G_a}}_{\text{Perturbation (Exploration)}}$$

- Random Perturbation

$$G_a \sim P_G$$

  - Gumbel distribution
  - Fréchet distribution
  - Weibull distribution
  - …



Follow-the-Perturbed-Leader (FTPL)

- Perturbed Policy

$$a_t := \arg\max_a \boxed{\hat{r}_a} + \frac{1}{\sqrt{n_a}} \boxed{G_a}$$

Greedy (Exploitation) Perturbation (Exploration)

- It is hard to obtain policy distribution explicitly
- How to control trade-off?

$$\frac{1}{\sqrt{n_a}} \to 0 \text{ as } n_a \to \infty$$

Perturbation (Exploration) $\longrightarrow$ Greedy (Exploitation)

Follow-the-Perturbed-Leader (FTPL)

Distribution	$\sup_x h_{\mathcal{D}}(x)$	$\mathbb{E}[\max_{i=1}^{N} Z_i]$	$\boxed{O(\sqrt{TN \log N})}$ Param.
Gumbel$(\mu = 1, \beta = 1)$	1 as $x \to 0$	$\log N + \gamma_0$	N/A
Frechet $(\alpha > 1)$	at most 2α	$N^{1/\alpha}\Gamma(1 - 1/\alpha)$	$\alpha = \log N$
Weibull*$(\lambda = 1, k \leq 1)$	k at $x = 0$	$O((\frac{1}{k})!(\log N)^{\frac{1}{k}})$	$k = 1$ (Exponential)
Pareto*$(x_m = 1, \alpha)$	α at $x = 0$	$\alpha N^{1/\alpha}/(\alpha - 1)$	$\alpha = \log N$
Gamma$(\alpha \geq 1, \beta)$	β as $x \to \infty$	$\log N + (\alpha - 1) \log \log N - \log \Gamma(\alpha) + \beta^{-1}\gamma_0$	$\beta = \alpha = 1$ (Exponential)

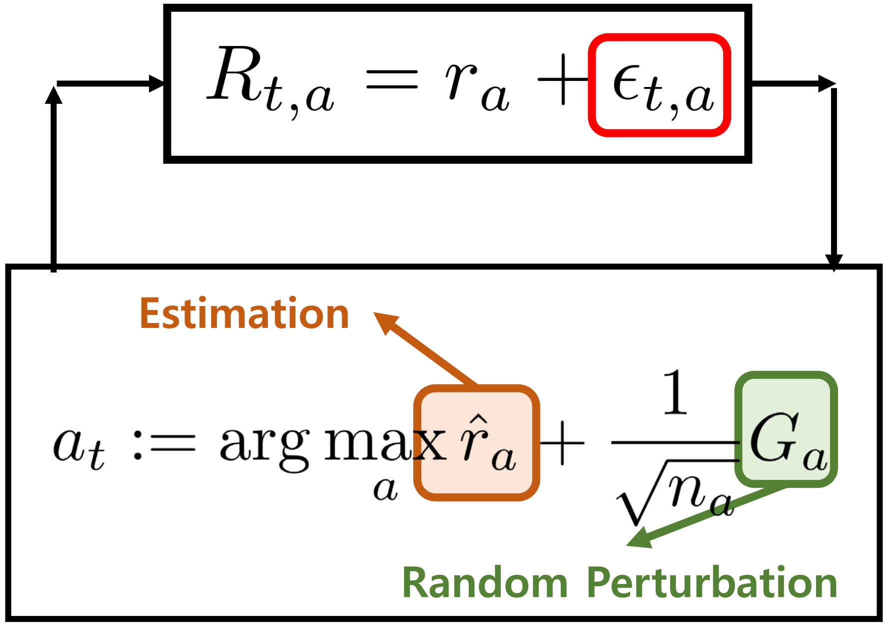
Abernethy, Jacob, Chansoo Lee, and Ambuj Tewari. "Fighting bandits with a new kind of smoothness." *arXiv preprint arXiv:1512.04152* (2015).

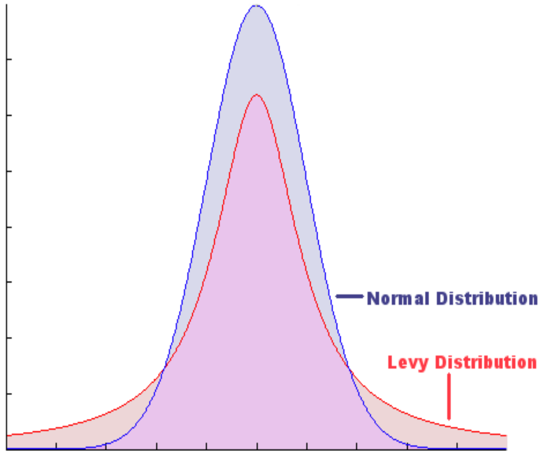
- Reminder! $\mathcal{L}_T \geq \Omega\left(\sqrt{KT}\right)$

$$\mathcal{L}_T \geq \Omega\left(\ln(T) \sum_a \frac{\Delta_a}{D_{kl}(R_a | R_{a^\star})}\right)$$

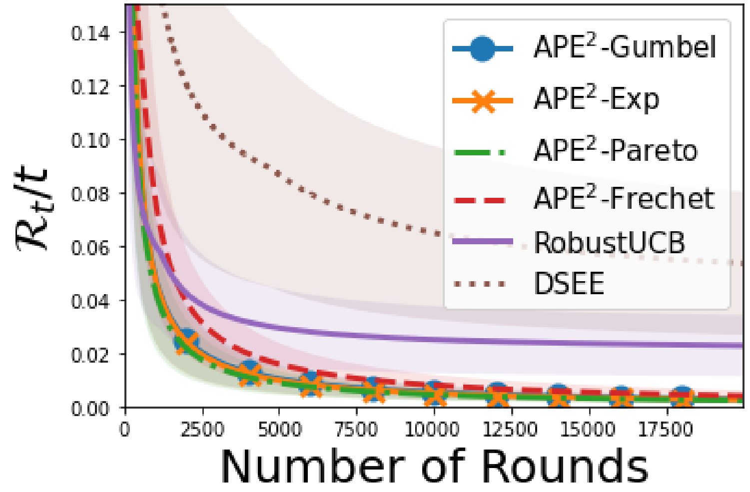
Adaptively Perturbed Exploration

- Perturbed Exploration for Heavy Tailed Noise





Heavy tailed Noise



Convergence Speed

K. Lee (SNU), H. Yang (UNIST), S. Lim (UNIST) and S. Oh (SNU), " Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy Tailed Rewards", **NeurIPS**, Dec. 2020.

Adaptively Perturbed Exploration

- Perturbed Exploration for Heavy Tailed Noise

Dist. on G	Prob. Dep. Bnd. $O(\cdot)$	Prob. Indep. Bnd. $O(\cdot)$	Low. Bnd. $\Omega(\cdot)$	Opt. Params.	Opt. Bnd. $\Theta(\cdot)$
Weibull	$\sum_{a\neq a^\star} A_{c,\lambda,a}\left(\ln\left(B_{c,a}T\right)\right)^{\frac{p}{k(p-1)}}$	$C_{K,T}\ln\left(K\right)^{\frac{1}{k}}$	$C_{K,T}\ln\left(K\right)$	$k=1,\lambda\geq 1$	$\boxed{K^{1-1/p}T^{1/p}\ln\left(K\right)}$
Gamma	$\sum_{a\neq a^\star} A_{c,\lambda,a}\alpha^{p/(p-1)}\ln\left(B_{c,a}T\right)^{p/(p-1)}$	$C_{K,T}\frac{\ln\left(\alpha K^{1+p/(p-1)}\right)^{p/(p-1)}}{\ln(K)^{\frac{1}{p-1}}}$	$C_{K,T}\ln\left(K\right)$	$\alpha=1,\lambda\geq 1$	
GEV	$\sum_{a\neq a^\star} A_{c,\lambda,a}\ln_\zeta\left(B_{c,a}T\right)^{p/(p-1)}$	$C_{K,T}\frac{\ln_\zeta\left(K^{\frac{2p-1}{p-1}}\right)^{p/(p-1)}}{\ln_\zeta(K)^{\frac{1}{p-1}}}$	$C_{K,T}\ln_\zeta\left(K\right)$	$\zeta=0,\lambda\geq 1$	
Pareto	$\sum_{a\neq a^\star} A_{c,\lambda,a}\left[B_{c,a}T\right]^{\frac{p}{\alpha(p-1)}}$	$C_{K,T}\alpha^{1+\frac{p^2}{\alpha(p-1)^2}}K^{\frac{1}{\alpha(p-1)}}$	$C_{K,T}\alpha K^{\frac{1}{\alpha}}$	$\alpha=\lambda=\ln(K)$	
Fréchet	$\sum_{a\neq a^\star} A_{c,\lambda,a}\left[B_{c,a}T\right]^{\frac{p}{\alpha(p-1)}}$	$C_{K,T}\alpha^{1+\frac{p^2}{\alpha(p-1)^2}}K^{\frac{1}{\alpha(p-1)}}$	$C_{K,T}\alpha K^{\frac{1}{\alpha}}$	$\alpha=\lambda=\ln(K)$	

Conclusion

- Efficiency of Exploration Methods
  - Regret Analysis / Regret Lower Bounds
- Follow-the-Regularized-Leader
  - Multi-modal optimal actions
  - Various applications
- Follow-the-Perturbed-Leader
  - Simple implementation

	FTRL	FTPL
Multi-Armed Bandit	O	O
Contextual Bandit	O	O (Linear Model)
Planning	O	?
Reinforcement Learning	O	?