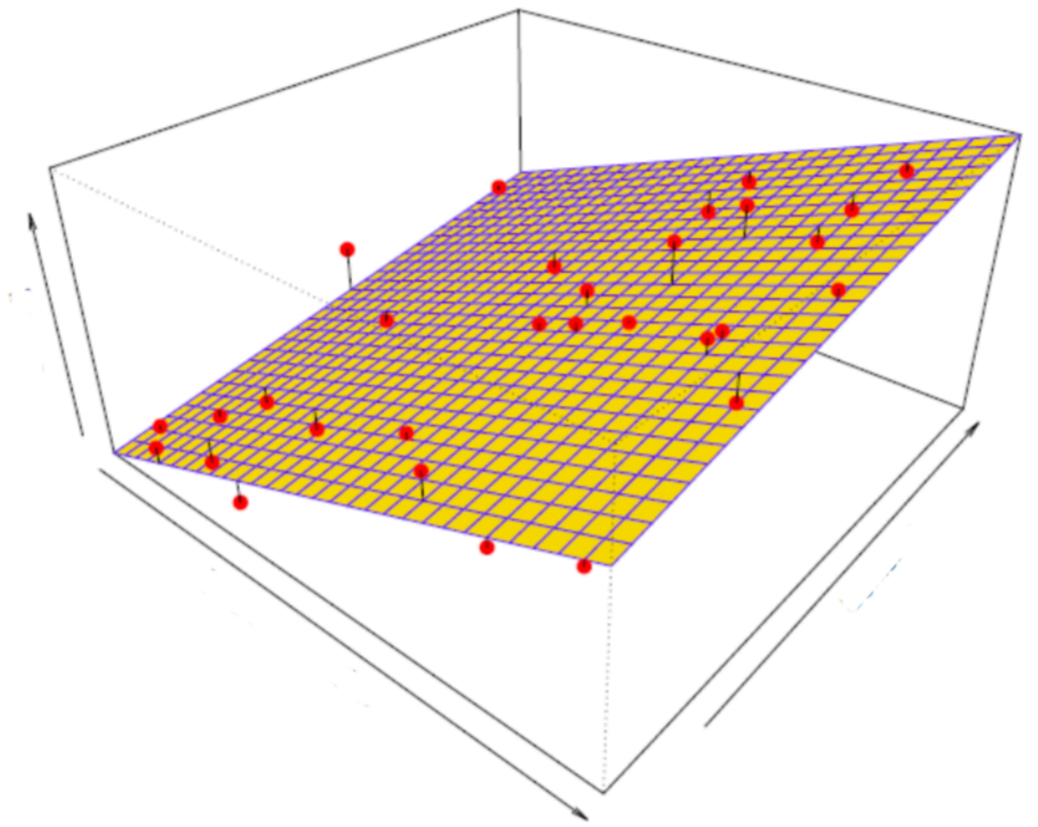
Introduction to Exploration in Reinforcement Learning

CS 234 Recitation

What is Exploration in Reinforcement Learning?

Machine Learning

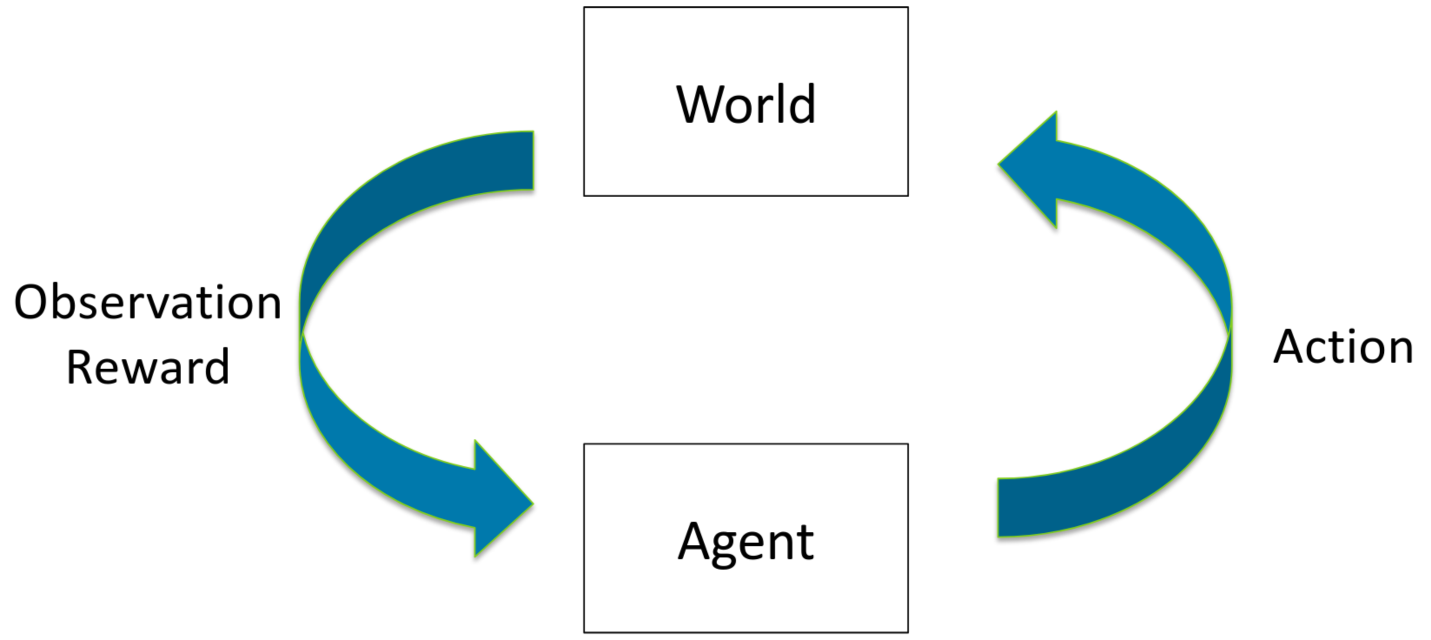
(Learning from data)



Data are given

Reinforcement Learning

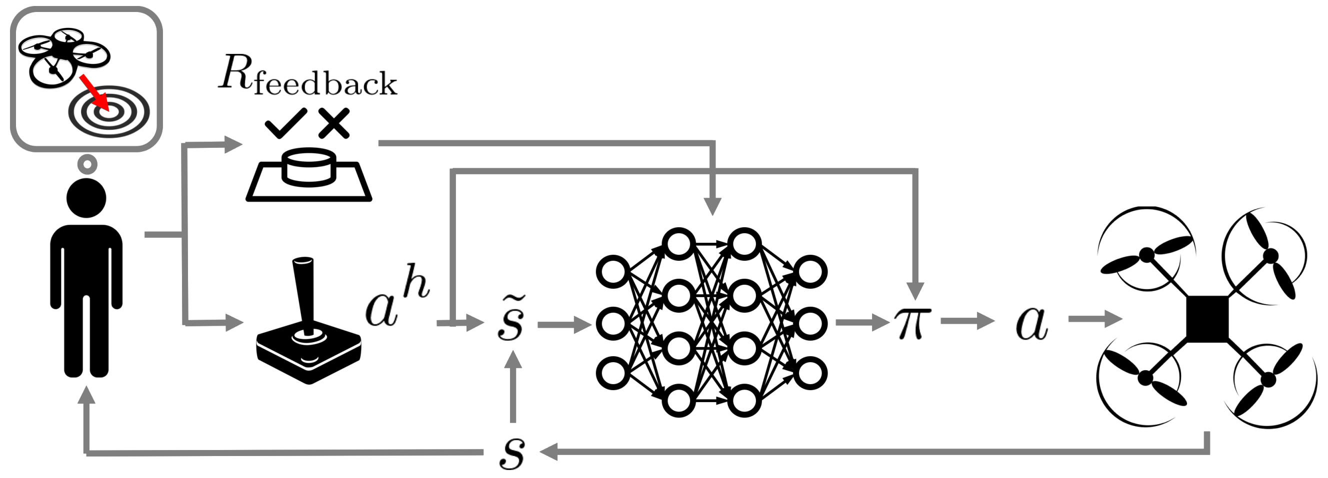
(Learning to make good sequences of decisions)

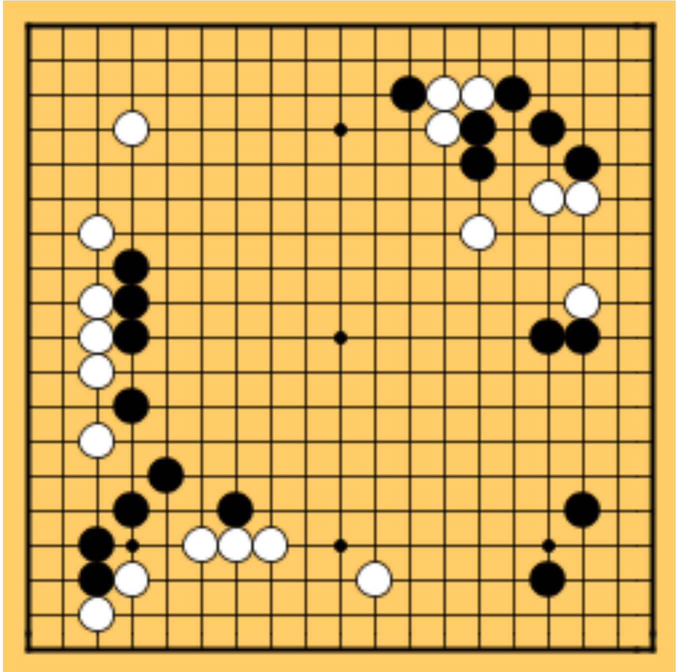


Data are collected by interacting with the world

Exploration
=
sample efficient data collection

Why do we need Efficient Exploration?





Some RL successes are impressive, but…

…need a lot of data

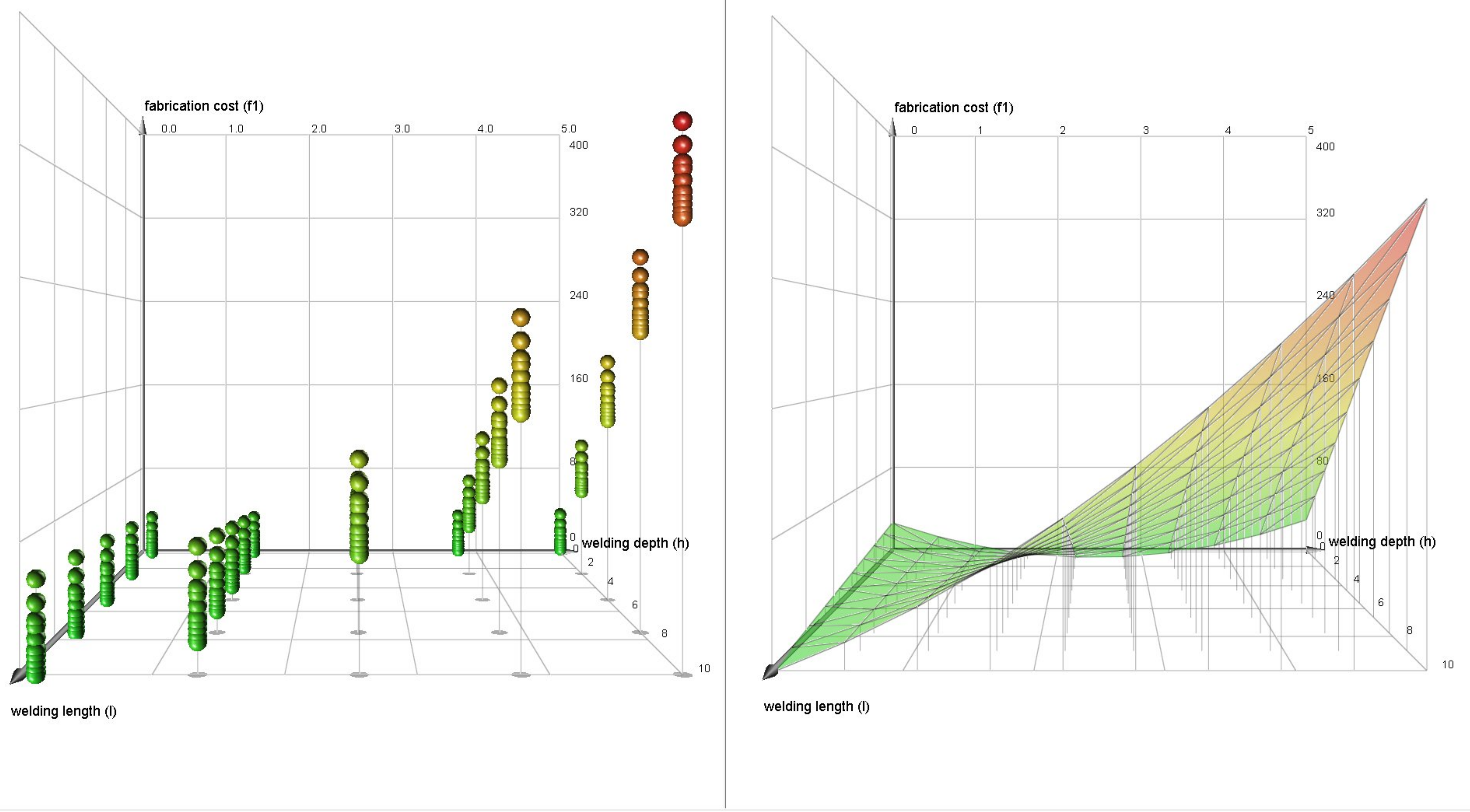
…require extensive fine tuning

Exploration:

Learn <u>efficiently</u> and <u>reliably</u>

Why is Exploration Hard in RL?

Design of Experiments



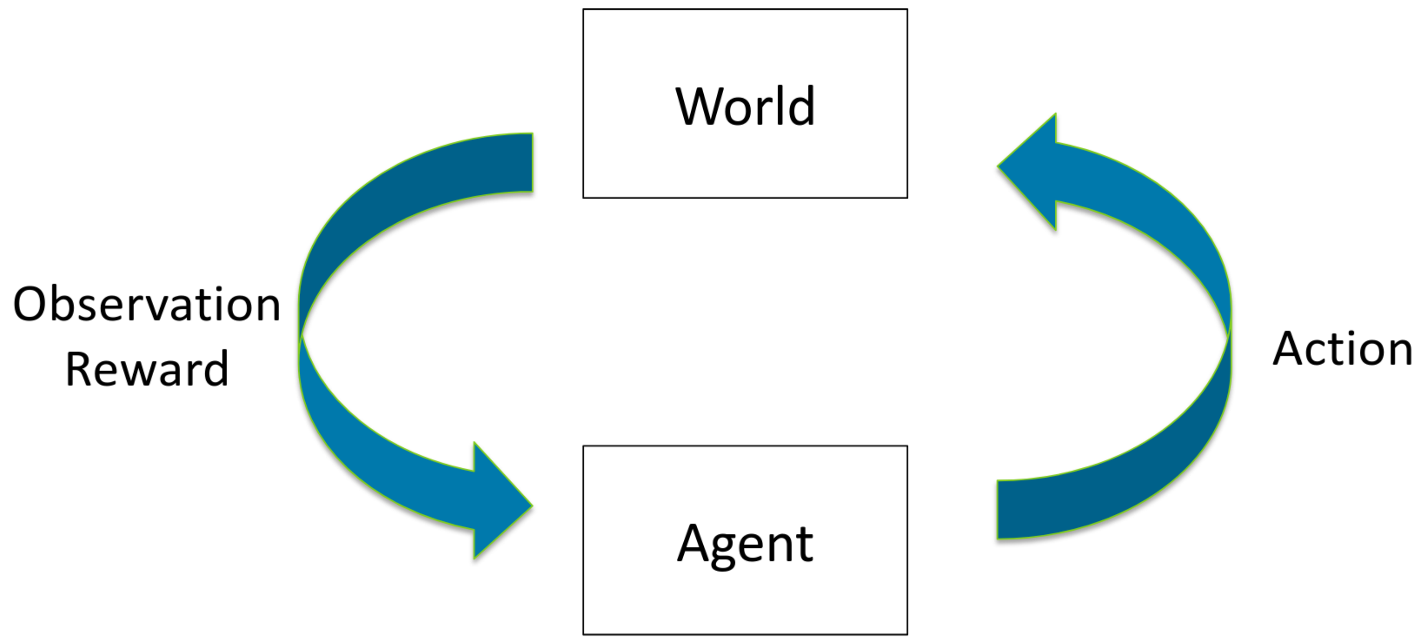
1) Pure Exploration

2) Deployment

Goal of Reinforcement Learning:

Cumulate as much ‘reward’ as possible while interacting with the system…

…while learning how the world works!



Learning while Deployed

Why is Exploration Hard?

Pure Exploitation: always play best known action / policy

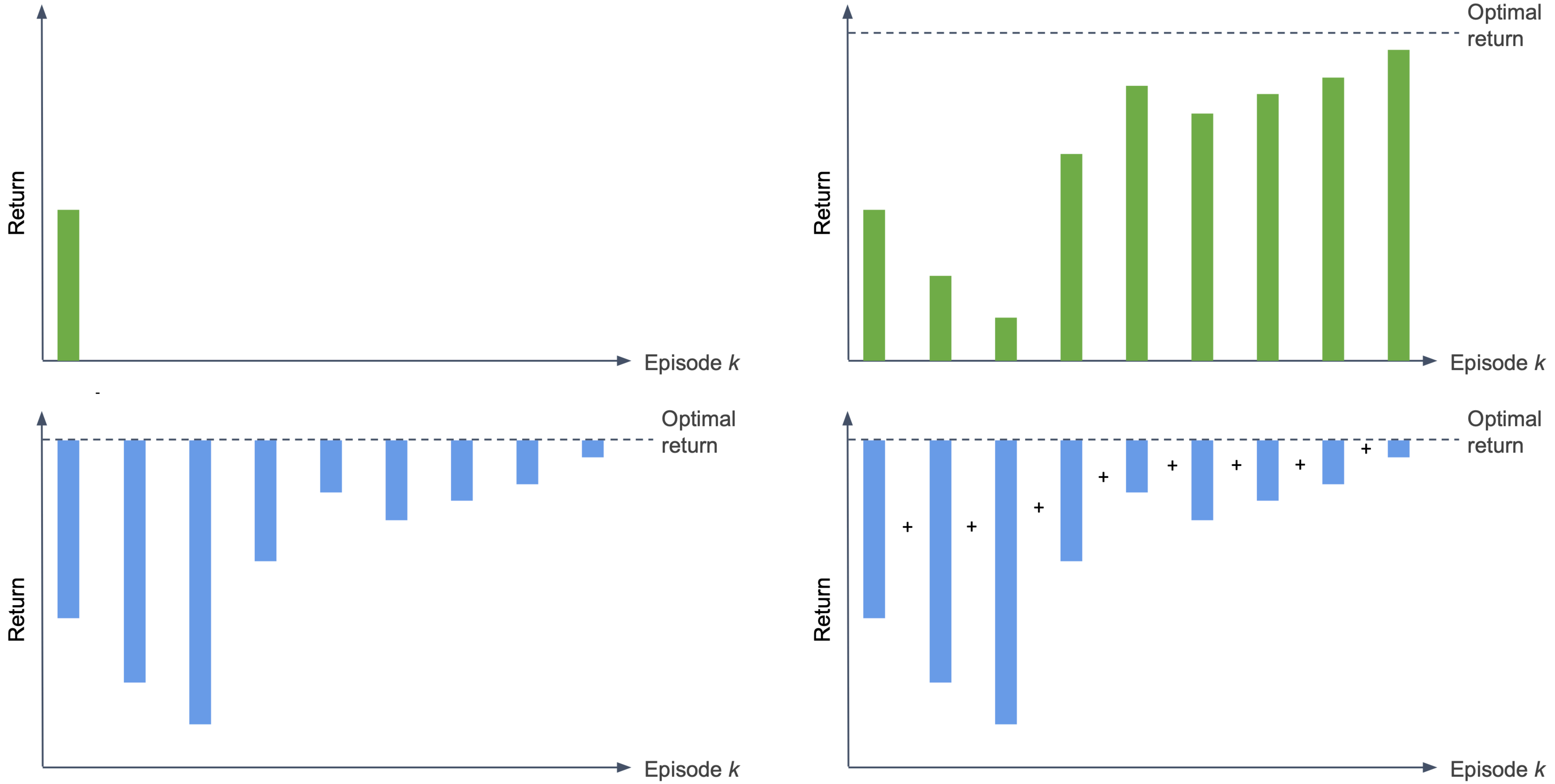
=> stuck in suboptimal polices forever

Pure Exploration: keep exploring indefinitely

=> never uses knowledge to accumulate reward

Need to balance exploration with exploitation

Performance Measure: Regret



Performance Measure: Regret

Regret: sum of losses compared to optimal policies

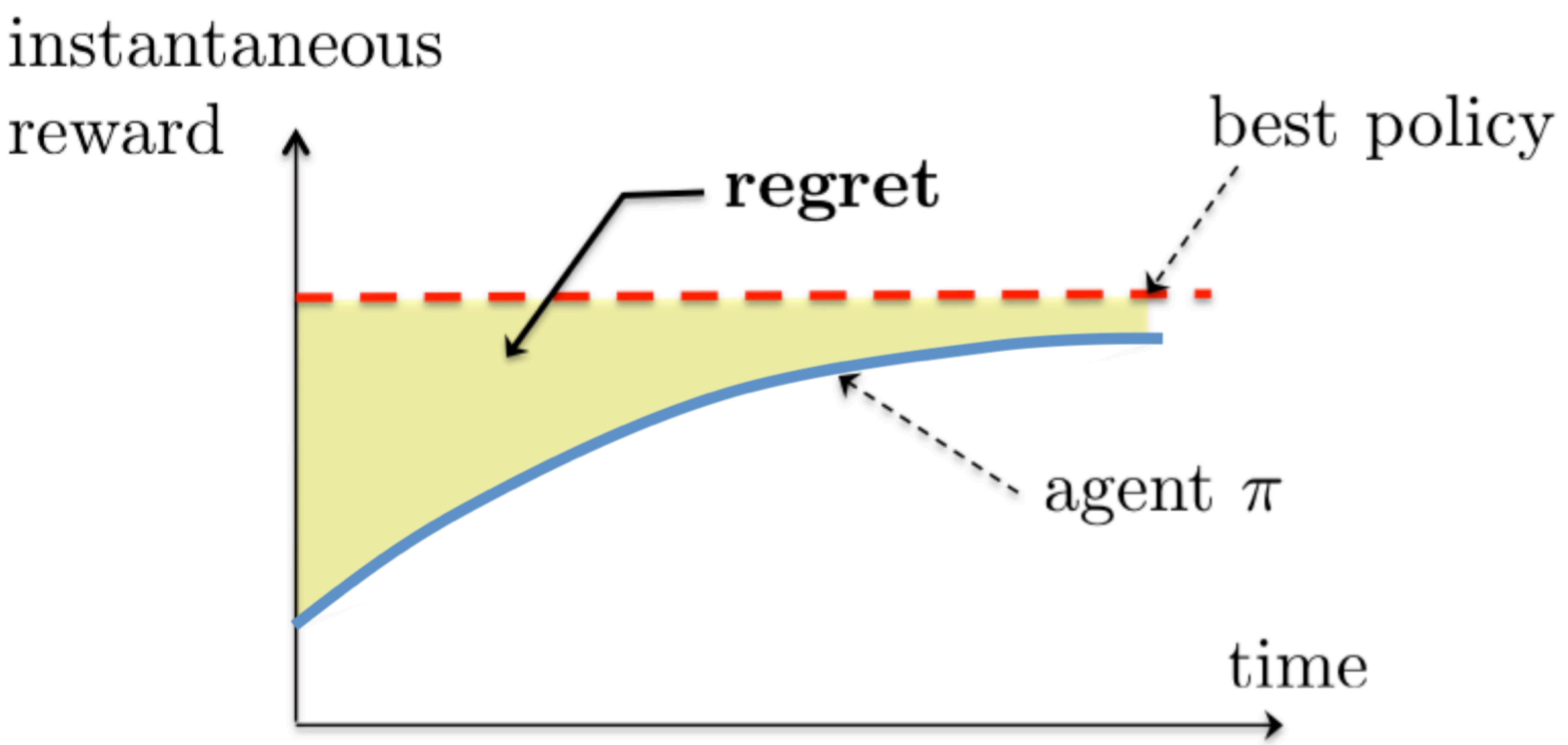
$$Regret(T) = T\mu^{\star} - \mathbb{E}_{\pi} \underbrace{\sum_{t=1}^{T} r_t}$$

Evaluation Time

Average Optimal Reward

Average Agent Cumulated Reward

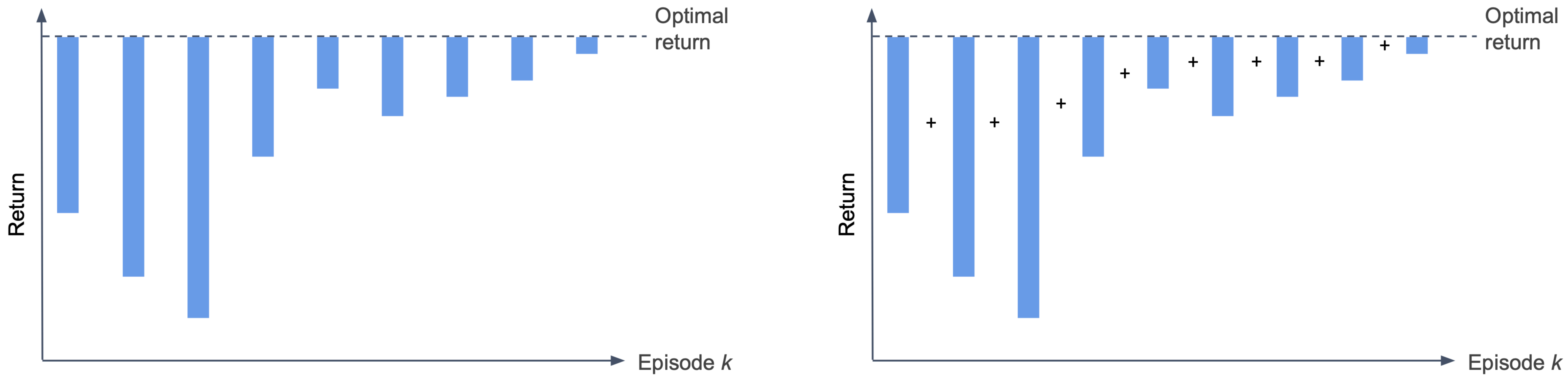
Remark: algorithm is being evaluated while learning

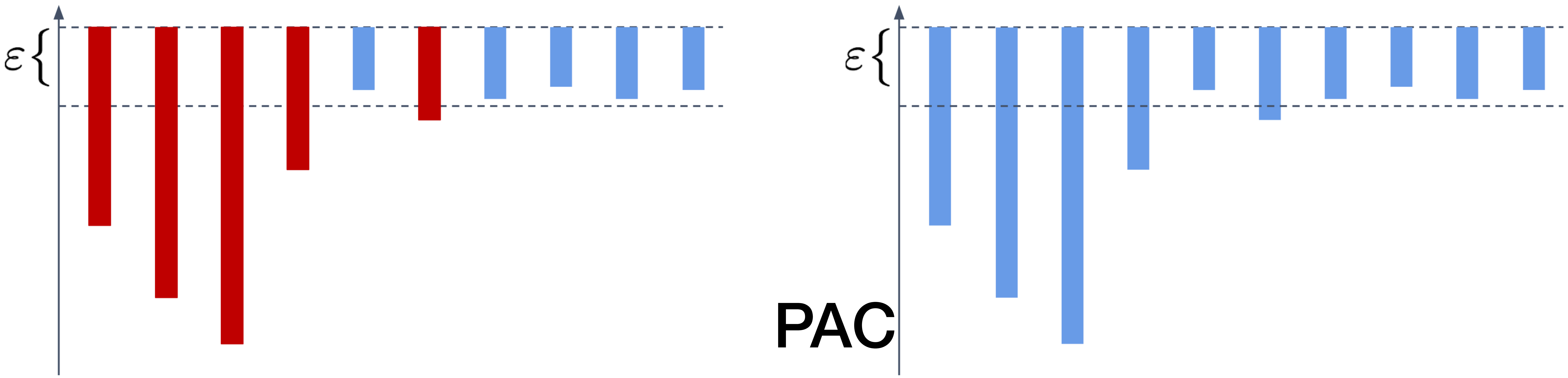


minimize Regret = maximize sum of Rewards

$$\min_{\pi} \left(Regret(T) \right) = \max_{\pi} \left(\mathbb{E}_{\pi} \sum_{t=1}^{T} r_t \right)$$

Regret


Optimal return
Return
Episode k
Optimal return
Return
+
Episode k


ε
ε

PAC

Ex I: Union Bound

2 Best Arm Identification in Multiarmed Bandit (35pts)

In this problem we focus on the Bandit setting with rewards bounded in $[0,1]$. A Bandit problem instance is defined as an MDP with just one state and action set $\mathcal{A}$. Since there is only one state, a "policy" consists of the choice of a single action: there are exactly $A = |\mathcal{A}|$ different deterministic policies. Your goal is to design a simple algorithm to identify a near-optimal arm with high probability.

Imagine we have n samples of a random variable x, $\{x_1, \dots, x_n\}$. We recall Hoeffding's inequality below, where $\overline{x}$ is the expected value of a random variable x, $\widehat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean (under the assumption that the random variables are in the interval [0,1]), n is the number of samples and $\delta > 0$ is a scalar:

$$\Pr\left(|\widehat{x} - \overline{x}| > \sqrt{\frac{\log(2/\delta)}{2n}}\right) < \delta.$$

Assuming that the rewards are bounded in $[0,1]$, we propose this simple strategy: allocate an identical number of samples $n_1 = n_2 = ... = n_A = n_{des}$ to every action, compute the average reward (empirical payout) of each arm $\widehat{r}_{a_1}, \dots, \widehat{r}_{a_A}$ and return the action with the highest empirical payout $\arg\max_a \widehat{r}_a$. The purpose of this exercise is to study the number of samples required to output an arm that is at least ϵ-optimal with high probability. Intuitively, as n_{des} increases the empirical payout $\widehat{r}_a$ converges to its expected value $\overline{r}_a$ for every action a, and so choosing the arm with the highest empirical payout $\widehat{r}_a$ corresponds to approximately choosing the arm with the highest expected payout $\overline{r}_a$.

(a) (15 pts) We start by defining a *good event*. Under this *good event*, the empirical payout of each arm is not too far from its expected value. Starting from Hoeffding inequality with n_{des} samples allocated to every action show that:

$$\Pr\left(\exists a \in \mathcal{A} \quad s.t. \quad |\widehat{r}_a - \overline{r}_a| > \sqrt{\frac{\log(2/\delta)}{2n_{des}}}\right) < A\delta.$$

In other words, the *bad event* is that at least one arm has an empirical mean that differs significantly from its expected value and this has probability at most $A\delta$.

More interesting algorithm: Identify near optimal arm with random stopping time

(a) (15 pts) We start by defining a *good event*. Under this *good event*, the empirical payout of each arm is not too far from its expected value *at a random stopping time* T. Starting from Hoeffding inequality with n_{des} samples allocated to every action *find* x such that:

$$\Pr\left(\exists a \in \mathcal{A} \quad s.t. \quad |\widehat{r}_a - \overline{r}_a| > \sqrt{\frac{\log(2x/\delta)}{2n_{des}}}\right) < \delta.$$

for the random stopping time n_{des}.

Solution

$$\Pr\left(\exists a \in \mathscr{A} \quad s.t. \quad |\hat{r}_a - \bar{r}_a| > \sqrt{\frac{\log(2x/\delta)}{2n_{des}}}\right) \leq \Pr\left(\exists a \in \mathscr{A}, \exists n \quad s.t. \quad |\hat{r}_a - \bar{r}_a| > \sqrt{\frac{\log(2x/\delta)}{2n}}\right)$$

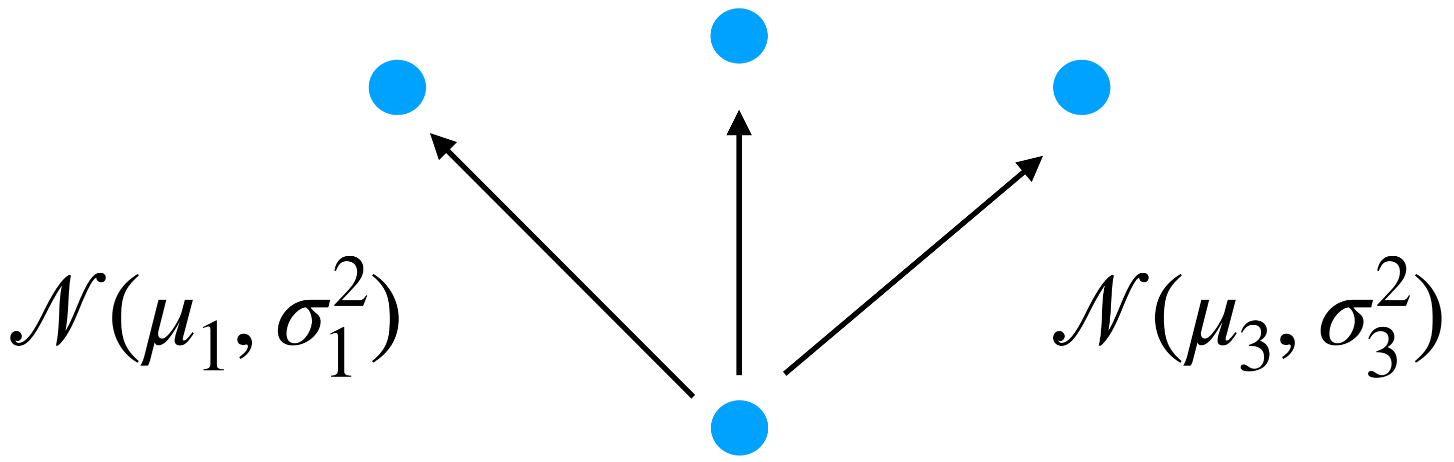
$$\leq \Pr\left(\bigcup_{a \in \mathscr{A}} \bigcup_{n} \quad s.t. \quad |\hat{r}_a - \bar{r}_a| > \sqrt{\frac{\log(2x/\delta)}{2n}}\right)$$

$$\leq \sum_{a \in \mathscr{A}} \sum_{n=1}^{\infty} \Pr\left(\quad |\hat{r}_a - \bar{r}_a| > \sqrt{\frac{\log(2x/\delta)}{2n}}\right) \leq \sum_{a \in \mathscr{A}} \sum_{n=1}^{\infty} \frac{\delta}{x} \leq \sum_{a \in \mathscr{A}} \sum_{n=1}^{\infty} \frac{\delta}{cAn^2} = \frac{\pi^2}{6}\frac{1}{c}\delta \leq \delta .$$

Posterior Sampling

Initialize prior over each arm a, $p(\mathcal{R}_a)$
loop
    For each arm a **sample** a reward distribution $\mathcal{R}_a$ from posterior
    Compute action-value function $Q(a) = \mathbb{E}[\mathcal{R}_a]$
    $a_t = \arg\max_{a \in \mathcal{A}} Q(a)$
    Observe reward r
    Update posterior $p(\mathcal{R}_a | r)$ using Bayes law
end loop

Example II: Posterior Sampling



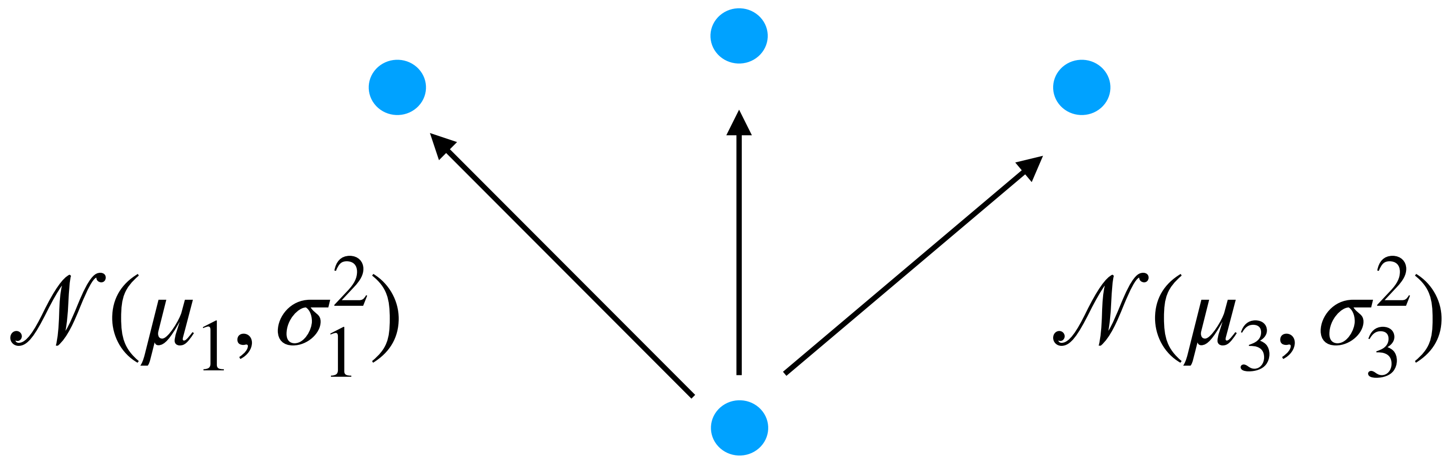
$$\sigma_1 = \sigma_2 = \ldots = \sigma$$

Assumption: Known Variance

Assume $x \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$ *and* $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. *Then:*

$$\mu \mid x \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}\, x \;+\; \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\, \mu_0 \;,\; \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\right)^{-1}\right)$$

Example II: Posterior Sampling



Can compute the posterior in closed form in few cases only

Normal-gamma distribution

From Wikipedia, the free encyclopedia

In probability theory and statistics, the **normal-gamma distribution** (or **Gaussian-gamma distribution**) is a bivariate four-parameter family of continuous probability distributions. It is the conjugate prior of a normal distribution with unknown mean and precision.[2]

normal-gamma

Parameters	μ location (real) $\lambda > 0$ (real) $\alpha > 0$ (real) $\beta > 0$ (real)
Support	$x \in (-\infty, \infty), \ \tau \in (0, \infty)$
PDF	$f(x, \tau \mid \mu, \lambda, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}} \tau^{\alpha - \frac{1}{2}} e^{-\beta\tau}$
Mean	[1] $\operatorname{E}(X) = \mu, \quad \operatorname{E}(\mathrm{T}) = \alpha\beta^{-1}$
Mode	$\left(\mu, \frac{\alpha - \frac{1}{2}}{\beta}\right)$
Variance	[1] $\operatorname{var}(X) = \left(\frac{\beta}{\lambda(\alpha - 1)}\right), \quad \operatorname{var}(\mathrm{T}) = \alpha\beta$

Contents [hide]

Definition [edit]

For a pair of random variables, (X,T), suppose that the conditional distribution of X given T is given by

$$X \mid T \sim N(\mu, 1/(\lambda T)),$$

meaning that the conditional distribution is a normal distribution with mean μ and precision λT — equivalently, with variance $1/(\lambda T)$.

Suppose also that the marginal distribution of T is given by

$$T \mid \alpha, \beta \sim \mathrm{Gamma}(\alpha, \beta),$$

where this means that T has a gamma distribution. Here λ, α and β are parameters of the joint distribution.

Then (X,T) has a normal-gamma distribution, and this is denoted by