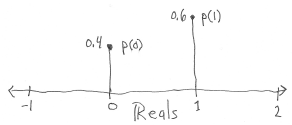
Random Variables, Expectations, Estimates, and a Learning Rule



Let $R \in \{0,1\}$ $p(0) = 0.4$ $p(1) = 0.6$ $p(r) = Pr[R=r]$

In general, consider a r.v. $R \in \mathcal{R} \subset \mathbb{R}eals$ $|\mathcal{R}| < \infty$

$R \sim p$ $p : \mathcal{R} \rightarrow [0,1]$ s.t. $\sum_{r \in \mathcal{R}} p(r) = 1$

The expectation, or <u>expected value</u> of R is

Defn: $\mathbb{E}[R] = \sum_{r \in \mathcal{R}} p(r) \cdot r$

Consider a sequence of r.v.s $R_t \sim p$ $t = 1, 2, 3, \ldots$
that are i.i.d. - identically, independently distributed

Consider the sequence of sample averages:

$$Q_{t+1} = \frac{R_1 + R_2 + \cdots + R_t}{t} \approx \mathbb{E}[R_t]$$

$$\lim_{t \to \infty} Q_t = \mathbb{E}[R_t]$$

$$Var[Q_t] = \mathbb{E}\left[(Q_t - \mathbb{E}[R_t])^2\right] \propto \frac{1}{\sqrt{t}}$$

Sample averages can be computed incrementally:

$$Q_{t+1} = Q_t + \frac{1}{t}[R_t - Q_t]$$

Which is a special case of our standard learning rule:

$$NewEst. = OldEst + StepSize \cdot [\underbrace{Target - OldEst}_{error}]$$

n-Armed Bandits

n actions $\in \{1, 2, 3, \ldots, n\} = \mathcal{A}$ $\quad |\mathcal{A}| = n$

A sequence of r.v.s $A_t \in \mathcal{A}$ $\quad A_t \sim \pi_t$ "policy" $\quad$ not iid

$t = 1, 2, 3, \ldots$ $\quad \pi_t : \mathcal{A} \to [0, 1]$, $\quad \sum_{a \in \mathcal{A}} \pi_t(a) = 1 \quad \forall a, t$

A sequence of r.v.s $R_t \in \mathcal{R} \subset \mathbb{R}$eals

$$p(r|a) = Pr\left[R_t = r \,\middle|\, A_t = a\right]$$

$$p : \mathcal{R} \times \mathcal{A} \longrightarrow [0, 1]$$

$$\sum_{r \in \mathcal{R}} p(r|a) = 1 \quad \forall a \in \mathcal{A}$$

The value of action a:

$$q_*(a) = E\left[R_t \,\middle|\, A_t = a\right] = \sum_{r \in \mathcal{R}} p(r|a) \cdot r$$

Estimates

$$Q_t(a) \approx q_*(a)$$

greedy policy: $A_t = \arg\max_a Q_t(a)$

OR:

$$\pi_t(a) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} Q_t(a') \\ 0 & \text{otherwise} \end{cases}$$

ϵ-greedy policy:

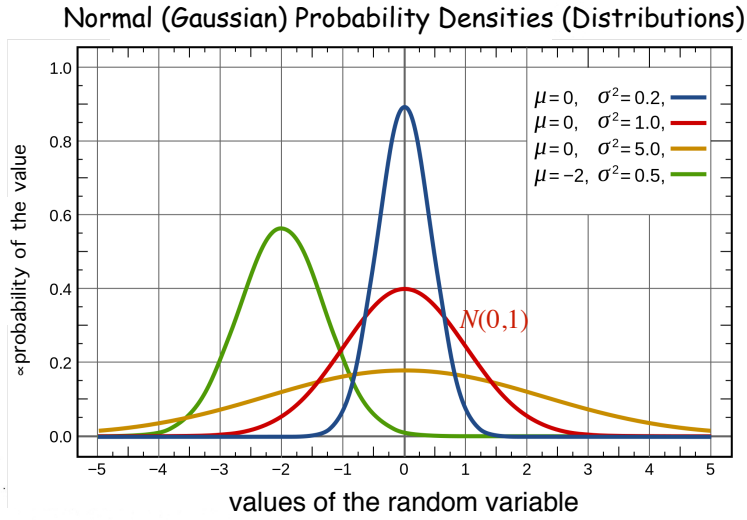
$$\pi_t(a) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{n} & \text{if } a = \arg\max_{a'} Q_t(a') \\ \frac{\epsilon}{n} & \text{otherwise} \end{cases}$$

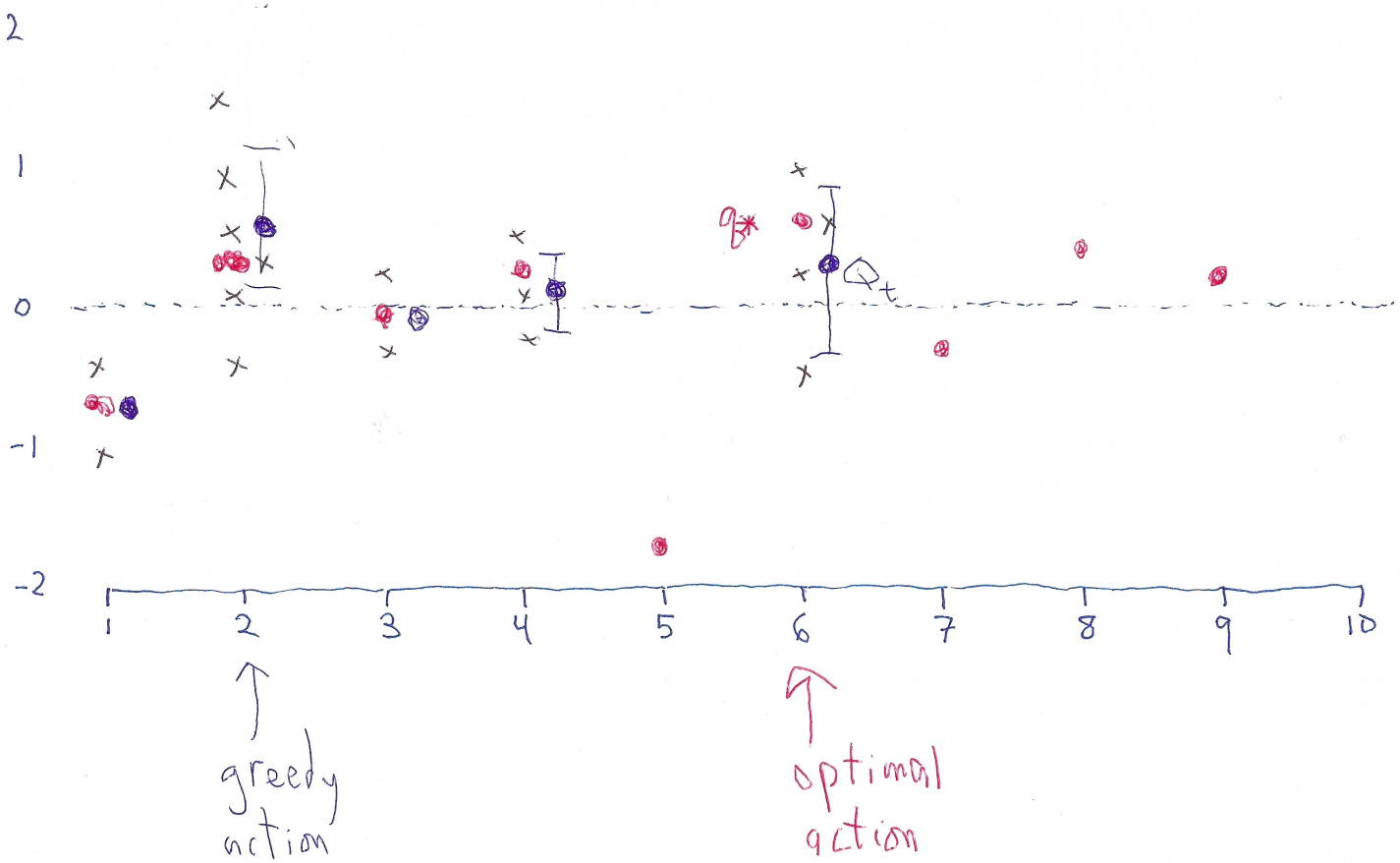
Ten-Armed Testbed

10 actions, $n = 10$

$q_*(a) \sim N(0, 1)$

mean, variance





$$Q_{t+1}(A_t) = Q_t(A_t) + \alpha\left[R_t - Q_t(A_t)\right]$$

and

$$Q_{t+1}(a) = Q_t(a) \quad \forall a \neq A_t \qquad a \in A$$

What you have learned from bandits

- ϵ-greedy policies
- the difference between a sample, an estimate, and a true expected value – R_t, Q_t, q_*
- the difference between the greedy action and the optimal action
- a learning rule; how learning can be seen as computing an average in an incremental way
- seen a complete example of goal-seeking – both the problem and the solution methods
- seen a complete example of mathematical formalization of an AI problem & solution