Learning and Acting

Learning is used to gain knowledge. In this course, we look into *machine* learning, so knowledge is a formal data structure. Where possible, we focus on knowledge representations that are *understandable* also to people, such as graph or rule-based representations. These are commonly termed *symbolic*.

Knowledge is used to act well. While human-understandability is only optional, a necessary condition for knowledge to be useful is that it should enable the machine (agent) or the human to perform good *actions* (=make good decisions). We first need to formalize what that means.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Actions

Let us use the following notation for any variable v and a number $k \in \mathbb{N}$:

$$v_{<k} = v_1, v_2, \ldots v_{k-1}$$

$$v_{\leq k} = v_1, v_2, \ldots v_k$$

Agent executes **actions** $y_k \in Y$ in discrete time $k \in \mathbb{N}$.

Y may be infinite but all $y \in Y$ must have a final description (i.e., Y is countable).

(Otherwise the agent could not communicate the chosen action in finite time.)

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Rewards

From its environment, it receives **rewards** $r_k \in R$, where R is a subset of a *bounded* real interval, i.e. $R = [a, b]$, $a, b \in \mathbb{R}$. Again, all $r \in R$ have to have a finite description (e.g., finite-precision decimal representation).

Reward at time k depends probabilistically on rewards up to time $k-1$, and actions up to time k:

$$r_k|_{y_{\leq k}, r_{<k}} \sim P(r_k \mid y_{\leq k}, r_{<k}) \tag{1}$$

Where the symbol $\sim$ means *'distributed as'*.

(Actions may influence rewards far into the future - consider a chess game with $R = \{0, 1\}$ and 1 indicating a win.)

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Actions and Rewards

The expression (1) describes the distribution of the random variable r_k assuming the history $y_{\leq k}, r_{<k}$. As it is clumsy, we adopt a simpler notation

$$r_k \overset{c}{\sim} P(r_k \mid y_{\leq k}, r_{<k}) \tag{2}$$

where $\overset{c}{\sim}$ means *'distributed as', assuming the conditions after the bar sign*.
This is a convenience notation only for this class, not used generally in literature!

The marginal probability of r_k is thus

$$r_k \sim P(r_k) = \sum_{y_{\leq k}, r_{<k}} P(r_k \mid y_{\leq k}, r_{<k})$$

(where the sum is over all pairs of sequences $y_{\leq k}, r_{<k}$) and the probability of a particular reward sequence $r_{\leq k}$ given action sequence $y_{\leq k}$ is

$$P(r_{\leq k} \mid y_{\leq k}) = P(r_1 \mid y_1) \cdot P(r_2 \mid r_1, y_{\leq 2}) \ldots P(r_m \mid r_{<m}, y_{\leq m})$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Example: Multi-Armed Bandit

Two-armed bandit

- k Game number (we play repeatedly)
- y_k One of two actions (pulling left or right lever)
- r_k Cash received minus cash thrown in at k. Depends on action at k but also on the history of actions of rewards. For example, after too much cash given out for the left-pull, the bandit lowers the mean reward for that action.

wiki

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Utility

The optimal action sequence $\overline{y}_1, \overline{y}_2, \ldots$ maximizes the **utility**, which is the expected (discounted) sum of rewards:

- for a *finite* time horizon $m \in \mathbb{N}$:

$$U^{y_{\leq m}} = \mathbb{E}\left(\sum_{k=1}^{m} r_k \,\middle|\, y_{\leq m}\right) = \sum_{r_{\leq m}} \left(P(r_{\leq m} \mid y_{\leq m}) \sum_{k=1}^{m} r_k\right) \tag{3}$$

- for an *infinite* horizon, we need to include a **discount** so that the sum converges. Typically, an exponential discount with base $0 < \gamma < 1$:

$$U^{y_{\leq \infty}} = \lim_{m\to\infty} \mathbb{E}\left(\sum_{k=1}^{m} r_k \gamma^{k-1} \,\middle|\, y_{\leq m}\right) = \lim_{m\to\infty} \sum_{r_{\leq m}} \left(P(r_{\leq m} \mid y_{\leq m}) \sum_{k=1}^{m} r_k \gamma^{k-1}\right) \tag{4}$$

Note: $\sum_{r_{\leq m}}$ *sums over all possible reward sequences* $r_{\leq m}$.

Exercises: finite, infinite

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Instant Rewards

In the general case, there is no obvious way to compute $\overline{y}_1, \overline{y}_2, \ldots$ without knowing the distribution $P(r_k \mid y_{\leq k}, r_{<k})$, e.g., without knowing the bandit's internals.

But consider the special case of **instant rewards**, which do not depend on anything before time k (a "memoryless" bandit). So instead of (2), we have

$$r_k \overset{c}{\sim} P(r_k \mid y_k) \tag{5}$$

Now utility ((3) or (4)) is maximized by constantly repeating the action

$$\overline{y} = \arg\max_y \mathbb{E}\left(r_k \mid y\right) \tag{6}$$

which is independent of k.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Exploration vs. Exploitation

Agent cannot follow (6) without knowing the distribution (5). It can however 'probe' the environment in time $k = 1, 2, \ldots K$ through random actions $y_1, y_2, \ldots, y_K \in Y$, collecting the rewards $r_1, r_2, \ldots, r_K$. K must be large enough so that each $x \in X$ occurs at least once in the action sequence.

For $k \geq K$, the agent switches from random actions to actions

$$y_{k+1} = \arg\max_y \widehat{\mathbb{E}}_k \left(r_{k+1} \mid y \right)$$

where the conditional expectation estimate $\widehat{\mathbb{E}}_k \left(r_{k+1} \mid y \right)$ is the average of all rewards obtained for action y from time 1 to time k.

The larger K, i.e. the longer the *exploration* period,

- the more accurate the estimate is
- the more time is spent behaving randomly, i.e. non-optimally

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Exploration vs. Exploitation (cont'd)

Dilemma: when to switch from exploration to *exploitation*?

Another option is to choose y_1 at random and then for $k \in \mathbb{N}$

- Make a random (exploration) action $y_{k+1} \in Y$ with probability $\epsilon > 0$.
- With prob. $(1 - \epsilon)$ use action that seems optimal:

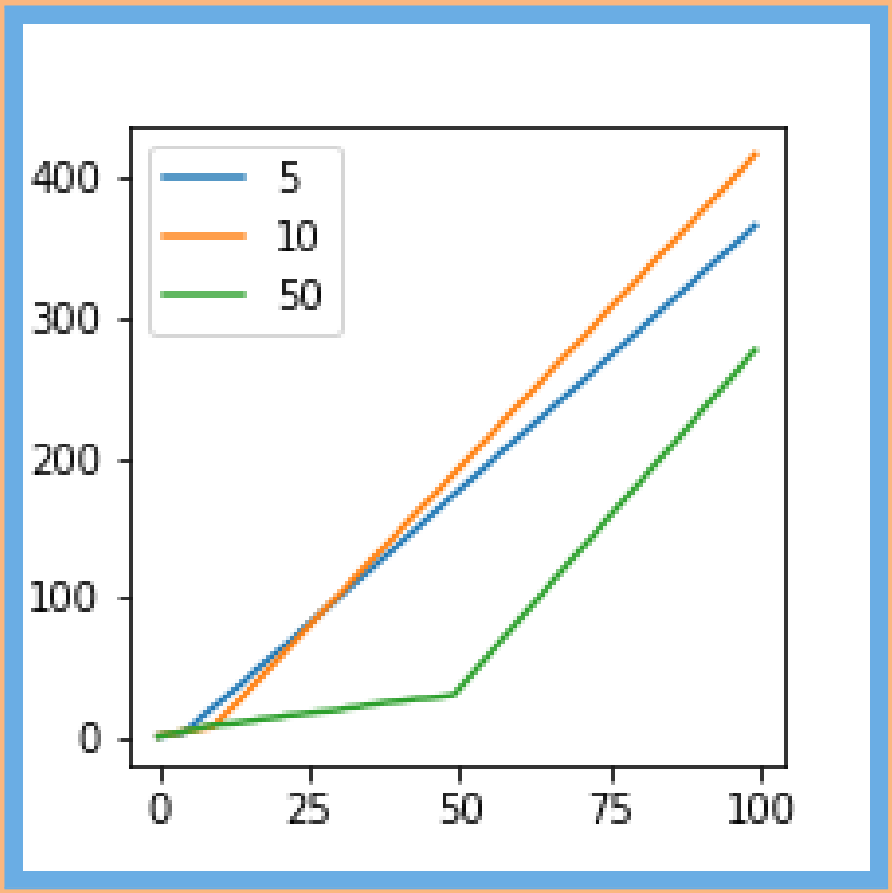
$$y_{k+1} = \arg\max_{y} \widehat{\mathbb{E}}_k \left(r_{k+1} \mid y \right)$$

Update the estimate $\widehat{\mathbb{E}}_k \left(. \mid . \right)$ at each k to the average of all rewards obtained for action y from time 1 to time k (use e.g. 0 if average undefined. i.e., y has not been used yet).

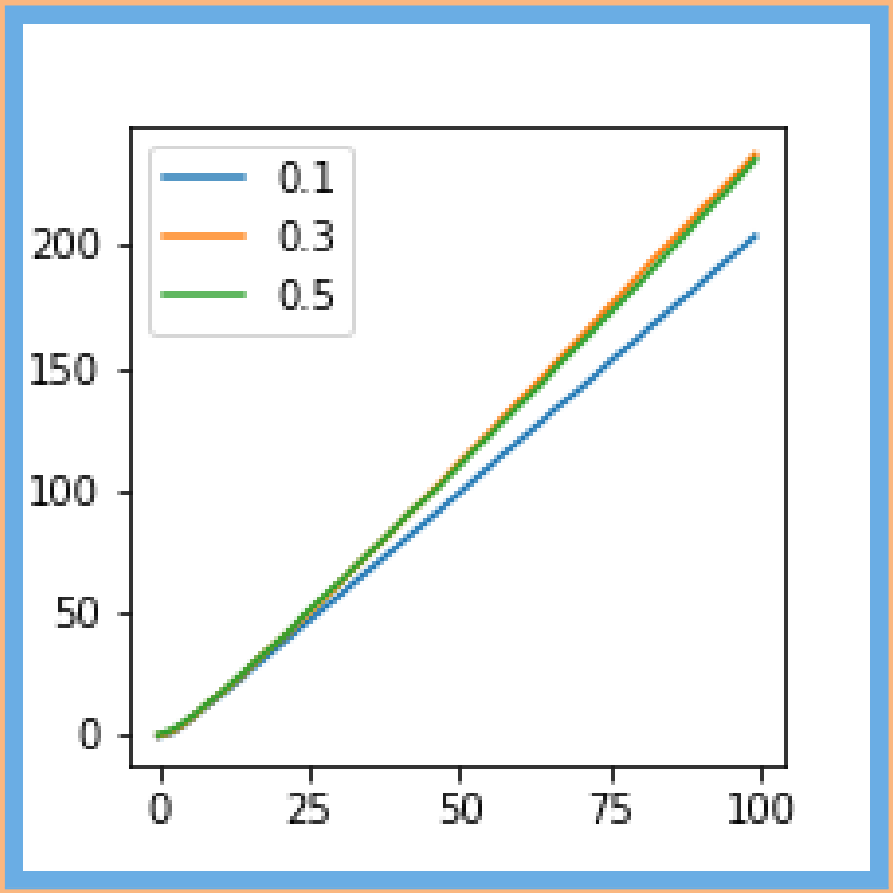
The dilemma is now how large ϵ could be. For an infinite horizon, an option is to let $\epsilon \to 0$ as $k \to \infty$ but keeping $\epsilon > 0$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Exploration vs. Exploitation: Example



Total reward for agent switching from exploration to exploitation after $K = 5, 10, 50$ steps.



Total reward for agent making exploration actions with probability $\epsilon = 0.1, 0.3, 0.5$ steps.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Observations

So far we have considered a 'blind' agent which does not observe anything about its environment, except for the rewards. Now we consider that at each k, it also receives an **observation** x_k from some set X. X may be infinite but its elements must have a finite representation.

The tuple (x_k, r_k) received by the agent is called **percept**, written xr_k for short. Analogically to (2), the percept generally depends on the entire history:

$$xr_k \overset{c}{\sim} P(xr_k \mid y_{\leq k}, xr_{<k}) \tag{7}$$

Despite the short notation xr_k, the above is a *joint* probability of x_k and r_k, which can be written using the probability chain rule as the product

$$P(x_k \mid r_k, xr_{<k}, y_{\leq k}) \tag{8}$$
$$\cdot P(r_k \mid xr_{<k}, y_{\leq k}) \tag{9}$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Example: Stock Market

k Day number

x_k Available market data on day k at closing time

y_k Investor's action (buy, sell, ..) on day k before closing time

r_k Cash earned or lost by investor on day k before closing time

$P(x_k \mid r_k, xr_{<k}, y_{\leq k})$ Current market data depend on the entire history of market data and investor's actions and rewards *(as well as other, unobserved factors - that is why it is a probability, not a function.).*

$P(r_k \mid xr_{<k}, y_{\leq k})$ Current reward depends on entire history of actions *(think a Bitcoin bought 5 years ago)*, market data except current x_k, and rewards.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Utility with Observations

Involving observations, the utility definitions (3) and (4) change.

- for a *finite* time horizon $m \in \mathbb{N}$, (3) turns into:

$$U^{y_{\leq m}} = \mathbb{E}\left(\sum_{k=1}^{m} r_k \;\middle|\; y_{\leq m}\right) = \sum_{xr_{\leq m}} \left(P(xr_{\leq m} \mid y_{\leq m}) \sum_{k=1}^{m} r_k\right) \tag{10}$$

- for an *infinite* time horizon, (4) turns into:

$$U^{y_1,y_2,\cdots} = \lim_{m\to\infty} \mathbb{E}\left(\sum_{k=1}^{m} r_k \gamma^{k-1} \;\middle|\; y_{\leq m}\right) = \lim_{m\to\infty} \sum_{xr_{\leq m}} \left(P(xr_{\leq m} \mid y_{\leq m}) \sum_{k=1}^{m} r_k \gamma^{k-1}\right) \tag{11}$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Instant Rewards with Observations

We now revisit the instant rewards assumption, this time with observations x_k.

With this assumption, the agent is rewarded only for how well it reacted to the *last seen observation* so for $k > 1$, (9) is replaced by

$$r_k \overset{c}{\sim} P(r_k \mid x_{k-1}, y_k)$$

i.e., we have for $k \in \mathbb{N}$:

$$\begin{aligned} r_1 &\overset{c}{\sim} P(r_1 \mid y_1) \\ r_{k+1} &\overset{c}{\sim} P(r_{k+1} \mid x_k, y_{k+1}) \end{aligned} \tag{12}$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Instant Rewards with Observations (cont'd)

The instant rewards assumption enabled to identify the optimal action (6) when observations were not part of the framework.

With observations, there is no longer an obvious way to compute optimal actions even if we know or can estimate the distribution (12). In particular,

$$y_{k+1} = \arg\max_{y \in Y} \mathbb{E}\left(r_{k+1} \mid x_k, y\right) \tag{13}$$

does *not* necessarily yield the optimal action y_{k+1}.

This is because we still assume in (8) that actions influence future observations. An effect of choosing y_{k+1} by (13) may be that the agent will receive 'worse' observations (allowing smaller rewards) in the future.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Example: Who Wants to Be a Millionaire?

k Question number

x_k Question

y_k Answer (one of a list of options) to question x_{k-1} *Formally, the k 'clock' ticks between a question and the subsequent answer.*

r_k Cash earned or lost

$P(x_k \mid r_k, xr_{<k}, y_{\leq k})$ Current question depends on history: correct answers cause more difficult questions to come. (Also: high rewards entail such questions to come.)

$P(r_k \mid x_{k-1}, y_k)$ Current reward depends only on the last question (in particular, the difficulty of it) and the answer to it (in particular, on its correctness w.r.t. the question).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

i.i.d. Observations

We will now consider one more assumption: observations are **i**ndependent of history and at each k they are sampled from the **i**dentical **d**istribution $P(x_k)$.

$$x_k \sim P(x_k) \tag{14}$$

This prevents the environment from 'playing tricks' on the agent. In the millionaire example, this would mean that questions are drawn randomly from a bucket and do not get progressively harder.

With the assumptions of instant rewards and i.i.d. observations, the optimal action is

$$\overline{y}_{k+1} = \arg\max_{y \in Y} \mathbb{E}\left(r_{k+1} \mid x_k, y\right) \tag{15}$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Policy

Without observations, the optimal action (6) was a constant. Here, the optimal action in (15) is a *function* of x_k. In general, a function

$$\pi : X \to Y \tag{16}$$

mapping an observation to an action is called a **policy**. The optimal policy will be denoted $\overline{\pi}(x)$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Sequential vs. Non-Sequential: Summary

Decision processes where rewards depend on history are called **sequential**.

No observations	
sequential	**non-sequential**
$r_k \overset{c}{\sim} P(r_k \mid y_{\leq k}, r_{<k})$ (2)	$r_k \overset{c}{\sim} P(r_k \mid y_k)$ (5)
	$\overline{y} = \arg\max_y \mathbb{E}(r_k \mid y)$ (6)

With observations	
sequential	**non-sequential**
$r_k \overset{c}{\sim} P(r_k \mid xr_{<k}, y_{\leq k})$ (9)	$r_1 \overset{c}{\sim} P(r_1 \mid y_1)$
	$r_{k+1} \overset{c}{\sim} P(r_{k+1} \mid x_k, y_{k+1})$ (12)
$x_k \overset{c}{\sim} P(x_k \mid r_k, xr_{<k}, y_{\leq k})$ (8)	$x_k \sim P(x_k)$ (14)
	$\overline{y}_{k+1} = \arg\max_y \mathbb{E}(r_{k+1} \mid x_k, y)$ (15)

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Scenarios in this course

In increasing generality:

1. *Concept and distribution learning from i.i.d. data*: non-sequential
2. *Online concept learning*: sequential (but additional assumptions)
3. *Reinforcement learning*: sequential (but additional assumptions)
4. *Universal AI*: sequential ("fully")

For didactic reasons, we will proceed in the 2, 1, 3, 4 order.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE