# 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 and Rewards

[label=l1] Agent executes **actions** $y_k \in Y$ in discrete time $k \in \mathbb{N}$. From its environment, it receives **rewards** $r_k \in R$, where $R$ is a (subset of a) *bounded* real interval, i.e. $R = [a, b]$, $a, b \in \mathbb{R}$. Actions may influence rewards far into the future (consider a chess game with $R = \{ 0, 1 \}$ and 1 indicating a win.)

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

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

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

Rewards depend probabilistically on all rewards and actions up to (and including) the current time:

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

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

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Actions and Rewards (cont'd)

[label=l1] 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

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

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

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

(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

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

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Example: Multi-Armed Bandit

[label=l1]
**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 $y_1^*, y_2^*, \ldots$ maximizes the **utility**, which is the expected (discounted) sum of rewards:

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

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

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

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

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

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Instant Rewards

In the general case, there is no obvious way to compute $y_1^*, y_2^*, \ldots$ without knowing the distribution $P(r_k|y_{\leq k}, r_{<k})$, 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 $r_k \overset{c}{\sim} P(r_k|y_k)$, or dropping the index:

$$r \overset{c}{\sim} P(r|y) \tag{5}$$

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

$$y^* = \arg\max_y \mathbb{E}\left(r|y\right) \tag{6}$$

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 $k = 1, 2, \ldots m$ through random actions $y_1, y_2, \ldots, y_m \in Y$, collecting the rewards $r_1, r_2, \ldots, r_m$. With no prior knowledge, $P(r|y)$ can then be estimated, i.e. *learned*, as

$$\widehat{P}(r|y) = \frac{|\{1 \leq k \leq m \mid r_k = r, y_k = y\}|}{|\{1 \leq k \leq m \mid y_k = y\}|}$$

The larger $m$, i.e. the longer the *exploration* period,

- the more accurate the estimate is, increasing chances to identify $y^*$
- the more time is spent behaving randomly, i.e. non-optimally

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

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Exploration vs. Exploitation (cont'd)

Another option:

- Make a random (exploration) action $y \in Y$ with probability $\epsilon$
- With prob. $(1 - \epsilon)$ use action that seems optimal:

$$y^* = \arg\max_y \widehat{\mathbb{E}}\,(r|y)$$

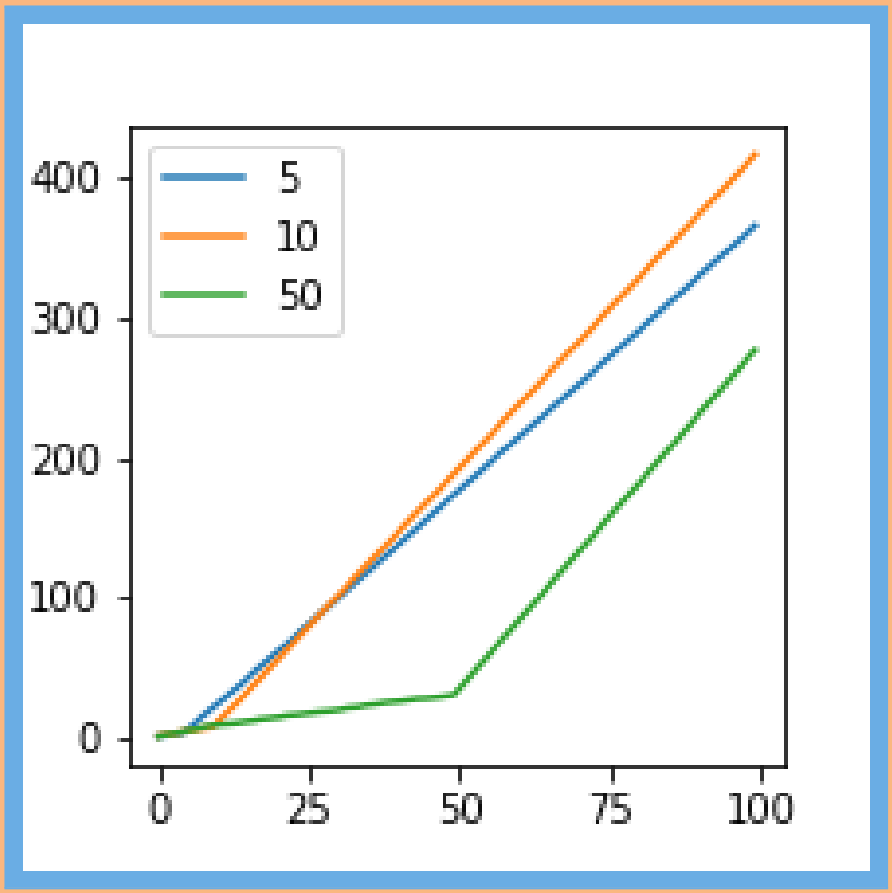
where $\widehat{\mathbb{E}}\,(r|y)$ is expectation with respect to $\widehat{P}(r|y)$.

Update $\widehat{P}$ at each $k$.

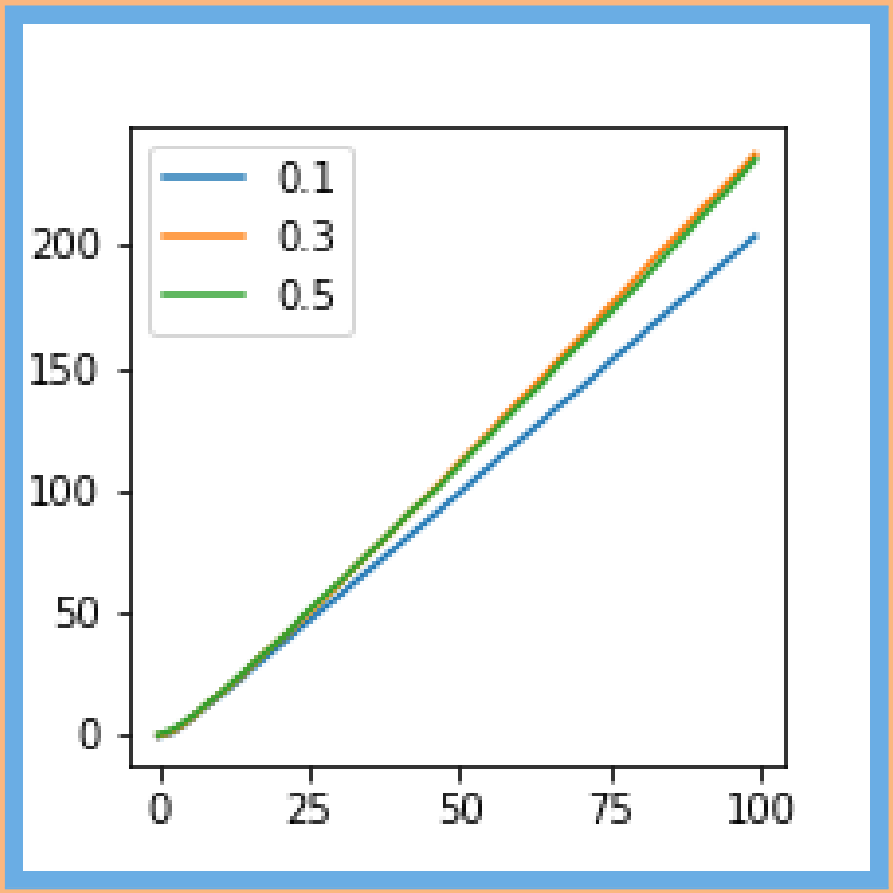
The dilemma is now how large $\epsilon$ could be. For an infinite horizon, an option is to let $\epsilon \to 0$ as $k \to \infty$ but keeping $\epsilon > 0$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Exploration vs. Exploitation: Example




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

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:

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

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

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

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

$P(x_k|r_k, xr_{<k}, y_{\leq k})$ 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.).*

$P(r_k|xr_{<k}, y_{\leq k})$ 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:

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

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

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

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 (9) is replaced by

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

Just as we did for (5), we may drop the indexes to simplify:

$$r \overset{c}{\sim} P(r | x, y) \tag{12}$$

but keeping in mind that $x$ *is one time-step behind* $y$ and $r$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Instant Rewards with Observations (cont'd)

The instant rewards assumption enabled to identify the optimal actions (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,

$$y^*(x) = \arg\max_y \mathbb{E}\left(r|x,y\right) \tag{13}$$

does *not* necessarily yield optimal actions.

This is because we still assume in (8) that actions influence future observations. An effect of choosing $y^*(x)$ 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

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

$P(r_k|x_{k-1}, y_k)$ 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 $P(r_k)$. As the distribution is the same for all $k$, we drop the index:

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

*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.*

Under the assumptions of instant rewards and i.i.d. observations, (13) prescribes the optimal (utility-maximizing) actions.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Policy

Unlike in (6), the optimal action $y^*(x)$ is a *function* of $x$ rather than a constant. In general, a function

$$y : X \to Y \tag{15}$$

mapping an observation to an action is called a **policy**.

*For simplicity, we use the same letter for the policy function $y(x)$ and the value $y$ it yields.*

# Sequential vs. Non-Sequential: Summary

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

| | **sequential** | **non-sequential** |
|---|---|---|
| **no observations** | $r_k \overset{c}{\sim} P(r_k \mid y_{\leq k}, r_{<k})$ (2) | $r \overset{c}{\sim} P(r \mid y)$ (5) |
| optimal action | ? | $y^* = \arg\max_y \mathbb{E}(r \mid y)$ (6) |
| **with observations** | $r_k \overset{c}{\sim} P(r_k \mid x r_{<k}, y_{\leq k})$ (9) | $r \overset{c}{\sim} P(r \mid x, y)$ (12) |
| | $x_k \overset{c}{\sim} P(x_k \mid r_k, x r_{<k}, y_{\leq k})$ (8) | $x \overset{c}{\sim} P(x)$ (14) |
| optimal policy | ? | $y^*(x) = \arg\max_y \mathbb{E}(r \mid x, y)$ (13) |

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 with instant rewards (12)
3. *Reinforcement learning*: sequential
4. *Universal AI*: sequential

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

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Classification

Non-sequential decision making where $Y$ is discrete (usually finite) is called **classification**.

*Note: when $Y = \mathbb{R}$, it is called 'regression', which we will not elaborate.*

$P(r|x, y)$ is usually defined by first defining a distribution $P(x, \bar{y})$ ($\bar{y} \in Y$) and a function $L : Y \times Y \to \mathbb{R}$ and letting

$$r(x) = -L(y(x), \bar{y}) \tag{16}$$

*Note that $r(x)$ is random even if $x$ is fixed, because $\bar{y}$ is random.*

$L$ is called a *loss function*.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Example: Classification of Handwritten Numbers

$X = \mathbb{R}^{16\times 16}$ ($x$ are pixel vectors), $Y = \{\, 0, \ldots, 9 \,\}$

Some suitable loss functions

$$L_{0,1}(y(x), \bar{y}) = \begin{cases} 0 \text{ if } y(x) = \bar{y} \\ 1 \text{ otherwise} \end{cases} \qquad L(y(x), \bar{y}) = |y(x) - \bar{y}|$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE