TÉCNICO LISBOA

Planning, Learning and Decision Making

Lecture 1. Reinforcement learning: Model-based methods

On to today’s stuff…

Stochastic approximation

Challenge 1

- Amy had the following grades in the first 3 labs+homework:

HW1	HW2	HW3
19.4	14.8	17.1

- What's her current lab grade?

$$(19.4 + 14.8 + 17.1) / 3 = 17.1$$

Great!

Challenge 1

- Her current lab average (after 3 labs) is 17.1
- Her fourth lab grade was 12.7
- What's her updated average?

$$(17.1 \times 3 + 12.7) / 4 = 16$$

Previous average corresponds to 3 grades

What is the average?

- Is the value *x* simultaneously closer to all samples:

$$\min \sum_{n=1}^{N} (x - x_n)^2$$

Derive and equate to 0

$$\sum_{n=1}^{N} x = \sum_{n=1}^{N} x_n$$

What is the average?

- Is the value *x* simultaneously closer to all samples:

$$\min \sum_{n=1}^{N} (x - x_n)^2$$

$$\Big\downarrow \text{ Derive and equate to 0}$$

$$x = \frac{1}{N} \sum_{n=1}^{N} x_n$$

What is the average?

- Is the value *x* simultaneously closer to all samples
- From the observed samples, it's the best prediction for the "next sample"

$$\bar{x}_N = \frac{1}{N} \sum_{n=1}^{N} x_n$$

How to recompute the average with a new sample?

New average

- If we observe a new sample x_{N+1}

$$\bar{x}_{N+1} = (\bar{x}_N \times N + x_{N+1})/(N+1)$$

Previous average corresponds to *N* samples

New average

- If we observe a new sample x_{N+1}

$$
\begin{aligned}
\bar{x}_{N+1} &= \frac{N}{N+1}\bar{x}_N + \frac{1}{N+1}x_{N+1} \\
&= \frac{N+1-1}{N+1}\bar{x}_N + \frac{1}{N+1}x_{N+1} \\
&= \left(1 - \frac{1}{N+1}\right)\bar{x}_N + \frac{1}{N+1}x_{N+1} \\
&= \bar{x}_N + \frac{1}{N+1}(x_{N+1} - \bar{x}_N)
\end{aligned}
$$

New average

- If we observe a new sample x_{N+1}

$$\bar{x}_{N+1} = \bar{x}_N + \frac{1}{N+1}\boxed{(x_{N+1} - \bar{x}_N)}$$

Error

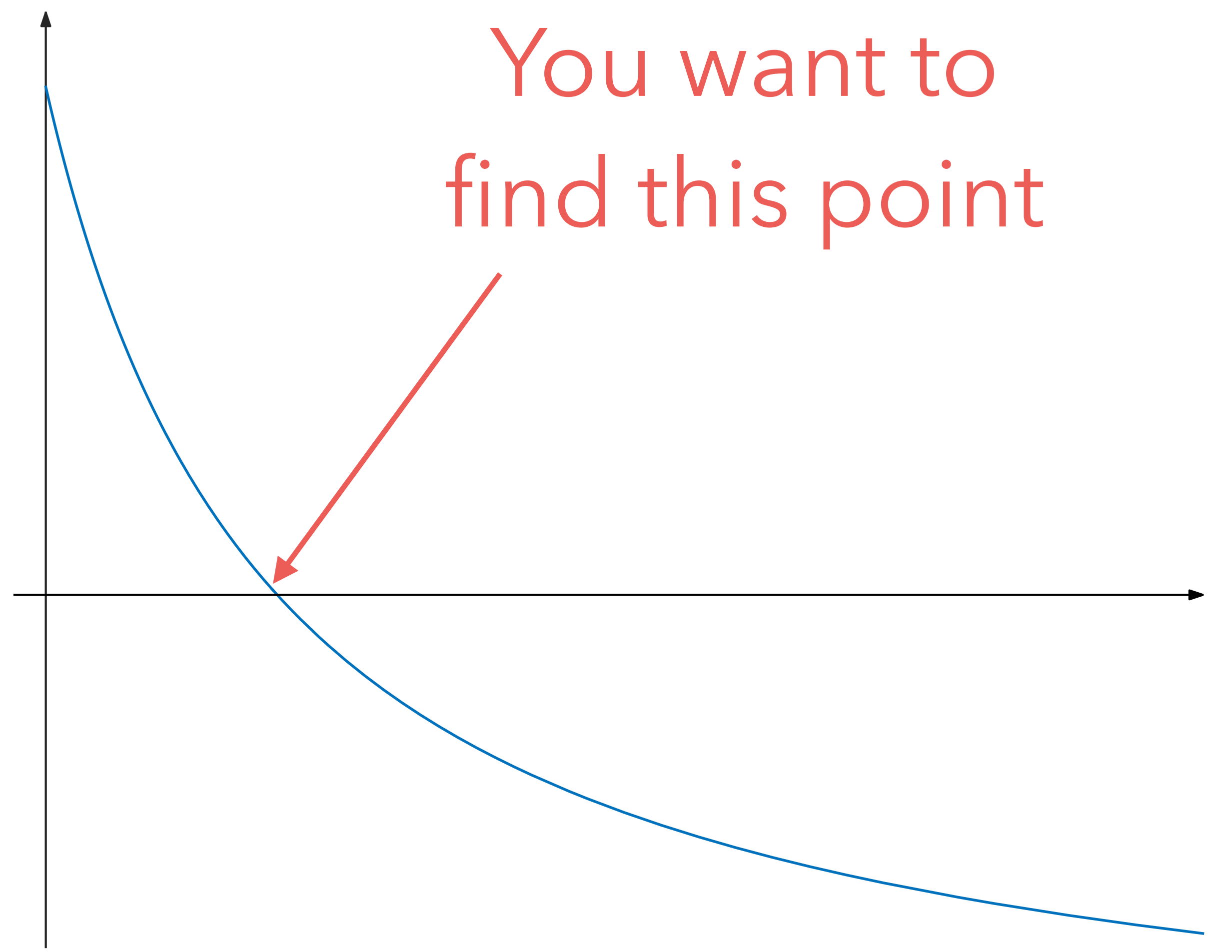
Step

New estimate

Previous estimate

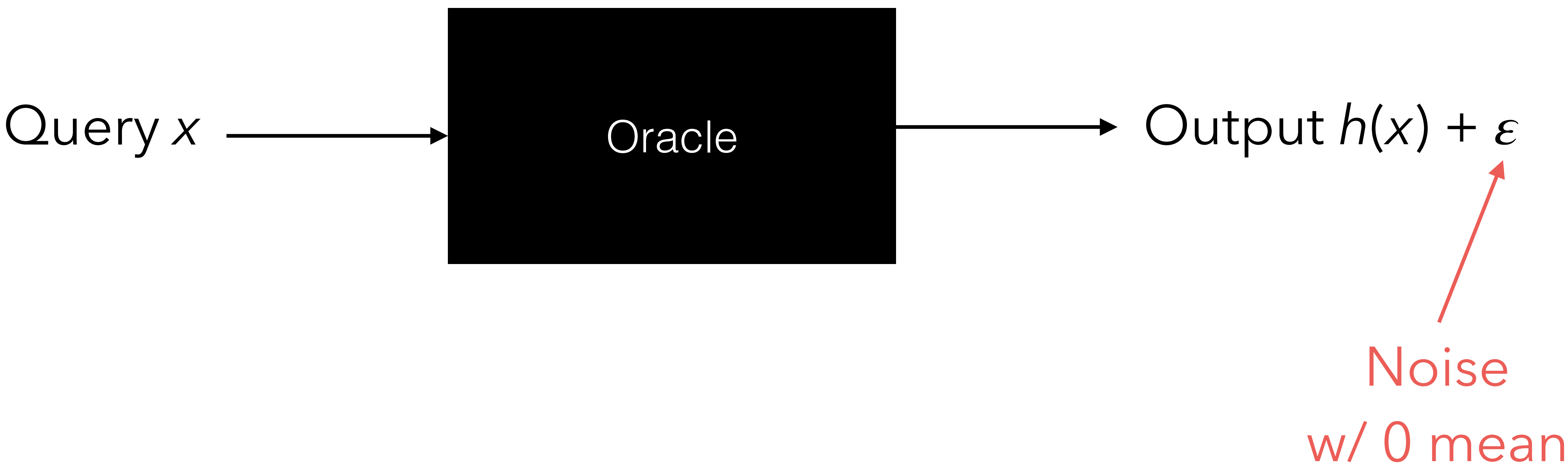
Challenge 2

- Consider the function:



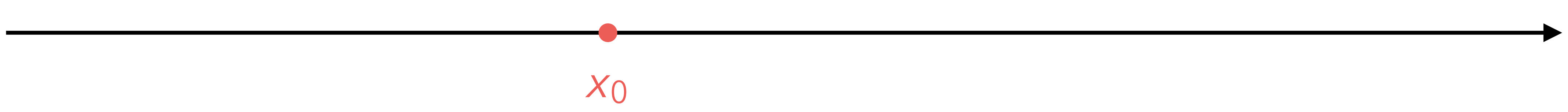
Challenge 2

- You can query a black box:

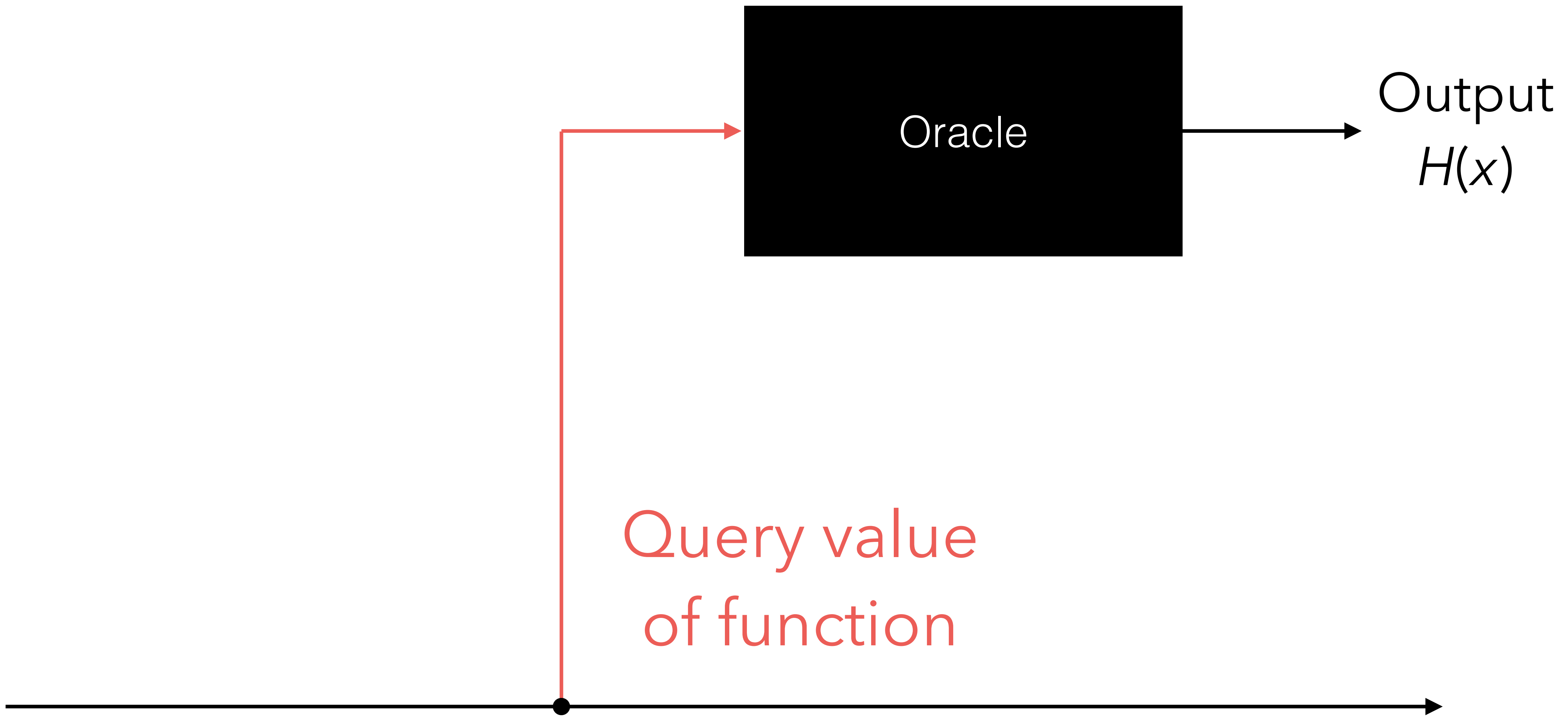


How do you solve this?

Idea



Idea



Idea

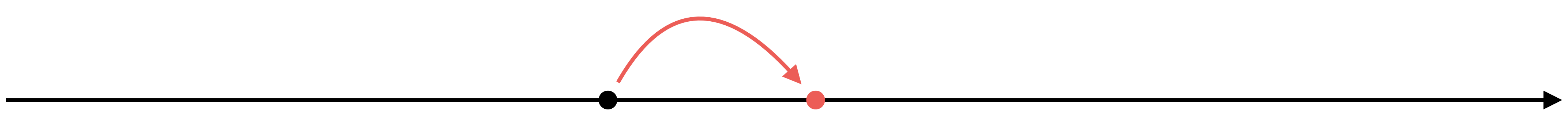
If $H(x) < 0$,
move back

Idea

If $H(x) \ll 0$,
move **far** back

Idea

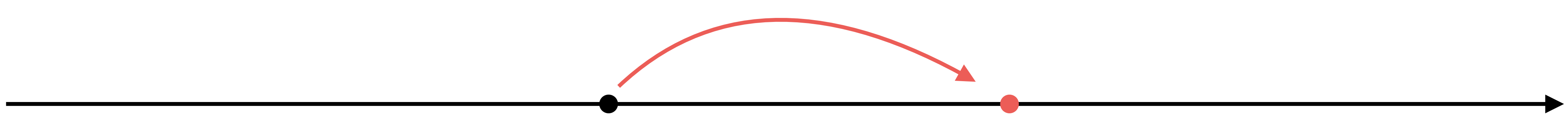
If $H(x) > 0$,
move forward



Idea

If $H(x) \gg 0$,

move **far** forward



Idea

- Compute the sequence

$$x_{n+1} = x_n + \alpha_n H(x_n)$$

Idea

- Compute the sequence

$$x_{n+1} = x_n + \alpha_n \boxed{(H(x_n) - 0)}$$

New estimate

Previous estimate

Step

Error

Stochastic approximation

- Iterative algorithms to compute the solution to the equation

$$\mathbb{E}\left[H(x)\right] = 0$$

where H is some function that can be queried

- Take the general form

$$\begin{aligned} x_{n+1} &= x_n + \alpha_n H(x_n) \\ &= x_n + \alpha_n h(x_n) + \alpha_n \boxed{(H(x_n) - h(x_n))} \end{aligned}$$

Zero-mean noise

Stochastic approximation

- Iterative algorithms to compute the solution to the equation

$$\mathbb{E}\left[H(x)\right] = 0$$

where *H* is some function that can be queried

- Take the general form

$$x_{n+1} = x_n + \alpha_n H(x_n)$$

- Example: Computing the mean

$$\bar{x}_{N+1} = \bar{x}_N + \frac{1}{N+1}(x_{N+1} - \bar{x}_N)$$

Stochastic approximation

- Iterative algorithms to compute the solution to the equation

$$\mathbb{E}\left[H(x)\right] = 0$$

where *H* is some function that can be queried

- Take the general form

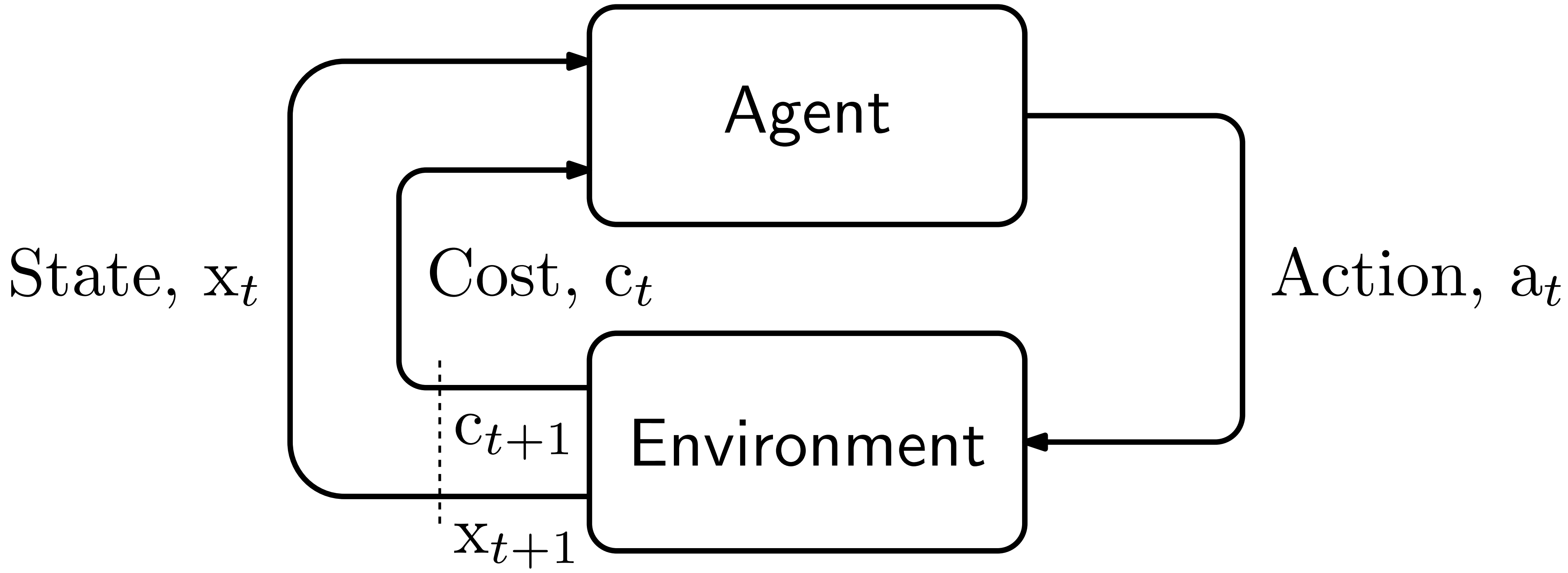
$$x_{n+1} = x_n + \alpha_n H(x_n)$$

- Example: Computing the mean

$$\bar{x}_{N+1} = \bar{x}_N + \alpha_n (x_{N+1} - \bar{x}_N)$$

Stochastic approximation

Markov decision process



Computing J^π

- We have that

$$J^\pi(x) = c_\pi(x) + \gamma \sum_{y \in \mathcal{X}} \mathrm{P}_\pi(y \mid x) J^\pi(y)$$

which is equivalent to

$$J^\pi(x) = \sum_{a \in \mathcal{A}} \pi(a \mid x) \left[\boxed{c(x, a)} + \gamma \sum_{y \in \mathcal{X}} \boxed{\mathrm{P}_a(y \mid x)} J^\pi(y) \right]$$

We must know the cost

We must know the transition probabilities

Computing Q*

- We have that

$$Q^*(x,a) = \boxed{c(x,a)} + \gamma \sum_{y \in \mathcal{X}} \boxed{\mathsf{P}_a(y \mid x)} \min_{a' \in \mathcal{A}} Q^*(y,a')$$

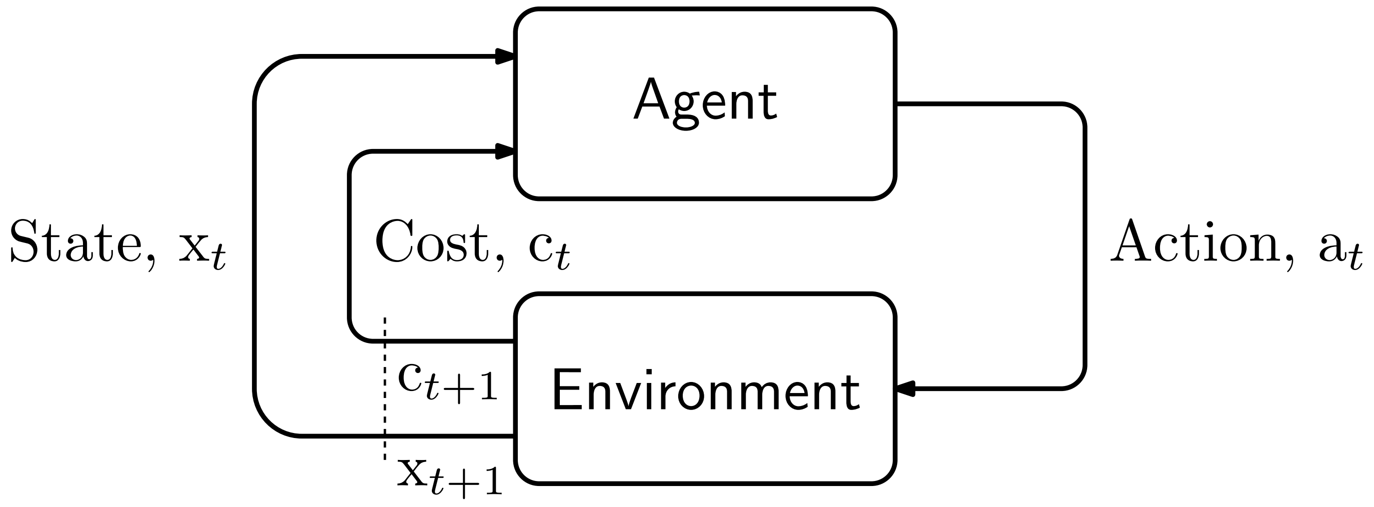
We must know the cost

We must know the transition probabilities

What if we don't?

Interactive learning

- We let the agent into the environment
- At each moment, the agent observes the state x_t
- The agent then selects an action a_t
- The agent observes the resulting cost c_t
- The process repeats



Interactive learning

- At each step, the agent collects a “data point”:

$$(x_t, a_t, c_t, x_{t+1})$$

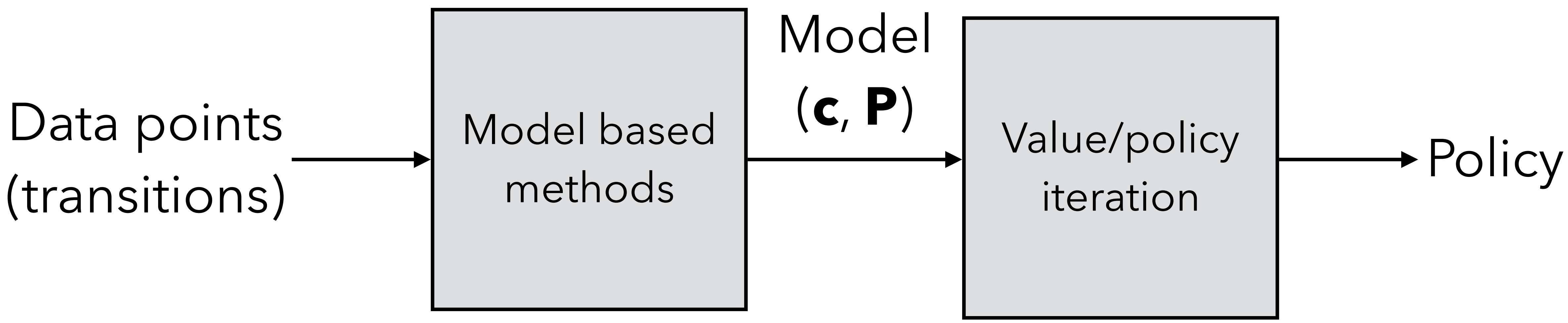
- Agent must compute the optimal policy by collecting many such data points
- The agent learns from “reward and punishment” (in the cost)
  - This form of learning is called **reinforcement learning**

TÉCNICO LISBOA

How?

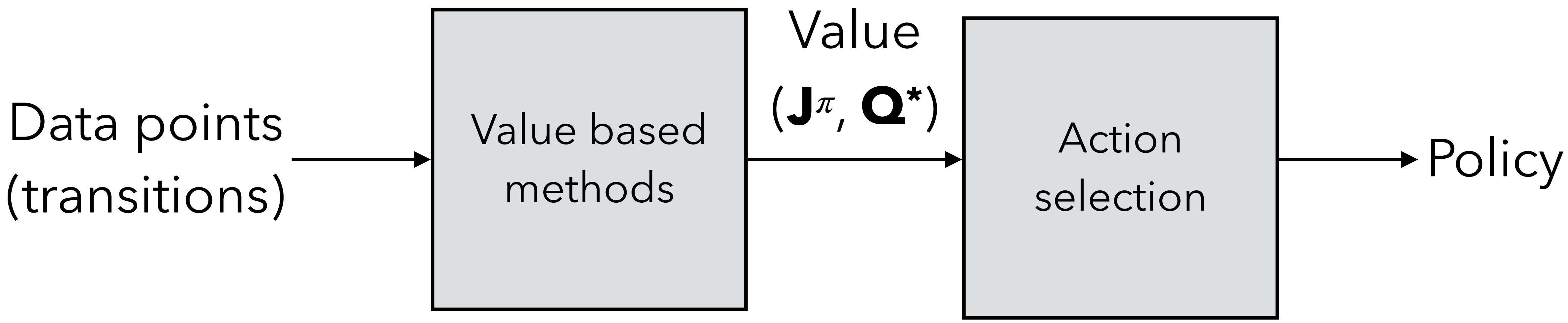
Three families of approaches

- Model-based methods:



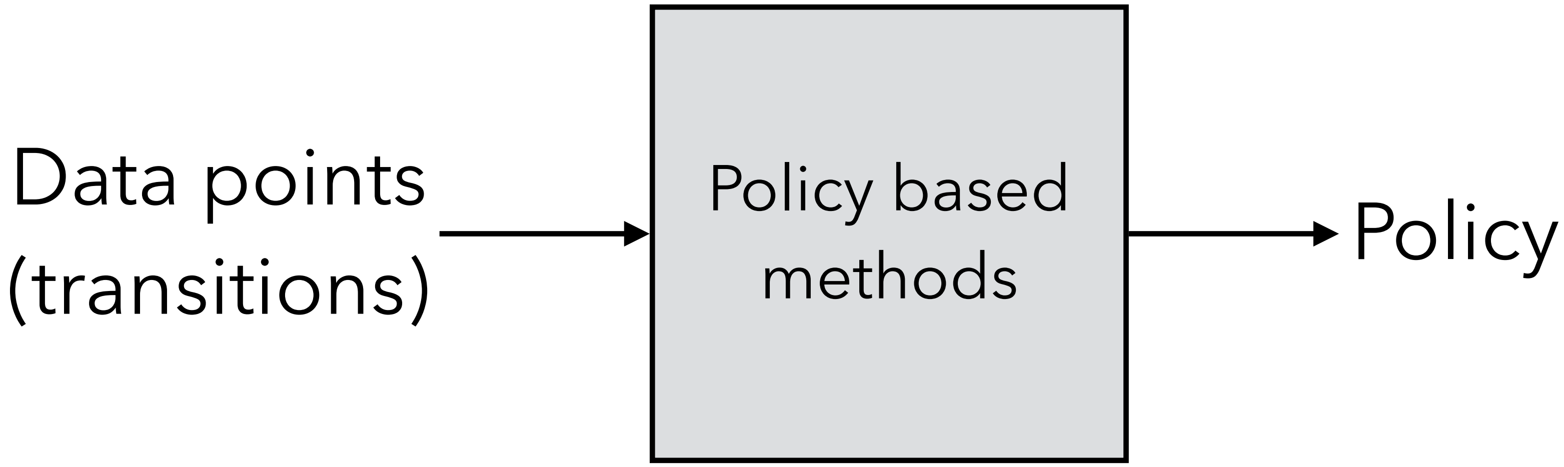
Three families of approaches

- Value-based methods:



Three families of approaches

- Policy-based methods:



Model-based methods

Estimating c

- We can estimate the cost c by keeping track of the observed costs at different states and actions
- At each step t, we just set

$$c(x_t, a_t) = c_t$$

What if there is noise in the costs?

Estimating c

- We can estimate the cost c by keeping track of the observed costs at different states and actions
- At each step *t*, we just set

$$\bar{c}_{t+1}(x_t, a_t) = \bar{c}_t(x_t, a_t) + \alpha_t(c_t - \bar{c}_t(x_t, a_t))$$

We just compute
the average!

Estimating P

- What about **P**?
- The transition probabilities can be seen as

$$\mathrm{P}(y \mid x, a) = \frac{N(x, a, y)}{N(x, a)}$$

N. of transitions from *x* to *y* after selecting *a*

N. of times *a* was selected in *x*

It's also an average!

Estimating P

- What about **P**?
- The transition probabilities can be seen as

$$\mathrm{P}(y \mid x, a) = \frac{1}{N(x,a)} \sum_{t=1}^{N} \mathbb{I}(\mathrm{x}_t = x, \mathrm{a}_t = a, \mathrm{x}_{t+1} = y)$$

It's also an average!

Estimating P

- We can estimate the transition probabilities **P** by keeping track of the how often we transition between states
- At each step *t*, we just set

$$\bar{\mathsf{P}}_{t+1}(y \mid x_t, a_t) = \bar{\mathsf{P}}_t(y \mid x_t, a_t) + \alpha(\mathbb{I}(\mathrm{x}_{t+1} = y) - \bar{\mathsf{P}}_t(y \mid x_t, a_t))$$

Use VI or PI with the model

- Once you have estimates for **P** and **c**
  - You can use VI to compute

$$J^\pi(x) = c_\pi(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_\pi(y \mid x) J^\pi(y)$$

or

$$Q^*(x, a) = c(x, a) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(y \mid x) \min_{a' \in \mathcal{A}} Q^*(y, a')$$

Use VI or PI with the model

- Once you have estimates for **P** and **c**
  - You can use PI to compute

$$\pi^*(x) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \left[c(x,a) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x, a) J^*(y) \right]$$

Does this work?

- We are computing averages:
  - We estimate $c(x, a)$ as an average (for each x and a)
  - We estimate $\mathbf{P}(\cdot \mid x, a)$ as an average (for each x and a)
- How many “data points” do we need for each x and a?
  - An infinite number!

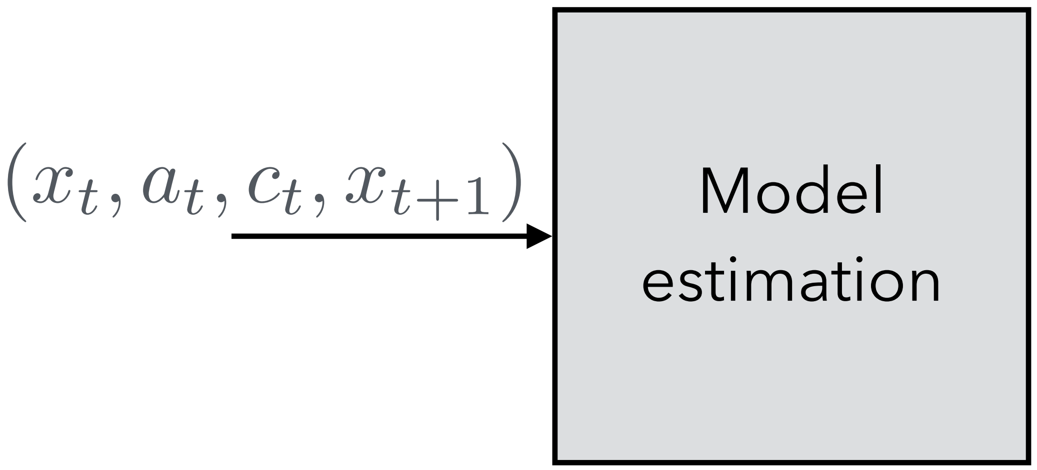
- The model-based approach described converges to the true parameters $\mathbf{P}$ and $\mathbf{c}$ as long as every state and action are visited infinitely often.

So when do we run VI?

Model based RL

- In practice, we interleave steps of model learning with steps of value/policy iteration

Model based RL



TÉCNICO LISBOA

Model based RL

Model estimation

$$\bar{c}_{t+1}(x_t, a_t) = \bar{c}_t(x_t, a_t) + \alpha_t(c_t - \bar{c}_t(x_t, a_t))$$

$$\bar{\mathrm{P}}_{t+1}(y \mid x_t, a_t) = \bar{\mathrm{P}}_t(y \mid x_t, a_t) + \alpha(\mathbb{I}(\mathrm{x}_{t+1} = y) - \bar{\mathrm{P}}_t(y \mid x_t, a_t))$$

Updated model

Model based RL

$$\bar{c}_{t+1}(x_t, a_t) = \bar{c}_t(x_t, a_t) + \alpha_t(c_t - \bar{c}_t(x_t, a_t))$$

$$\bar{\mathsf{P}}_{t+1}(y \mid x_t, a_t) = \bar{\mathsf{P}}_t(y \mid x_t, a_t) + \alpha(\mathbb{I}(\mathrm{x}_{t+1} = y) - \bar{\mathsf{P}}_t(y \mid x_t, a_t))$$

VI update

Model based RL

VI update

$$Q_{t+1}(x_t, a_t) = \bar{c}_{t+1}(x_t, a_t) + \gamma \sum_{y \in \mathcal{X}} \bar{\mathbf{P}}_{t+1}(y \mid x_t, a_t) \min_{a' \in \mathcal{A}} Q_t(y, a')$$

Updated Q-function

Model based RL

- Given a sample (x_t, c_t, x_{t+1}), where the action was selected from π,
- Compute

$$\bar{\mathsf{P}}_{t+1}(y \mid x_t) = \bar{\mathsf{P}}_t(y \mid x_t) + \alpha(\mathbb{I}(\mathrm{x}_{t+1} = y) - \bar{\mathsf{P}}_t(y \mid x_t))$$

$$\bar{c}_{t+1}(x_t) = \bar{c}_t(x_t) + \alpha_t(c_t - \bar{c}_t(x_t))$$

- Compute

$$J_{t+1}(x_t) = \bar{c}_{t+1}(x_t) + \gamma \sum_{y \in \mathcal{X}} \bar{\mathbf{P}}_{t+1}(y \mid x_t) J_t(y)$$

Update only
affected entries

Model based RL

- Given a sample (x_t, a_t, c_t, x_{t+1})
- Compute

$$\bar{\mathsf{P}}_{t+1}(y \mid x_t, a_t) = \bar{\mathsf{P}}_t(y \mid x_t, a_t) + \alpha(\mathbb{I}(\mathrm{x}_{t+1} = y) - \bar{\mathsf{P}}_t(y \mid x_t, a_t))$$

$$\bar{c}_{t+1}(x_t, a_t) = \bar{c}_t(x_t, a_t) + \alpha_t(c_t - \bar{c}_t(x_t, a_t))$$

- Compute

$$Q_{t+1}(x_t, a_t) = \bar{c}_{t+1}(x_t, a_t) + \gamma \sum_{y \in \mathcal{X}} \bar{\mathbf{P}}_{t+1}(y \mid x_t, a_t) \min_{a' \in \mathcal{A}} Q_t(y, a')$$

Update only
affected entries