AARHUS UNIVERSITY

Data Science 6

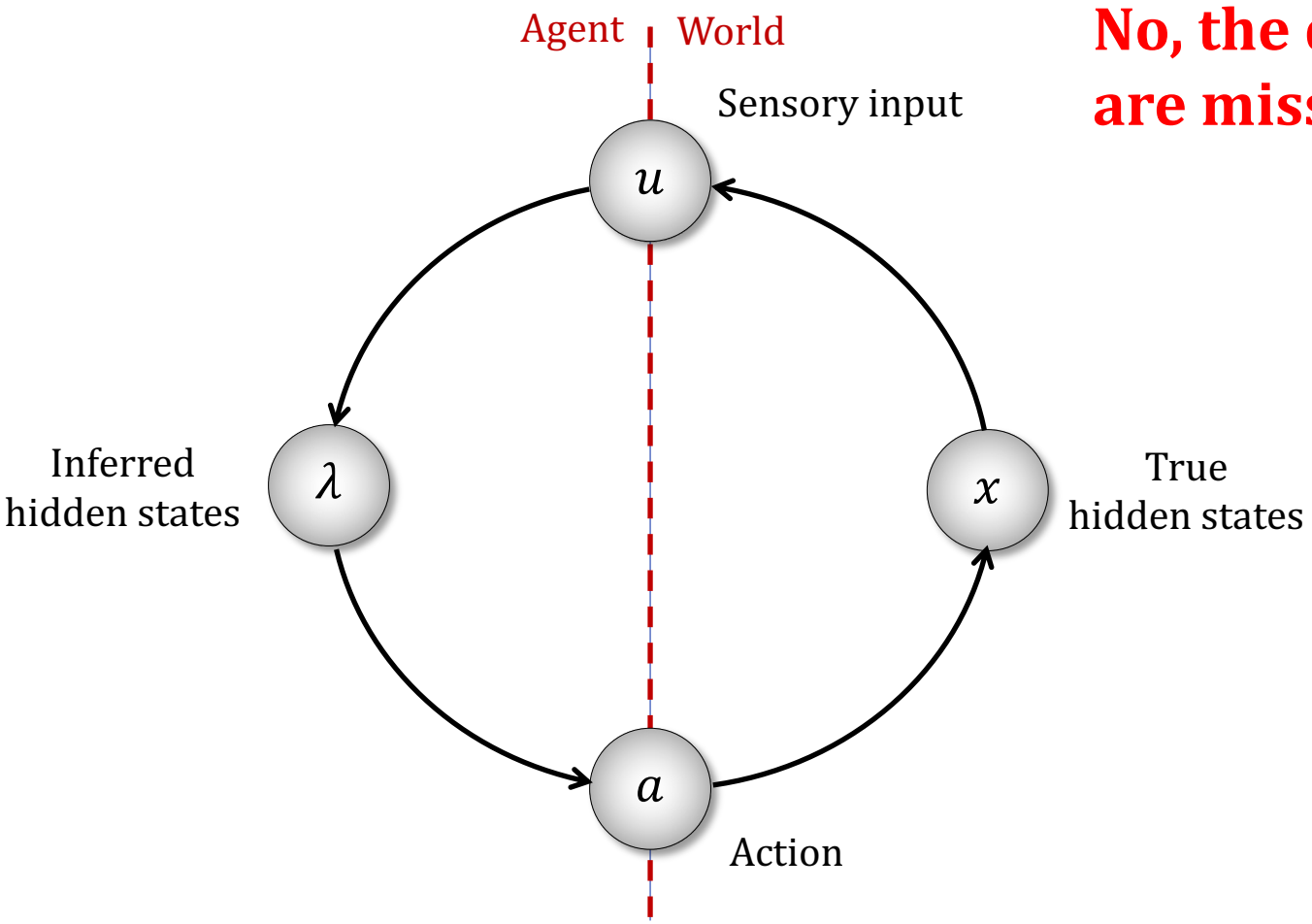
Chris Mathys

Master's Degree Programme in Cognitive Science

Spring 2023

Sequential inputs: learning rates, moving average, exponential smoothing, the Kalman filter

But: does inference as we've described it adequately describe the situation of actual biological agents?



No, the dynamics are missing!

What about dynamics?

Up to now, we've only looked at inference on static quantities, but biological agents live in a continually changing world.

In our example, the boat's position changes and with it the angle to the lighthouse.

How can we take into account that old information becomes obsolete? If we don't, our learning rate becomes smaller and smaller because our eqations were derived under the assumption that we're accumulating information about a stable quantity.

What's the simplest way to keep the learning rate from going too low?

Keep it constant!

So, taking the update equation for the mean of our observations as our point of departure...

$$\bar{u}_n = \bar{u}_{n-1} + \frac{1}{n}(u_n - \bar{u}_{n-1}),$$

... we simply replace $\frac{1}{n}$ with a constant α (and $\bar{u}$ with a generic value q):

$$q_n = q_{n-1} + \alpha(u_n - q_{n-1}).$$

This is called *Rescorla-Wagner learning* [although it wasn't this line of reasoning that led Rescorla & Wagner (1972) to their formulation].

How are we treating observations in Rescorla-Wagner learning?

Rewriting the RW learning rule reveals that q_n is an average weighted by α of the observation x_n and the previous value q_{n-1}

$$
\begin{aligned}
q_n &= q_{n-1} + \alpha(u_n - q_{n-1}) \\
&= (1-\alpha)q_{n-1} + \alpha u_n \\
&= (1-\alpha)\big((1-\alpha)q_{n-2} + \alpha u_{n-1}\big) + \alpha u_n \\
&= (1-\alpha)^2 q_{n-2} + (1-\alpha)\alpha u_{n-1} + \alpha u_n \\
&= (1-\alpha)^3 q_{n-3} + (1-\alpha)^2 \alpha u_{n-2} + (1-\alpha)\alpha u_{n-1} + \alpha u_n \\
&= (1-\alpha)^n q_0 + \alpha \sum_{i=1}^{n} (1-\alpha)^{n-i} u_i
\end{aligned}
$$

Recursively unpacking the content of q_n reveals that **observations u_i are exponentially discounted into the past.**

How are we treating observations in Rescorla-Wagner learning?

Taking $q_0 = 0$ and $\gamma \coloneqq 1 - \alpha$, we get

$$q_n = q_{n-1} + \alpha(u_n - q_{n-1})$$

$$= (1 - \alpha)q_{n-1} + \alpha u_n$$

$$= \alpha \sum_{i=1}^{n} (1 - \alpha)^{n-i} u_i$$

$$= (1 - \gamma) \sum_{i=1}^{n} \gamma^{n-i} u_i$$

$$= (1 - \gamma) \sum_{i=0}^{n-1} \gamma^{i} u_{n-i}$$

Does a constant learning rate solve our problems?

Partly: it implies a certain rate of forgetting because it amounts to taking only the $n = \frac{1}{\alpha}$ last data points into account. But...

... if the learning rate is supposed to reflect uncertainty in Bayesian inference, then how do we

(a) know that α reflects the right level of uncertainty at any one time, and

(b) account for changes in uncertainty if α is constant?

What we really need is an adaptive learning that accurately reflects uncertainty.

Needed: an adaptive learning rate that accurately reflects uncertainty

This requires us to think a bit more about what kinds of uncertainty we are dealing with.

A possible taxonomy of uncertainty is (cf. Yu & Dayan, 2003; Payzan-LeNestour & Bossaerts, 2011):

(a) **outcome uncertainty** that remains unaccounted for by the model, called *risk* by economists (π_ε in our Bayesian example); this uncertainty remains even when we know all parameters exactly,

(b) **informational** or *expected* uncertainty about the value of model parameters ($\pi_{\vartheta|x}$ in the Bayesian example),

(c) **environmental** or *unexpected* uncertainty owing to changes in model parameters (not accounted for in our Bayesian example, hence unexpected).

An adaptive learning rate that accurately reflects uncertainty

Various efforts have been made to come up with an adaptive learning rate:

- Kalman (1960)
- Sutton (1992)
- Nassar et al. (2010)
- Payzan-LeNestour & Bossaerts (2011)
- Mathys et al. (2011)
- Wilson et al. (2013)

We will look at two of these:

- **The Kalman filter** is optimal for linear dynamical systems, but realistic data usually require non-linear models.
- Mathys et al. use **a generic non-linear hierarchical Bayesian model** that allows us to derive update equations that are optimal in the sense that they minimize surprise.

Dealing with nonstationary environments: the Kalman filter

- We return to the Bayesian version of the lighthouse problem
- Relaxing the assumption that the underlying hidden state x is stationary and replacing it with a Gaussian random walk gives us the **Kalman filter:**

$$p\left(x^{(k)}|x^{(k-1)},\vartheta\right)=\mathcal{N}\left(x^{(k)};x^{(k-1)},\vartheta\right)$$

$$p\left(u^{(k)}|x^{(k)},\varepsilon\right)=\mathcal{N}\left(u^{(k)};x^{(k)},\varepsilon\right)$$

- Combining this with the **prior**

$$p\left(x^{(k-1)}\right)=\mathcal{N}\left(x^{(k-1)};\mu_x^{(k-1)},1/\pi_x^{(k-1)}\right),\ldots$$

Dealing with nonstationary environments: the Kalman filter

... and doing some algebra, we get the **posterior**

$$p\left(x^{(k)}\right) = \mathcal{N}\left(x^{(k)}; \mu_x^{(k)}, 1/\pi_x^{(k)}\right)$$

with

$$\pi_x^{(k)} = \frac{1}{\sigma_x^{(k-1)} + \vartheta} + \frac{1}{\varepsilon} = \hat{\pi}_x^{(k-1)} + \hat{\pi}_u$$

$$\mu_x^{(k)} = \mu_x^{(k-1)} + \frac{\hat{\pi}_u}{\pi_x^{(k)}}\left(u^{(k)} - \mu_x^{(k-1)}\right)$$

$$= \mu_x^{(k-1)} + \frac{\hat{\pi}_u}{\frac{1}{\sigma_x^{(k-1)} + \vartheta} + \hat{\pi}_u}\left(u^{(k)} - \mu_x^{(k-1)}\right)$$

The Kalman filter is optimal for linear dynamic systems.

Unfortunately, except for simple physical systems, **the world is not linear**. Living organisms need to be able to filter inputs whose rate of change changes, in other words: **processes whose volatility is volatile.**

Where would we need a model with an adaptive learning rate?

Task of Iglesias et al., *Neuron*, (2013)