FROM PERCEPTION TO ACTION: HIERARCHICAL GAUSSIAN FILTERS (HGFS)

Christoph Mathys

Computational Psychiatry Course

Zurich, 11 September 2024

INTERACTING MINDS CENTRE (IMC)

AARHUS UNIVERSITY

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS

NEW INFORMATION COMES AT US CONSTANTLY

Imagine the following situation:

You’re on a boat, you’re lost in a storm and trying to get back to shore. A lighthouse has just appeared on the horizon, but you can only see it when you’re at the peak of a wave. Your GPS etc., has all been washed overboard, but what you can still do to get an idea of your position is to measure the angle between north and the lighthouse. These are your measurements (in degrees):

76, 73, 75, 72, 77

What number are you going to base your calculation on?

Right. The mean: 74.6. How do you calculate that?

UPDATING THE MEAN OF A SERIES OF OBSERVATIONS

The usual way to calculate the mean $\bar{u}$ of $u_1, u_2, \dots, u_n$ is to take

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

This requires you to remember all u_i, which can become inefficient. Since the measurements arrive sequentially, we would like to update $\bar{u}$ sequentially as the u_i come in – without having to remember them.

It turns out that this is possible. After some algebra (see next slide), we get

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

UPDATING THE MEAN OF A SERIES OF OBSERVATIONS

- Proof of sequential update equation:

$$\bar{u}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} u_i = \frac{1}{n+1}\left(u_{n+1} + n \cdot \frac{1}{n}\sum_{i=1}^{n} u_i\right) =$$

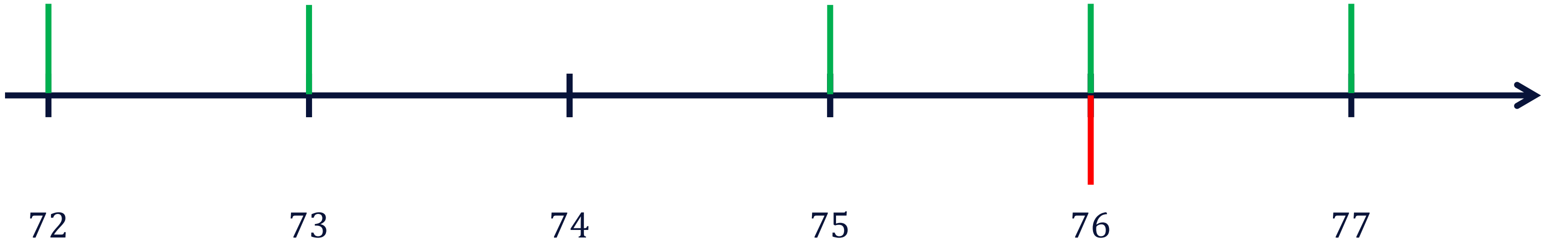
$$= \frac{1}{n+1}(u_{n+1} + n\bar{u}_n) = \frac{1}{n+1}(u_{n+1} - \bar{u}_n + (n+1)\bar{u}_n)$$

$$= \bar{u}_n + \tfrac{1}{n+1}(u_{n+1} - \bar{u}_n)$$

q.e.d.

UPDATING THE MEAN OF A SERIES OF OBSERVATIONS

- The seqential updates in our example now look like this:



$\bar{u}_1 = 76$

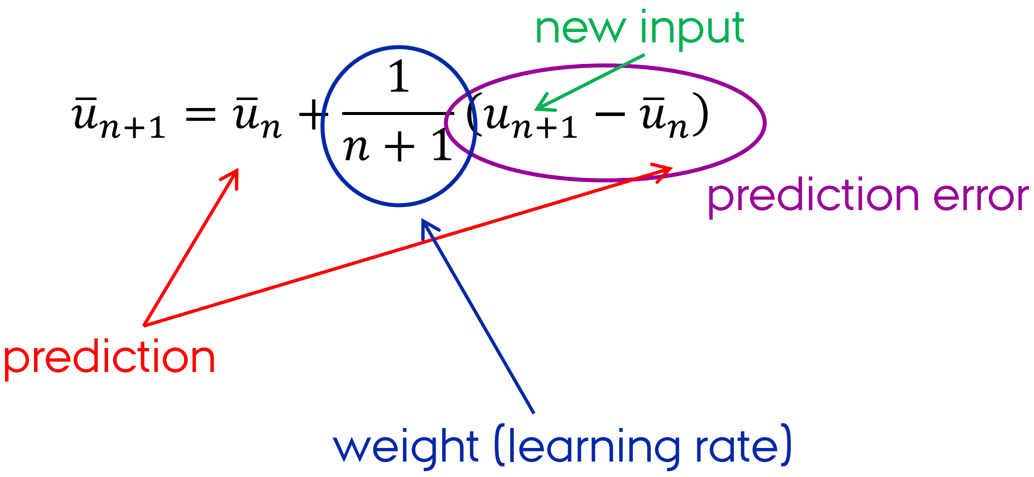
$\bar{u}_2 = 76 + \frac{1}{2}(73 - 76) = 74.5$

$\bar{u}_3 = 74.5 + \frac{1}{3}(75 - 74.5) = 74.\overline{6}$

$\bar{u}_4 = 74.\overline{6} + \frac{1}{4}(72 - 74.\overline{6}) = 74$

$\bar{u}_5 = 74 + \frac{1}{5}(77 - 74) = 74.6$

WHAT ARE THE BUILDING BLOCKS OF THE UPDATES WE'VE JUST SEEN?



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

new input

prediction error

prediction

weight (learning rate)

Is this a general pattern?

More specifically, does it generalize to Bayesian inference?

Indeed, it turns out that in many cases, Bayesian inference can be based on parameters that are updated using precision-weighted prediction errors.

UPDATES IN A SIMPLE GAUSSIAN MODEL

Think boat, lighthouse, etc., again, but now we’re doing Bayesian inference.

Before we make the next observation, our belief about the true value of the state x can be described by a Gaussian prior:

$$p(x) \sim \mathcal{N}(\mu_x, \pi_x^{-1})$$

The likelihood of an observation u is also Gaussian, with precision π_ε :

$$p(u|x) \sim \mathcal{N}(x, \pi_\varepsilon^{-1})$$

Bayes’ rule now tells us that the posterior is Gaussian again:

$$p(x|u) = \frac{p(u|x)p(x)}{\int p(u|x')p(x')\mathrm{d}x'} \sim \mathcal{N}\left(\mu_{x|u}, \pi_{x|u}^{-1}\right)$$

UPDATES IN A SIMPLE GAUSSIAN MODEL

Here's how the updates to the sufficent statistics μ and π describing our belief look like:

$$\pi_{x|u} = \pi_x + \pi_\varepsilon$$

$$\mu_{x|u} = \underbrace{\mu_x}_{\text{prediction}} + \underbrace{\frac{\pi_\varepsilon}{\pi_{x|u}}}_{\text{weight (learning rate)}=\frac{\text{how much we're learning here}}{\text{how much we already know}}} \overbrace{(u - \mu_x)}^{\text{prediction error}}$$

The mean is updated by an uncertainty-weighted (more specifically: precision-weighted) prediction error.

The size of the update is proportional to the likelihood precision and inversely proportional to the posterior precision.

This pattern is not specific to the univariate Gaussian case, but generalizes to Bayesian updates for all exponential families of likelihood distributions with conjugate priors (i.e., to all formal descriptions of inference you are ever likely to need).

HOW TO REVEAL THE PRECISION-WEIGHTING OF PREDICTION ERRORS WHEN SIMPLE EXPONENTIAL-FAMILY LIKELIHOODS WILL NOT DO

- Formulate the problem hierarchically (i.e., imitate evolution: when it built a brain that supports a mind which is a model of its environment, it came up with a (largely) hierarchical solution)

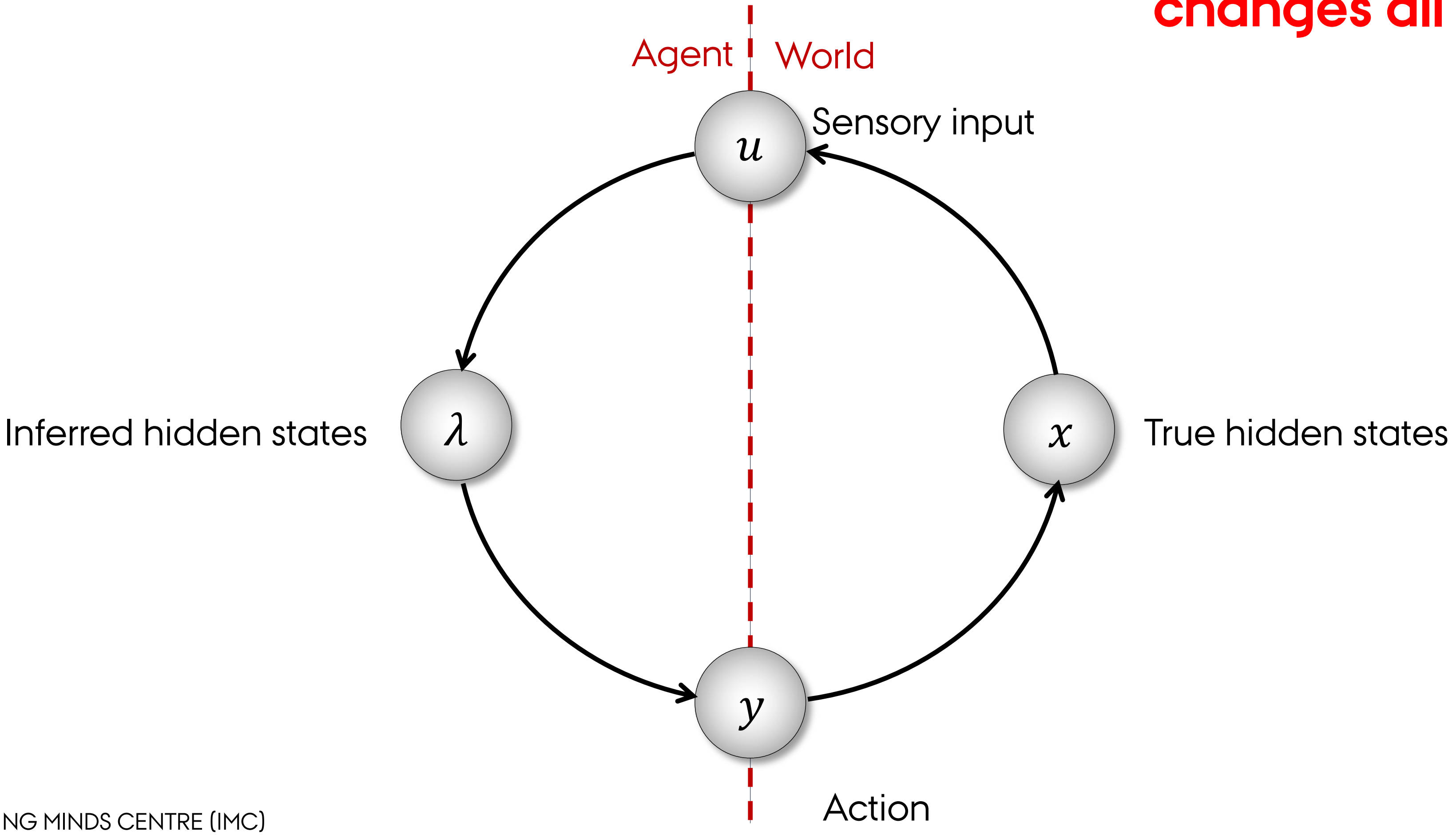
- Separate levels using a mean-field approximation

- Derive update equations

- Example: HGF

DOES INFERENCE AS WE'VE DESCRIBED IT ADEQUATELY DESCRIBE THE SITUATION OF ACTUAL BIOLOGICAL AGENTS?

No, the environment changes all the time!



THE ENVIRONMENT IS DYNAMIC

Up to now, we've only looked at inference on stationary 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 stationary 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].

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.

However, an optimal learning rate

a) Balances the need to learn faster as uncertainty increases with the need to learn more slowly as observation noise increases

b) Takes account of all sources of uncertainty (outcome, informational, environmental)

What we really need is an adaptive learning rate that accurately reflects the changing nature of the environment.

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

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

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

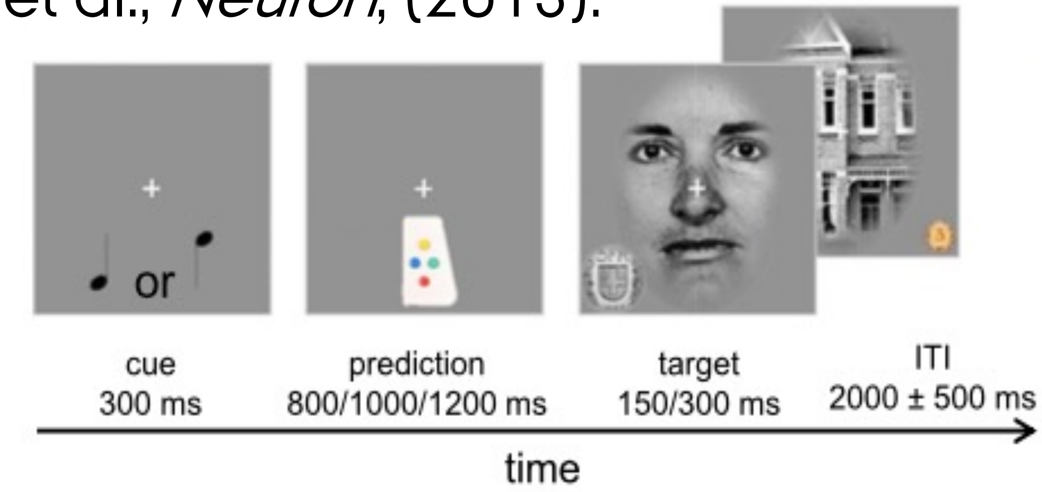
$$= \mu_x^{(k-1)} + \frac{\hat{\pi}_u}{\frac{1}{\sigma_x^{(k-1)} + \vartheta} + \hat{\pi}_u}\left(u^{(k)} - \mu_x^{(k-1)}\right)$$

The Kalman filter is optimal for linear dynamic systems.

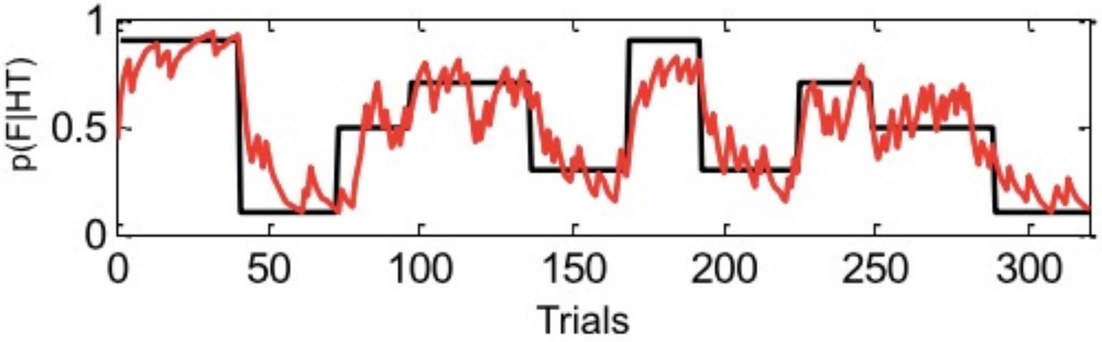
Unfortunately, except for simple physical systems, **the world is not linear**. Living organisms need to be able to filter inputs whose rate of change changes, in other words: **processes whose volatility is volatile.**

WHERE WOULD WE NEED A MODEL WITH AN ADAPTIVE LEARNING RATE?

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



Changes in cue strength (black), and posterior expectation of visual category (red)

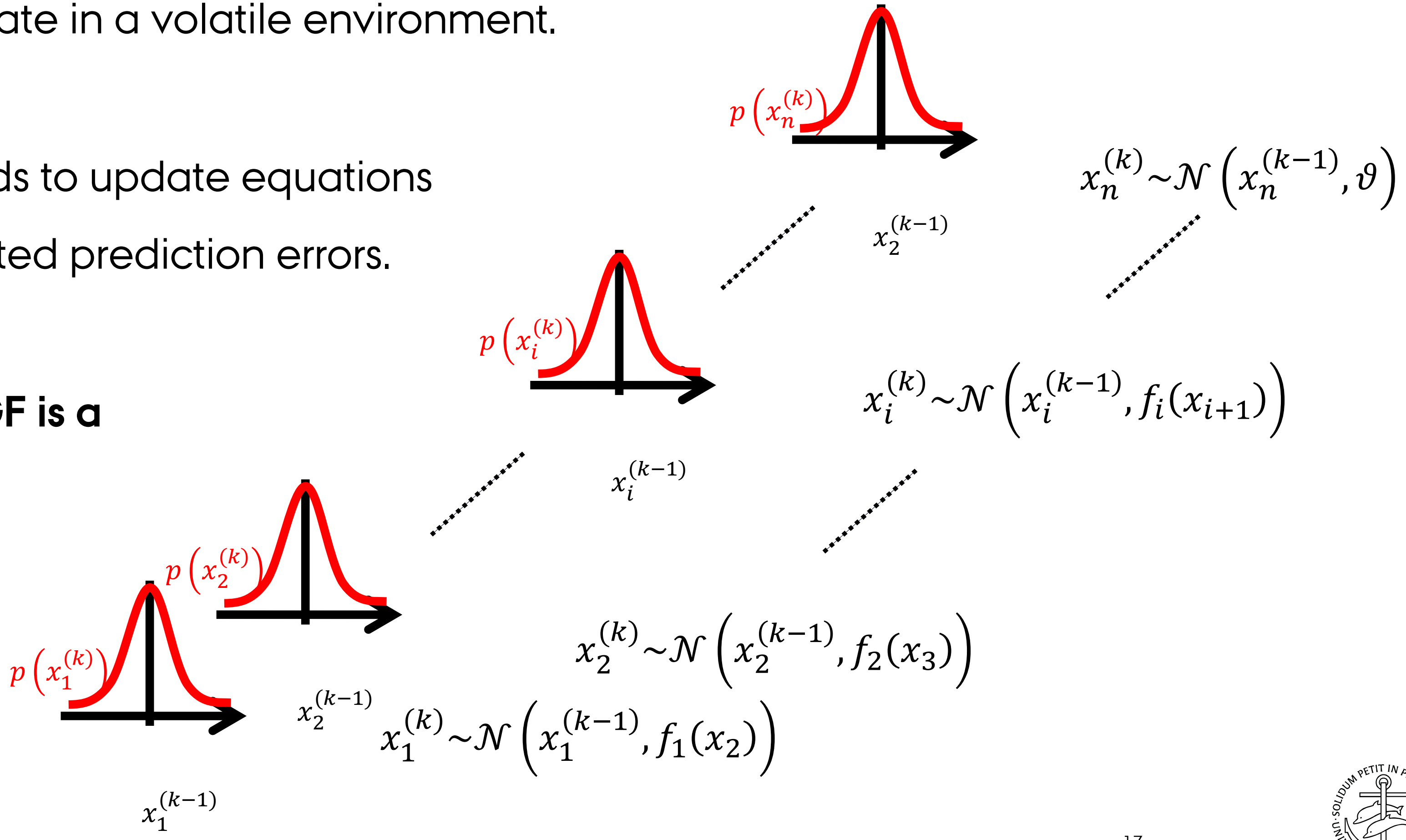


THE HIERARCHICAL GAUSSIAN FILTER (HGF, MATHYS ET AL., 2011; 2014)

- The HGF provides a generic solution to the problem of adaptingone’s learning rate in a volatile environment.

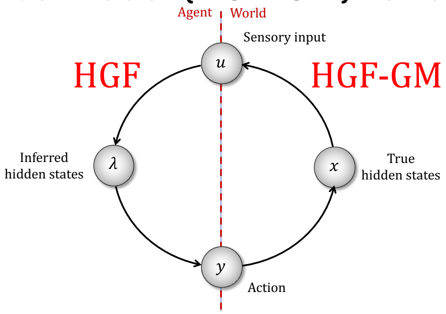
Variational inversion leads to update equations that are precision-weighted prediction errors.

With only 1 level, the HGF is a Kalman filter.


$p\left(x_n^{(k)}\right)$
$x_n^{(k)} \sim \mathcal{N}\left(x_n^{(k-1)}, \vartheta\right)$
$x_2^{(k-1)}$
$p\left(x_i^{(k)}\right)$
$x_i^{(k)} \sim \mathcal{N}\left(x_i^{(k-1)}, f_i(x_{i+1})\right)$
$x_i^{(k-1)}$
$p\left(x_2^{(k)}\right)$
$x_2^{(k)} \sim \mathcal{N}\left(x_2^{(k-1)}, f_2(x_3)\right)$
$p\left(x_1^{(k)}\right)$
$x_2^{(k-1)}$
$x_1^{(k)} \sim \mathcal{N}\left(x_1^{(k-1)}, f_1(x_2)\right)$
$x_1^{(k-1)}$

VARIATIONAL INVERSION AND UPDATE EQUATIONS

- Important distinction: **generative model** (HGF-GM) vs its inversion, **inference model** (HGF proper)



- Inversion of HGF-GM proceeds by introducing a mean field approximation and fitting quadratic approximations to the resulting variational energies (Mathys et al., 2011).
- This leads to **simple one-step update equations** (HGF proper) for the sufficient statistics (mean and precision) of the approximate Gaussian posteriors of the states x_i.
- The updates of the means have the same structure as value updates in Rescorla-Wagner learning:

$$\Delta\mu_i \propto \frac{\hat{\pi}_{i-1}}{\pi_i}\delta_{i-1}$$

Prediction error

Precisions determine learning rate

- The updates are **precision-weighted prediction errors**.

UPDATES AT THE FIRST LEVEL

- At the outcome level (i.e., at the very bottom of the hierarchy), we have

$$u^{(k)} \sim \mathcal{N}\left(x_1^{(k)}, \hat{\pi}_u^{-1}\right)$$

This gives us the following update for our belief on x_1 (our quantity of interest):

$$\pi_1^{(k)} = \hat{\pi}_1^{(k)} + \hat{\pi}_u$$

$$\mu_1^{(k)} = \mu_1^{(k-1)} + \frac{\hat{\pi}_u}{\pi_1^{(k)}}\left(u^{(k)} - \mu_1^{(k-1)}\right)$$

The familiar structure again – but now with a learning rate that is responsive to all kinds of uncertainty, including environmental (unexpected) uncertainty.

THE LEARNING RATE (‘KALMAN GAIN’) IN THE HGF

Unpacking the learning rate, we see:

outcome uncertainty

$$\frac{\hat{\pi}_u}{\pi_1^{(k)}} = \frac{\hat{\pi}_u}{\hat{\pi}_1^{(k)} + \hat{\pi}_u} = \frac{\hat{\pi}_u}{\frac{1}{\sigma_1^{(k-1)} + \exp\left(\kappa_1 \mu_2^{(k-1)} + \omega_1\right)} + \hat{\pi}_u}$$

informational uncertainty

environmental uncertainty (instead of the constant ϑ in the Kalman filter)