Theoretical Neuroscience Group,
Department of Physiology, University of Bern, Switzerland
Institute of Neuroinformatics, University of Zürich and ETH Zürich, Switzerland

The Neural Particle Filter

Scalability and Biological Implementation

Simone Carlo Surace
surace@ini.uzh.ch

April 17, 2018
SIAM Conference on Uncertainty Quantification
Session MS36
Controlled Interacting Particle Systems for Nonlinear Filtering

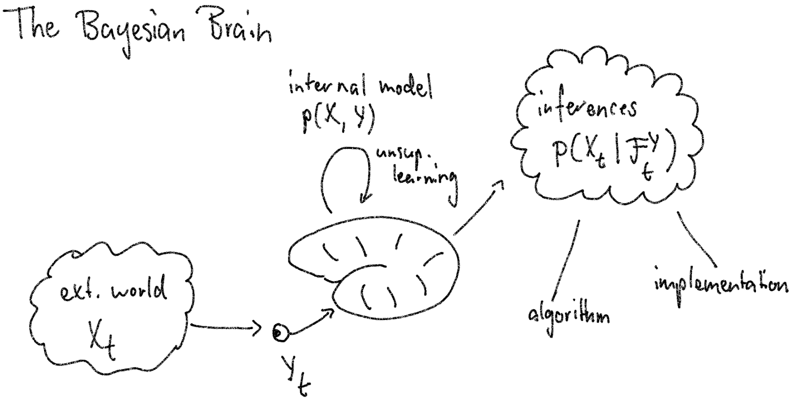
Thanks

Theoretical Neuroscience Group
Jean-Pascal Pfister, Anna Kutschireiter

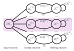
UNIVERSITÄT BERN

uzh | eth | zürich

FNSNF SWISS NATIONAL SCIENCE FOUNDATION

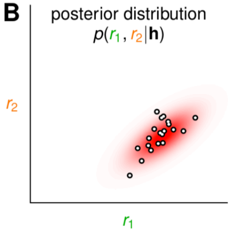

The Bayesian Brain
internal model
p(X, Y)
unsup. learning
inferences
$p(X_t \mid \mathcal{F}_t^Y)$
ext. world
X_t
Y_t
algorithm
implementation

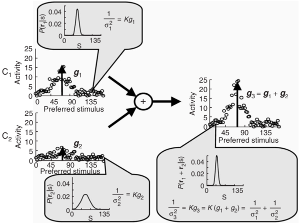
Content



Static inference

- Neural sampling: Aitchison & Lengyel (2016, 2017)
- Probabilistic Population Codes (PPC): Ma et al. (2006)



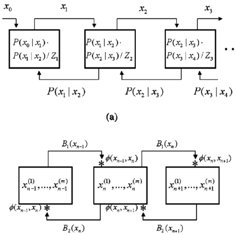


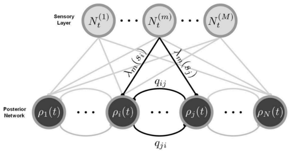
Existing work

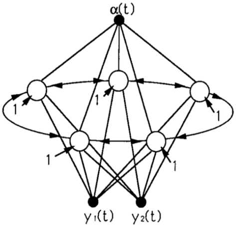
Neural-like algorithms to perform Bayesian inference

Filtering

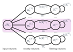
- Probabilistic Population Codes: Sokoloski (2017)
- Sampling: Lee Mumford (2003)
- Direct interpretation of the filtering equation as neuronal dynamics: Bobrowski et al. (2009), Legenstein et al. (2014)
- Synthetic approach: Ting-Ho Lo (1994)







Research goals
Desiderata



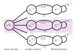
1. Solve the nonlinear filtering problem, at least approximately (hard just by itself),

Research goals

Desiderata

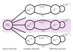
1. Solve the nonlinear filtering problem, at least approximately (hard just by itself),
2. using a sampling-based representation,

Research goals
Desiderata



1. Solve the nonlinear filtering problem, at least approximately (hard just by itself),
2. using a sampling-based representation,
3. with a biologically plausible algorithm:
   - neural dynamics,
   - local learning rules

Research goals
Desiderata



1. Solve the nonlinear filtering problem, at least approximately (hard just by itself),
2. using a sampling-based representation,
3. with a biologically plausible algorithm:
   - neural dynamics,
   - local learning rules
4. with a scalable algorithm.

Weighted particle filters

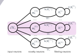
Why are they not suitable as models for the brain?

Why are standard (weighted) particle filters not on the list?

Why are standard (weighted) particle filters not on the list?
They are sampling-based...

Weighted particle filters

Why are they not suitable as models for the brain?



Generative model

$$dX_t = f(X_t)dt + g(X_t)dW_t, \quad (1)$$

$$dY_t = h(X_t)dt + dV_t, \quad (2)$$

Weighted particle filters

Why are they not suitable as models for the brain?

Generative model

$$dX_t = f(X_t)dt + g(X_t)dW_t, \tag{1}$$
$$dY_t = h(X_t)dt + dV_t, \tag{2}$$

Particle ensemble & unnormalized weights

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)}, \quad i = 1, ..., N \tag{3}$$
$$dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)})dY_t, \tag{4}$$

Bain & Crisan (2009), after Crisan & Lyons (1999)

Weighted particle filters

Why are they not suitable as models for the brain?

Generative model

$$dX_t = f(X_t)dt + g(X_t)dW_t, \tag{1}$$

$$dY_t = h(X_t)dt + dV_t, \tag{2}$$

Particle ensemble & unnormalized weights

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)}, \quad i = 1, ..., N \tag{3}$$

$$dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)})dY_t, \tag{4}$$

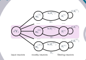
Estimate

$$\hat{\varphi}_t = \mathbb{E}[\varphi(X_t)|\mathcal{F}_t^Y] \approx \sum_{i=1}^{N} m_t^{(i)} \varphi(Z_t^{(i)}), \quad m_t^{(i)} = \frac{M_t^{(i)}}{\sum_{j=1}^{N} M_t^{(j)}}. \tag{5}$$

Bain & Crisan (2009), after Crisan & Lyons (1999)

Weighted particle filters

Why are they not suitable as models for the brain?



Generative model

$$dX_t = f(X_t)dt + g(X_t)dW_t, \tag{1}$$

$$dY_t = h(X_t)dt + dV_t, \tag{2}$$

Particle ensemble & unnormalized weights

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)}, \quad i = 1, ..., N \tag{3}$$

$$dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)})dY_t, \tag{4}$$

Estimate

$$\hat{\varphi}_t = \mathbb{E}[\varphi(X_t)|\mathcal{F}_t^Y] \approx \sum_{i=1}^{N} m_t^{(i)} \varphi(Z_t^{(i)}), \quad m_t^{(i)} = \frac{M_t^{(i)}}{\sum_{j=1}^{N} M_t^{(j)}}. \tag{5}$$

Bain & Crisan (2009), after Crisan & Lyons (1999)

Weighted particle filters

Why are they not suitable as models for the brain?

Generative model

$$dX_t = f(X_t)dt + g(X_t)dW_t, \tag{1}$$
$$dY_t = h(X_t)dt + dV_t, \tag{2}$$

Particle ensemble & unnormalized weights

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)}, \quad i = 1, \ldots, N \tag{3}$$
$$dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)})dY_t, \tag{4}$$

Estimate

$$\hat{\varphi}_t = \mathbb{E}[\varphi(X_t)|\mathcal{F}_t^Y] \approx \sum_{i=1}^{N} m_t^{(i)} \varphi(Z_t^{(i)}), \quad m_t^{(i)} = \frac{M_t^{(i)}}{\sum_{j=1}^{N} M_t^{(j)}}. \tag{5}$$

Bain & Crisan (2009), after Crisan & Lyons (1999)

Weighted particle filters

Why are they not suitable as models for the brain?

Generative model

$$dX_t = f(X_t)dt + g(X_t)dW_t, \tag{1}$$

$$dY_t = h(X_t)dt + dV_t, \tag{2}$$

Particle ensemble & unnormalized weights

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)}, \quad i = 1, ..., N \tag{3}$$

$$dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)})dY_t, \tag{4}$$

Estimate

$$\hat{\varphi}_t = \mathbb{E}[\varphi(X_t)|\mathcal{F}_t^Y] \approx \sum_{i=1}^{N} m_t^{(i)} \varphi(Z_t^{(i)}), \quad m_t^{(i)} = \frac{M_t^{(i)}}{\sum_{j=1}^{N} M_t^{(j)}}. \tag{5}$$

Bain & Crisan (2009), after Crisan & Lyons (1999)

Weighted particle filters
Why are they not suitable as models for the brain?

$$m_t^{(i)} = \frac{M_t^{(i)}}{\sum_{j=1}^N M_t^{(j)}}, \quad dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)}) dY_t, \quad dY_t = h(X_t)dt + dV_t.$$

Weighted particle filters

Why are they not suitable as models for the brain?

$$m_t^{(i)} = \frac{M_t^{(i)}}{\sum_{j=1}^{N} M_t^{(j)}}, \quad dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)}) dY_t, \quad dY_t = h(X_t)dt + dV_t.$$

Normalized weight dynamics

$$dm_t^{(i)} = m_t^{(i)}(h(Z_t^{(i)}) - \bar{h}_t) \cdot (h(X_t) - \bar{h}_t)dt + m_t^{(i)}(h(Z_t^{(i)}) - \bar{h}_t) \cdot dV_t$$

Weighted particle filters

Why are they not suitable as models for the brain?

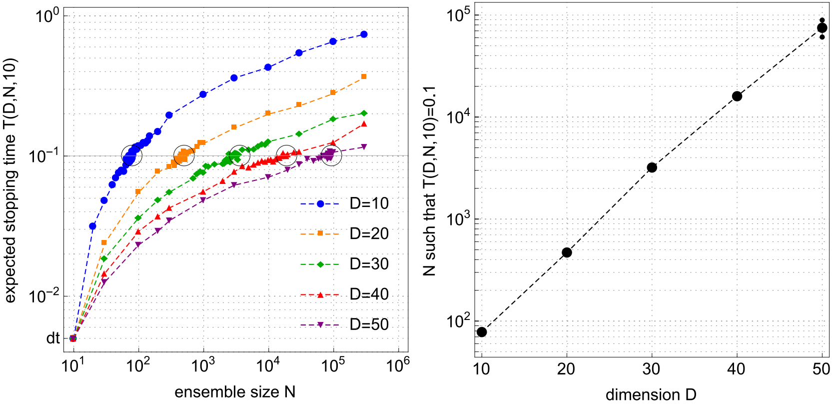
$$m_t^{(i)} = \frac{M_t^{(i)}}{\sum_{j=1}^N M_t^{(j)}}, \quad dM_t^{(i)} = M_t^{(i)} h(Z_t^{(i)}) dY_t, \quad dY_t = h(X_t)dt + dV_t.$$

Normalized weight dynamics

$$dm_t^{(i)} = m_t^{(i)}(h(Z_t^{(i)}) - \bar{h}_t) \cdot (h(X_t) - \bar{h}_t)dt + m_t^{(i)}(h(Z_t^{(i)}) - \bar{h}_t) \cdot dV_t$$

- $m_t^{(i)} \to \{0, 1\}$, with a rate proportional to D_Y.

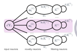
arXiv (2017), soon to appear SIAM Review



arXiv (2017), soon to appear SIAM Review

Weighted particle filters

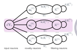
Why are they not suitable as models for the brain?



Why are standard (weighted) particle filters not on the list?
They are sampling-based...

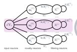
Weighted particle filters

Why are they not suitable as models for the brain?



Why are standard (weighted) particle filters not on the list?
They are sampling-based...
But they are not scalable with the dimensionality of the observations! $\rightarrow$ Curse of dimensionality (COD)

Weighted particle filters
Why are they not suitable as models for the brain?



Why are standard (weighted) particle filters not on the list? They are sampling-based...
But they are not scalable with the dimensionality of the observations! $\to$ Curse of dimensionality (COD)

How to improve the situation

- Resampling: biologically implausible, cannot keep up with the COD.
- Smarter dynamics for the particles
  - $dZ_t^{(i)} = ... + ...dY_t$, but importance weights cannot be defined (mutual singularity of measures),
  - $dt \to 0$ limits of 'optimal proposal' (Doucet et al., 2000) are trivial,
  - open question: are there other ways to incorporate observations into the dynamics, while preserving importance weights (FPF is fundamentally different). Goal: minimize rate of weight degeneracy!

Surace et al. (2017)

Ansatz (NPF)

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)} + K_t^{(i)}(dY_t - h(Z_t^{(i)})dt) \tag{6}$$

- In contrast to the FPF, this was not derived from a variational principle:

Kutschireiter et al. (2017)

Neural Particle Filter

A sampling-based algorithm without importance weights

Ansatz (NPF)

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)} + K_t^{(i)}(dY_t - h(Z_t^{(i)})dt) \tag{6}$$

- In contrast to the FPF, this was not derived from a variational principle:
  - It is a priori unclear how to set the gain K_t,

Kutschireiter et al. (2017)

Ansatz (NPF)

$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)} + K_t^{(i)}(dY_t - h(Z_t^{(i)})dt) \qquad (6)$$

- In contrast to the FPF, this was not derived from a variational principle:
  - It is a priori unclear how to set the gain K_t,
  - The hedging term in the feedback is missing (less interactions between particles),

Kutschireiter et al. (2017)

Neural Particle Filter

A sampling-based algorithm without importance weights

Ansatz (NPF)

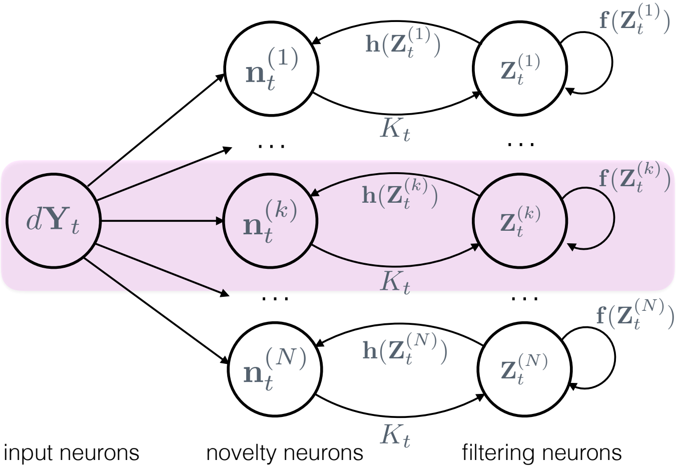
$$dZ_t^{(i)} = f(Z_t^{(i)})dt + g(Z_t^{(i)})dB_t^{(i)} + K_t^{(i)}(dY_t - h(Z_t^{(i)})dt) \tag{6}$$

- In contrast to the FPF, this was not derived from a variational principle:
  - It is a priori unclear how to set the gain K_t,
  - The hedging term in the feedback is missing (less interactions between particles),
- As the FPF, this is fundamentally detached from the weighted particle filter (even for very small K_t, the measures are still mutually singular).

Kutschireiter et al. (2017)

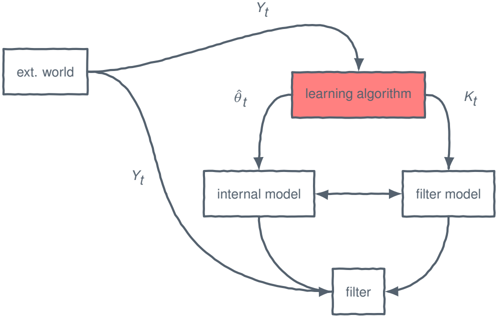
Neural Particle Filter

A sampling-based algorithm without importance weights



Kutschireiter et al. (2017)

Gain computation
Online maximum likelihood parameter estimation



Stochastic Gradient Ascent:

$$\mathcal{L}_t(\theta) = \int_0^t \hat{h}_s(\theta) dY_s - \frac{1}{2} \int_0^t \hat{h}_s^2(\theta) ds$$

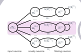
$$\partial_\theta \mathcal{L}_t(\theta) = \int_0^t \hat{h}_s^\theta(\theta)(dY_s - \hat{h}_s(\theta) ds)$$

$$d\hat{\theta}_t \propto \tilde{h}_t^\theta (dY_t - \tilde{h}_t dt)$$

Surace & Pfister (2016), Kutschireiter et al. (2017)

Gain computation

Online maximum likelihood parameter estimation



Update rules for elements of the gain matrix

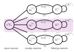
$$dK_t^{(i,j,k)} = \sum_{l=1}^{D_Y} \partial_j h_l(Z_t^{(i)}) \xi_t^{(i,j,k)} \left(dY_t^{(l)} - \frac{1}{N} \sum_{n=1}^{N} h_l(Z_t^{(n)}) dt \right) \tag{7}$$

- The processes $\xi_t^{(i,j,k)}$ measure the rate of change of the particles with respect to the gain matrix elements (sensitivity equations),

Surace & Pfister (2016), Kutschireiter et al. (2017)

Gain computation

Online maximum likelihood parameter estimation



Update rules for elements of the gain matrix

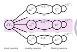
$$dK_t^{(i,j,k)} = \sum_{l=1}^{D_Y} \partial_j h_l(Z_t^{(i)}) \xi_t^{(i,j,k)} \left(dY_t^{(l)} - \frac{1}{N} \sum_{n=1}^{N} h_l(Z_t^{(n)}) dt \right) \qquad (7)$$

- The processes $\xi_t^{(i,j,k)}$ measure the rate of change of the particles with respect to the gain matrix elements (sensitivity equations),
- There are $N \times D_X \times D_Y$ additional variables,

Surace & Pfister (2016), Kutschireiter et al. (2017)

Gain computation

Online maximum likelihood parameter estimation



Update rules for elements of the gain matrix

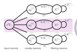
$$dK_t^{(i,j,k)} = \sum_{l=1}^{D_Y} \partial_j h_l(Z_t^{(i)}) \xi_t^{(i,j,k)} \left(dY_t^{(l)} - \frac{1}{N} \sum_{n=1}^{N} h_l(Z_t^{(n)}) dt \right) \tag{7}$$

- The processes $\xi_t^{(i,j,k)}$ measure the rate of change of the particles with respect to the gain matrix elements (sensitivity equations),
- There are $N \times D_X \times D_Y$ additional variables,
- There are a lot of nonlocal interactions between the variables (biology).

Surace & Pfister (2016), Kutschireiter et al. (2017)

Gain computation

Online maximum likelihood parameter estimation



Update rules for elements of the gain matrix

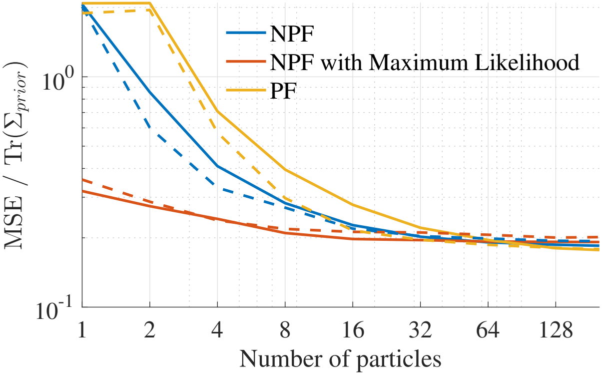
$$dK_t^{(i,j,k)} = \sum_{l=1}^{D_Y} \partial_j h_l(Z_t^{(i)}) \xi_t^{(i,j,k)} \left(dY_t^{(l)} - \frac{1}{N} \sum_{n=1}^{N} h_l(Z_t^{(n)}) dt \right) \qquad (7)$$

- The processes $\xi_t^{(i,j,k)}$ measure the rate of change of the particles with respect to the gain matrix elements (sensitivity equations),
- There are $N \times D_X \times D_Y$ additional variables,
- There are a lot of nonlocal interactions between the variables (biology).
- But it works:
  - Performance of the filter is surprisingly good,
  - The behavior of the gain matrix is as expected, i.e. it properly reflects uncertainty
  - The performance is similar as for the empirical gain

Surace & Pfister (2016), Kutschireiter et al. (2017)

Gain computation

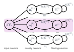
Online maximum likelihood parameter estimation



Kutschireiter et al. (2017)

Gain computation

Online maximum likelihood parameter estimation



What about Online Expectation Maximization (EM)?

Gain computation
Online maximum likelihood parameter estimation

What about Online Expectation Maximization (EM)?

Online EM

A variant of EM due to Mongillo & Deneve (2008) and Cappé (2011) for HMMs.

What about Online Expectation Maximization (EM)?

Online EM

A variant of EM due to Mongillo & Deneve (2008) and Cappé (2011) for HMMs.

- EM requires smoothing, online EM requires *forward* (recursive) smoothing,
- In continuous time, the forward smoothing problem leads to a modified Zakai equation similar to the Zakai equation for filtering,
- The associated modified Kushner-Stratonovich equation can be used to derive a modified FPF.

What about Online Expectation Maximization (EM)?

Online EM

A variant of EM due to Mongillo & Deneve (2008) and Cappé (2011) for HMMs.

- EM requires smoothing, online EM requires *forward* (recursive) smoothing,
- In continuous time, the forward smoothing problem leads to a modified Zakai equation similar to the Zakai equation for filtering,
- The associated modified Kushner-Stratonovich equation can be used to derive a modified FPF.
- Unfortunately, this does not work for parameters that only appear in the filter!

Gain computation

Online maximum likelihood parameter estimation

What about Online Expectation Maximization (EM)?

Online EM

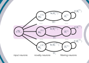
A variant of EM due to Mongillo & Deneve (2008) and Cappé (2011) for HMMs.

- EM requires smoothing, online EM requires *forward* (recursive) smoothing,
- In continuous time, the forward smoothing problem leads to a modified Zakai equation similar to the Zakai equation for filtering,
- The associated modified Kushner-Stratonovich equation can be used to derive a modified FPF.
- Unfortunately, this does not work for parameters that only appear in the filter!
- The modified FPF also has a gain that needs to be set!

Summary and Conclusions

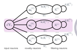
- We phrased perceptual inference in a dynamic setting as nonlinear filtering problem.

Summary and Conclusions



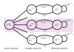
- We phrased perceptual inference in a dynamic setting as nonlinear filtering problem.
- When a sampling based representation is desired, it needs to be unweighted.

Summary and Conclusions



- We phrased perceptual inference in a dynamic setting as nonlinear filtering problem.
- When a sampling based representation is desired, it needs to be unweighted.
- The NPF, which is very similar to the FPF, is biologically interpretable and avoids the scaling problem in high dimensions.

Summary and Conclusions



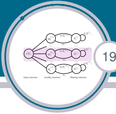
- We phrased perceptual inference in a dynamic setting as nonlinear filtering problem.
- When a sampling based representation is desired, it needs to be unweighted.
- The NPF, which is very similar to the FPF, is biologically interpretable and avoids the scaling problem in high dimensions.
- The main difficulty is the computation of the gain, for which various methods have been proposed.

Summary and Conclusions

- We phrased perceptual inference in a dynamic setting as nonlinear filtering problem.
- When a sampling based representation is desired, it needs to be unweighted.
- The NPF, which is very similar to the FPF, is biologically interpretable and avoids the scaling problem in high dimensions.
- The main difficulty is the computation of the gain, for which various methods have been proposed.
- A maximum likelihood approach to gain computation does not produce biologically meaningful update rules, but may be of interest for other applications if the dimensionality of the problem is moderate.

Online EM

Details



Jensen's inequality:

$$\mathcal{L}_t^Y(\theta) \geq \mathcal{L}_t^Y(\theta') + Q(\theta, \theta'),$$

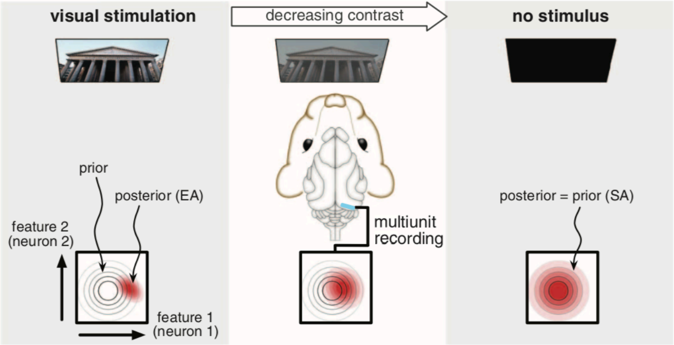
$$Q(\theta, \theta') = \mathbb{E}_{\theta'}\left[\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta'}} \middle| \mathcal{F}_t^Y\right].$$

For a diffusion model,

$$\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta'}} = R(\theta') + \int_0^t \varphi(X_t, \theta) dX_t + \int_0^t \psi(X_t, \theta) dY_t + \int_0^t \zeta(X_t, \theta) dt.$$

Use generalized measure:

$$\tilde{\rho}_t[\varphi, \xi, \theta] \doteq \mathbb{E}_\theta^\dagger\left[\varphi(X_t) e^{\xi \cdot S_t} \frac{d\mathbb{P}_\theta}{d\mathbb{P}_\theta^\dagger} \middle| \mathcal{F}_t^Y\right], \quad \xi \in \mathbb{R}^{n_S}. \tag{8}$$


visual stimulation
decreasing contrast
no stimulus
prior
posterior (EA)
feature 2
(neuron 2)
feature 1
(neuron 1)
multiunit
recording
posterior = prior (SA)

Experimental evidence

The Bayesian Brain