A plausible model of recognition and postdiction in dynamic environment

Kevin Li, Maneesh Sahani

Gatsby Computational Neuroscience Unit, University College London

June 4, 2019

1. Introduction

Inference using an internal model (Helmholtz machine)

static world

$$p(\boldsymbol{z}, \boldsymbol{x}) \longrightarrow \boldsymbol{r}(\cdot)$$

$$\begin{tikzcd} \boldsymbol{z} \arrow[d] & \boldsymbol{r} \\ \boldsymbol{x} & \boldsymbol{x} \arrow[u] \end{tikzcd}$$

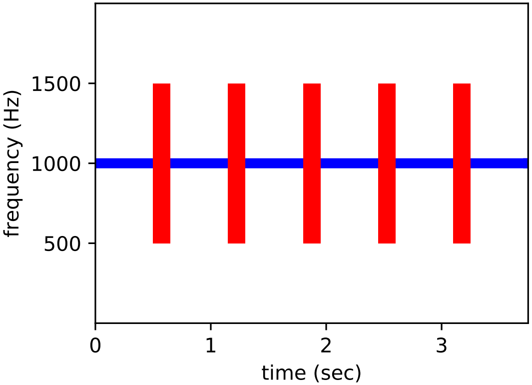
Dayan, Hinton, Neal & Zemel, 1995

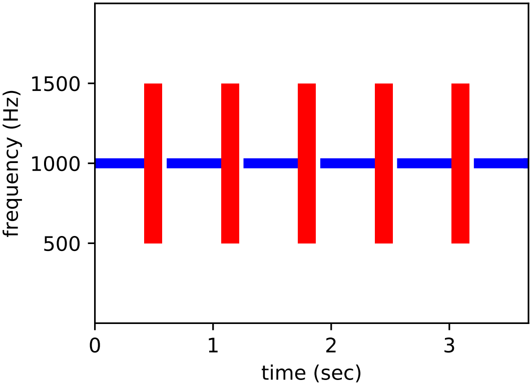
dynamic world

$$p(\boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}) \longrightarrow \boldsymbol{r}_t(\cdot)$$

$$\begin{tikzcd} {} \arrow[r] & \boldsymbol{z}_{t-1} \arrow[r] \arrow[d] & \boldsymbol{z}_t \arrow[r] \arrow[d] & {} & {} \arrow[r] & \boldsymbol{r}_{t-1} \arrow[r] & \boldsymbol{r}_t \arrow[r] & {} \\ & \boldsymbol{x}_{t-1} & \boldsymbol{x}_t & & & \boldsymbol{x}_{t-1} \arrow[u] & \boldsymbol{x}_t \arrow[u] & \end{tikzcd}$$

Illusion 1

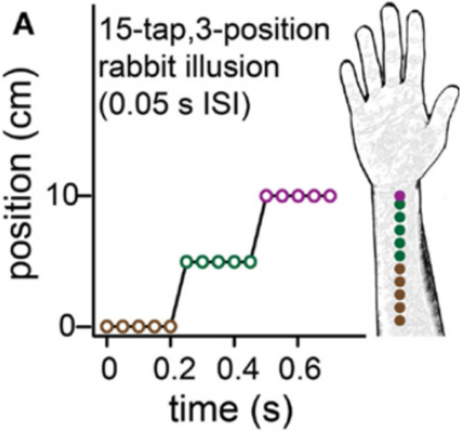

frequency (Hz)
1500
1000
500
0
1
2
3
time (sec)


frequency (Hz)
1500
1000
500
0
1
2
3
time (sec)

Illusion 2: cutaneous rabbit

In the course of designing some experiments on the cutaneous perception of mechanical pulses delivered to the back of the forearm, it was discovered that, under some conditions of timing, the taps produced seemed not to be properly localized under the contactors. [...] They will seem to be distributed, with more or less uniform spacing, from the region of the first contactor to that of the third. **There is a smooth progression of jumps up the arm, as if a tiny rabbit were hopping from elbow to wrist**.

Geldard & Sherrick, 1972, Science



In those examples:

- Percept of the past is changed by new, future observation
- "Law of Continuity"

What could be the neural basis for these *statistical computation*?

- **representing** beliefs as distributional objects
- **updating** *beliefs of the past* based on new evidence in *real time*?
- **learning** to do all the above

2. Distributed distributional code

DDC: a framework for neural representation of uncertainty

A DDC encodes a **probability distribution**:

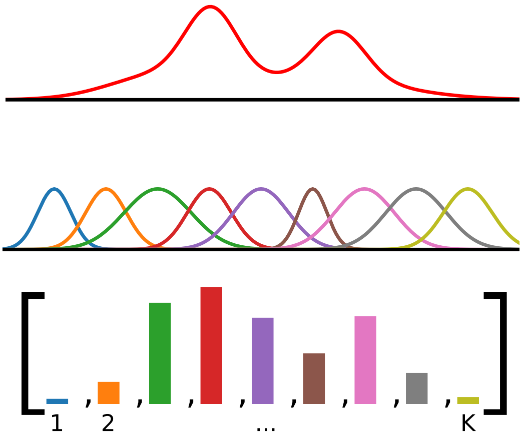
$$q(\boldsymbol{z})$$

by a set of **tuning functions**

$$\boldsymbol{\gamma}(\boldsymbol{z}) := [\gamma_1(\boldsymbol{z}), \gamma_2(\boldsymbol{z}), \gamma_3(\boldsymbol{z}), \ldots, \gamma_K(\boldsymbol{z})]$$

into a set of **expectations**

$$\boldsymbol{r} := \mathbb{E}_{q(\boldsymbol{z})}[\gamma_1(\boldsymbol{z}), \gamma_2(\boldsymbol{z}), \ldots, \gamma_K(\boldsymbol{z})]$$

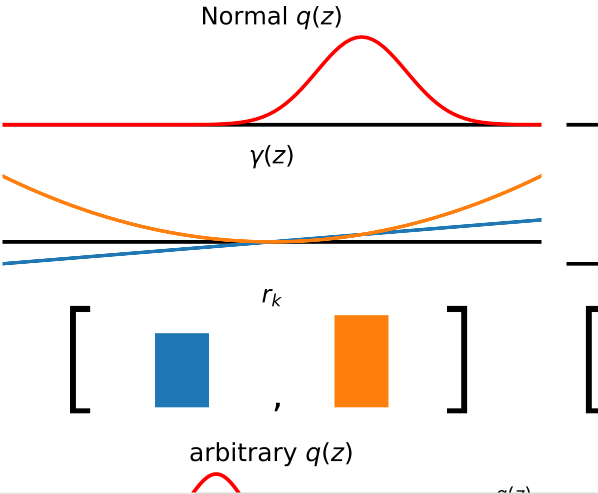


Zemel, Dayan & Pouget (1998); Sahani & Dayan (2003), Vértes & Sahani (2018)

Why DDC? It can encode a large family of distributions

given $\mathbb{E}_q\left[\gamma_k\left(\boldsymbol{z}\right)\right] = r_k, \quad \forall k \in \{1, 2, \ldots K\}$

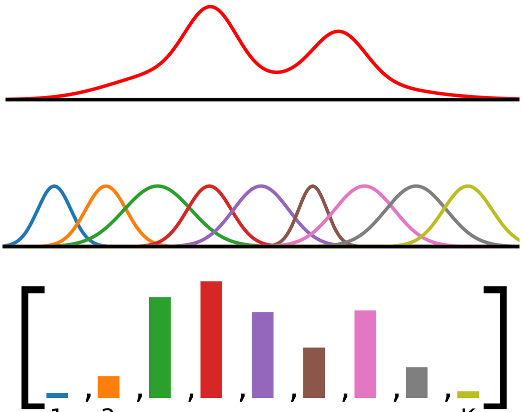
$$
\begin{aligned}
q(\boldsymbol{z}) &= \arg\max H[q] \\
&= \frac{1}{Z(\boldsymbol{r})} \exp\left[\sum_k \theta_k(\boldsymbol{r})\gamma_k(\boldsymbol{z})\right]
\end{aligned}
$$



Why DDC? It makes computation simple for neurons

$$\boldsymbol{r} := \mathbb{E}_{q(\boldsymbol{z})}\left[\gamma_1(\boldsymbol{z}), \gamma_2(\boldsymbol{z}), \ldots, \gamma_K(\boldsymbol{z})\right]$$

$$\boldsymbol{r} := \mathbb{E}_{q(\boldsymbol{z})}\left[\gamma_1(\boldsymbol{z}), \gamma_2(\boldsymbol{z}), \ldots, \gamma_K(\boldsymbol{z})\right]$$



Key computations involve expected values:

- Message passing:
  $q(z_2) = \mathbb{E}_{q(z_1)}\left[p(z_2|z_1)\right]$
- Marginalization:
  $q(z_2 \in (a,b)) = \mathbb{E}_{q(z_1,z_2)}\left[1(a < z_2 < b)\right]$
- Action evaluation:
  $Q(a) = \mathbb{E}_{q(s)}\left[R(s,a)\right]$

How to compute $\mathbb{E}\left[h(\boldsymbol{z})\right]$?

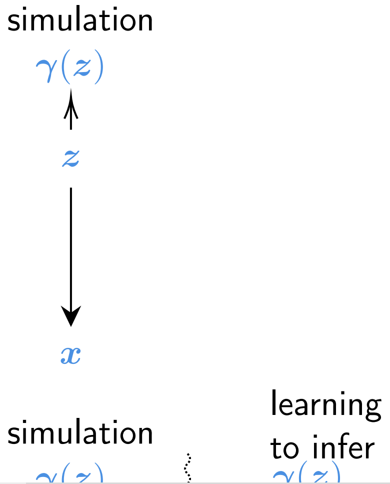
if $h(\boldsymbol{z}) \approx \sum \alpha_i \gamma_i(\boldsymbol{z}) \Rightarrow \mathbb{E}\left[h(\boldsymbol{z})\right] \approx \sum \alpha_i r_i$

Why DDC? It allows biologically plausible learning to infer

- Previously, we saw marginal $q(\boldsymbol{z}) \leftrightarrow \mathbb{E}_{q(\boldsymbol{z})}\left[\boldsymbol{\gamma}(\boldsymbol{z})\right]$
- For $p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$, $p(\boldsymbol{z}|\boldsymbol{x})$ is intractable
- How to obtain approx. $q(\boldsymbol{z}|\boldsymbol{x}) \leftrightarrow \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\boldsymbol{\gamma}(\boldsymbol{z})\right]$?
- Clue: posterior mean...
$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}\left[\boldsymbol{\gamma}(\boldsymbol{z})\right] = \arg\min_{\boldsymbol{\phi}} \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}\left[\|\boldsymbol{\gamma}(\boldsymbol{z}) - \boldsymbol{\phi}\|_2^2\right]$$
- "Amortize" using $\boldsymbol{\phi}_{\boldsymbol{W}}(\boldsymbol{x}) := \boldsymbol{W}\boldsymbol{\sigma}(\boldsymbol{x})$

$$\boxed{\begin{aligned} \boldsymbol{W}^* &= \arg\min_{\boldsymbol{W}} \mathbb{E}_{p(\boldsymbol{z},\boldsymbol{x})}\left[\|\boldsymbol{\gamma}(\boldsymbol{z}) - \boldsymbol{W}\boldsymbol{\sigma}(\boldsymbol{x})\|_2^2\right] \\ \boldsymbol{r}(\boldsymbol{x}) &:= \boldsymbol{W}^*\boldsymbol{\sigma}(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\boldsymbol{\gamma}(\boldsymbol{z})\right] \end{aligned}}$$

- Find $\boldsymbol{W}^*$ by the **delta rule**:
$$\Delta\boldsymbol{W} \propto (\boldsymbol{\gamma}(\boldsymbol{z}) - \boldsymbol{\phi}_{\boldsymbol{W}}(\boldsymbol{x}))\boldsymbol{\sigma}(\boldsymbol{x})^{\mathsf{T}}, \quad \{\boldsymbol{z}, \boldsymbol{x}\} \sim p$$



DDC Summary

Definition

DDC of $q(\boldsymbol{z})$ associated tuning functions $\boldsymbol{\gamma}(\boldsymbol{z})$ is $\boldsymbol{r} := \mathbb{E}_{q(\boldsymbol{z})}[\boldsymbol{\gamma}(\boldsymbol{z})]$

MaxEnt interpretation

$$\boldsymbol{r} \underset{\boldsymbol{\gamma}(\boldsymbol{z}),\text{MaxEnt}}{\overset{\mathbb{E}[\boldsymbol{\gamma}(\boldsymbol{z})]}{\leftrightharpoons}} q(\boldsymbol{z}) \propto \exp\left[\boldsymbol{\theta} \cdot \boldsymbol{\gamma}(\boldsymbol{z})\right]$$

Expectation approximation

$$h(\boldsymbol{z}) \approx \boldsymbol{\alpha} \cdot \boldsymbol{\gamma}(\boldsymbol{z}) \implies \mathbb{E}\left[h(\boldsymbol{z})\right] \approx \boldsymbol{\alpha} \cdot \boldsymbol{r}$$

Learning to infer given $p(z, x)$

$$\boldsymbol{r}(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\boldsymbol{\gamma}(\boldsymbol{z})\right] = \boldsymbol{W}^{*}\boldsymbol{\sigma}(\boldsymbol{x}), \qquad \Delta\boldsymbol{W} \propto (\boldsymbol{\gamma} - \boldsymbol{\phi}_{\boldsymbol{W}})\boldsymbol{\sigma}^{\intercal}, \quad \{\boldsymbol{z}, \boldsymbol{x}\} \sim p(\boldsymbol{z}, \boldsymbol{x})$$

3. Online recognition and postdiction

A generic dynamic internal model

We assume a generic internal model

$$\boldsymbol{z}_t = \boldsymbol{f}(\boldsymbol{z}_{t-1}, \xi^{(z)})$$
$$\boldsymbol{x}_t = \boldsymbol{g}(\boldsymbol{z}_t, \xi^{(x)})$$

$$\begin{tikzcd}
{} \arrow[r] & \boldsymbol{z}_{t-1} \arrow[r] \arrow[d] & \boldsymbol{z}_t \arrow[r] \arrow[d] & {} \\
& \boldsymbol{x}_{t-1} & \boldsymbol{x}_t &
\end{tikzcd}$$

Assumptions

- Discrete-time
- Markov property
- Stationarity

Representing and computing beliefs of the whole history

- Online recognition (filtering): maintain $q(\boldsymbol{z}_t|\boldsymbol{x}_{1:t})$ to allow postdiction
- Define **temporally extended encoding function** $\boldsymbol{\psi}_t := \boldsymbol{\psi}\left(\boldsymbol{z}_{1:t}\right)$ for DDC
- A plausible ψ_t

$$\boldsymbol{\psi}_1 := \boldsymbol{\gamma}\left(\boldsymbol{z}_1\right)$$
$$\boldsymbol{\psi}_t := \boldsymbol{U}\boldsymbol{\psi}_{t-1} + \boldsymbol{\gamma}\left(\boldsymbol{z}_t\right)$$

- ***Update beliefs about the past***: compute

$$\boldsymbol{r}_t = \mathbb{E}_{q(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})}\left[\boldsymbol{\psi}_t\right]$$

  from $\boldsymbol{r}_{t-1}$ and $\boldsymbol{x}_t$
- ***Postiction:*** readout statistics

$$h(\boldsymbol{z}_{t-\tau}) \approx \boldsymbol{\alpha} \cdot \boldsymbol{\psi}(\boldsymbol{z}_{1:t}) \implies \mathbb{E}_{q(\boldsymbol{z}_{t-\tau})}\left[h(\boldsymbol{z}_{t-\tau})\right] \approx \boldsymbol{\alpha} \cdot \boldsymbol{r}_t$$

Learning to postdict online

- Want $\boldsymbol{r}_t(\boldsymbol{x}_{1:t}) = \mathbb{E}_{q_t}\left[\boldsymbol{\psi}_t\left(\boldsymbol{z}_{1:t}\right)\right]$
- Recall for $p(\boldsymbol{z}, \boldsymbol{x})$

$$\boldsymbol{r} := \boldsymbol{\phi}_{\boldsymbol{W}^*}(\boldsymbol{x})$$

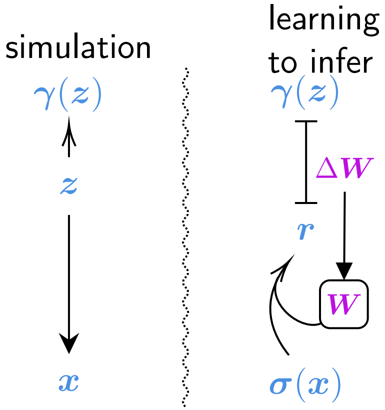
- Likewise, for SSM $p(\boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t})$

$$\begin{aligned}\boldsymbol{r}_t &:= \boldsymbol{\phi}_{\boldsymbol{W}_t}(\boldsymbol{r}_{t-1}, \boldsymbol{x}_t) \\ &= \boldsymbol{W}_t \cdot (\boldsymbol{r}_{t-1} \otimes \boldsymbol{\sigma}(\boldsymbol{x}_t))\end{aligned}$$

- Learning by the delta rule

$$\Delta \boldsymbol{W}_t \leftarrow (\boldsymbol{\psi}_t - \boldsymbol{\phi}_t)(\boldsymbol{r}_{t-1} \otimes \boldsymbol{\sigma}_t)^{\mathsf{T}}$$

$$\{\boldsymbol{\psi}_t, \boldsymbol{x}_t, \boldsymbol{r}_{t-1}\} \sim p(\boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}), \{\boldsymbol{h}_{\boldsymbol{W}_i}\}_{i=1}^{t-1}$$

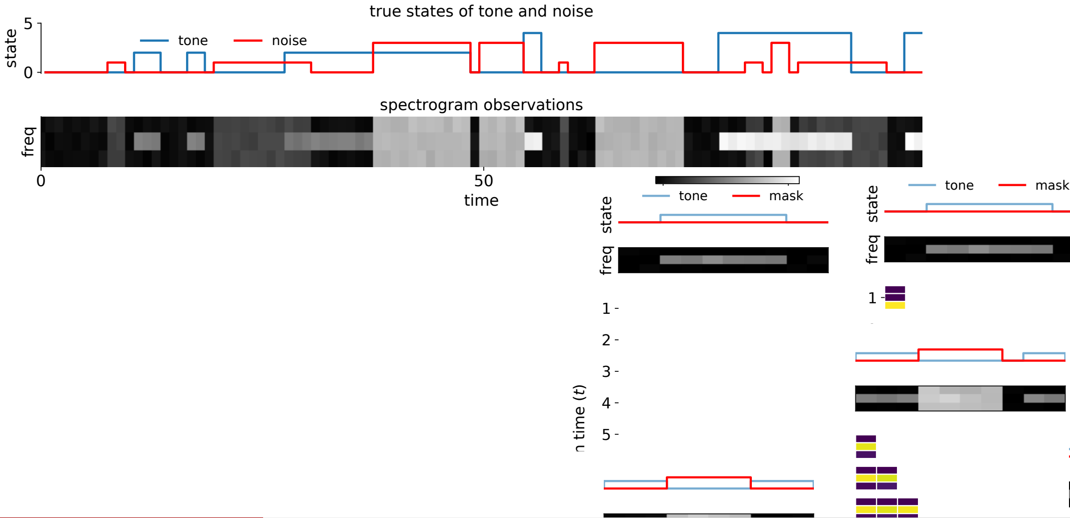


Simulation

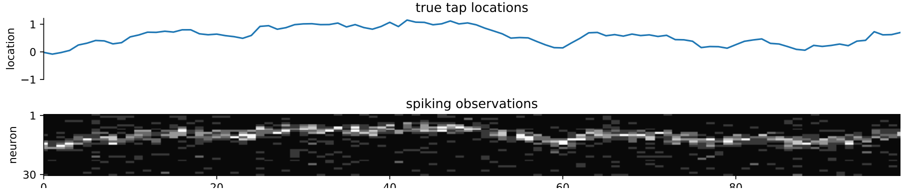
Learning to infer

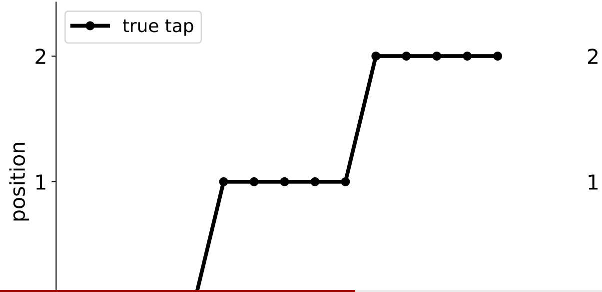
4. Testing DDC filtering on simulated experiments

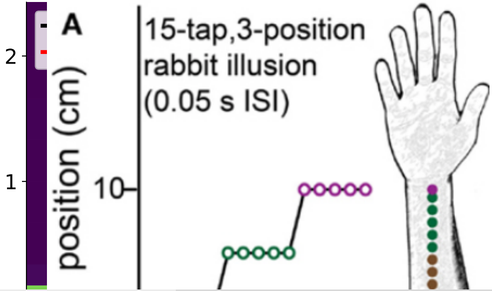
Auditory continuity illusion



Cutaneous rabbit







Summary of DDC filtering and Questions

Representation: DDC $\boldsymbol{r}_t := \mathbb{E}_{q(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})}\left[\boldsymbol{\psi}_t(\boldsymbol{z}_{1:t})\right]$
Computation: bilinear $\boldsymbol{r}_t := \boldsymbol{W}^* \cdot (\boldsymbol{r}_{t-1} \otimes \boldsymbol{\sigma}(\boldsymbol{x}_t))$ **, or linear** $\mathbb{E}_q\left[h(\boldsymbol{z}_{t-\tau})\right] \approx \boldsymbol{\alpha}^\intercal \boldsymbol{r}_t$
Learning to infer: delta rule $\Delta\boldsymbol{W} \propto (\boldsymbol{\psi}_t - \boldsymbol{\phi}_{\boldsymbol{W}})(\boldsymbol{r}_{t-1} \otimes \boldsymbol{\sigma}_t)$**, similar for readout** $\boldsymbol{\alpha}$
Questions:

- Is the encoding function $\boldsymbol{\psi}_t = \boldsymbol{U}\boldsymbol{\psi}_{t-1} + \boldsymbol{\gamma}(\boldsymbol{z}_t)$ the best?
- Does the brain encode the joint $q(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$?
- Any theoretical argument for using the bilinear rule?
- Is there a way to also adapt $\boldsymbol{U}$ in a plausible way?
- Can we encode the internal model by DDC? (talk to Eszter Vértes)