# A neurally plausible model for online recognition and postdiction

**Li K. Wenliang** **Maneesh Sahani**
Gatsby Computational Neuroscience Unit
University College London
London, W1T 4JG
{kevinli,maneesh}@gatsby.ucl.ac.uk

## Abstract

Humans and other animals are frequently near-optimal in their ability to integrate noisy and ambiguous sensory data to form robust percepts—which are informed both by sensory evidence and by prior experience about the causal structure of the environment. It is hypothesized that the brain establishes these structures using an internal model of how the observed patterns can be generated from relevant but unobserved causes. In dynamic environments, such integration often takes the form of *postdiction*, wherein later sensory evidence affects inferences about earlier percepts. As the brain must operate in current time, without the luxury of acausal propagation of information, how does such postdictive inference come about? Here, we propose a general framework for neural probabilistic inference in dynamic models based on the distributed distributional code (DDC) representation of uncertainty, naturally extending the underlying encoding to incorporate implicit probabilistic beliefs about both present and past. We show that, as in other uses of the DDC, an inferential model can be learnt efficiently using samples from an internal model of the world. Applied to stimuli used in the context of psychophysics experiments, the framework provides an online and plausible mechanism for inference, including postdictive effects.

## 1 Introduction

The brain must process a constant stream of noisy and ambiguous sensory signals from the environment, making accurate and robust real-time perceptual inferences crucial for survival. Many behavioral experiments suggest that humans and other animals achieve nearly Bayes-optimal performance across a range of contexts involving noise and uncertainty: e.g., when combining noisy signals across sensory modalities [35, 16, 1], making sensory decisions with consequences of unequal value [45], or inferring causal structure in the sensory environment [24]. Such perception becomes more challenging in dynamical environments (referred to as filtering). Beliefs about dynamical quantities must be continuously and rapidly updated on the basis of new sensory input, and very often informative sensory inputs will arrive after the time of the relevant state. Thus, perception in dynamical environments requires a combination of *prediction*—to ensure actions are not delayed relative to the external world—and *postdiction*—to ensure that perceptual beliefs about the past are correctly updated by subsequent sensory evidence [41, 12, 21, 18, 8, 34].

Behavioral [3, 7, 23, 33, 47] and physiological [17, 9, 10] findings suggest that the brain acquires an internal model of how relevant states of the world evolve, and how they give rise to the stream of sensory evidence. Recognition is then formally a process of statistical inference which inverts this internal model to form perceptual beliefs about the latent causes given these observations. While this type of statistical computation is well understood and accounts for nearly optimal perception in

Preprint. Under review.

experiments, it remains largely unknown how the brain carries out these computations in non-trivial but biologically relevant situations. First, to afford such computations, abstract distributional objects have to be instantiated or represented in the brain, whether they are continuous or discrete. A number of neural representations of distributions have been proposed, including sample-based representations [22, 36], so-called probabilistic population codes (PPCs) [28, 6] and other distributional representations [13, 48, 40]—however the plausibility or otherwise of these hypotheses remains debated.

Second, computations over these distributional representations need to be efficient and accurate to obtain informative and robust percpets. In dynamic inference, it is necessary for optimal evidence integration to retain both a point estimate and the associated uncertainty or multiplicity [40] regarding the unobserved quantities of interest. The causal process of realistic dynamical environment is usually highly nonlinear, making exact inference intractable and necessitating approximations. PPC makes explicit parametric assumptions about the inferred posterior distribution, and the resulting analytical solution are implemented by neural circuits, making approximations as needed [28, 7, 4, 29], including the case for the Kalman filter [5]. In the sampling approach, the neural particle filter [25] approximates the exact filtering solution for latent dynamics described by nonlinear stochastic differential equations. When the latent variables are discrete, authors in [26] propose that an ensemble of spiking neurons can be used to encode how likely each outcome is. Using energy-based models, a recurrent exponential family harmonium with binary latent variables [31] can be trained to perform filtering without explicit specification of the unknown latent dynamics, and this model architecture is generalized in [42] where a "neural Bayes rule" is proposed for the PPC representation.

Third, it is a challenge for the brain *learns* to perform inference for the many different tasks on which near-optimal recognition has been observed. In most of the aforementioned frameworks (except the energy based models), special neural circuits need to be wired for specific problems according to structures implied by approximate solutions. However, the sparse feedback given to subjects in behavioral experiments is unlikely to be sufficient for rewiring as dictated by those analyitical approximations. General artificial neural networks, when given sufficient computational capacity (e.g. using error-backpropagation), can be approximated the optimal solutions from supervision signals (errors) [14]; but it remains unknown as to what kind of representation of uncertainty is adopted in such networks, and what kind of recipe it follows for computation.

In this work, we introduce a powerful online recognition scheme that address the three aspects above. We first review the distributed distributional code [40, 44] as a representation of uncertainty in the brain. We then show that this representation allows efficient and accurate computation of probabilistic percepts over latent causes in a generic class of internal models. Importantly, each new observation is used to update not only beliefs about the present time, but also about the recent history—thus implementing a form of real-time *online* postdiction. This form of recognition is rarely considered in the literature, but is important for control and planning [15], and accounts for perceptual illusions in multiple modalities [41]. In addition, learning to infer is biologically plausible. We demonstrate in experiments that the proposed scheme reproduces interesting perceptual phenomena, including the auditory continuity illusion [8, 32], and positional smoothing associated with the flash-lag effect in vision [30, 34]. We also test its performance on tracking the hidden state of a nonlinear dynamical system from noisy and occluded observations.

## 2 Background

We describe the generic internal model of sequential observations considered in this work and, briefly, the distributed distributional code (DDC) [40, 44] used to perform recognition.

### 2.1 A generic internal model of the world for recognition

We make weak assumptions that the brain has a discrete-time internal model of the dynamic world which is stationary, Markovian and easy to samples. The latent transition dynamics and observation emission take a generic form as

$$\boldsymbol{z}_t = f(\boldsymbol{z}_{t\text{-}1}, \zeta_t^{(z)}) \tag{1a}$$

$$\boldsymbol{x}_t = g(\boldsymbol{z}_t, \zeta_t^{(x)}) \tag{1b}$$

where $f$ and $g$ are arbitrary functions that transform the conditioning variables and noise terms $\zeta_t^{(\cdot)}$ no explicit parametric assumptions are needed. Unlike the inferential sampling approach[22, 36] in which *posterior* samples must be drawn in real time, we require only off-line sampling in the generative process for learning to infer, as shown later in Section 3.2.

### 2.2 Distributed distributional code as a representation of uncertainty

Building on previous work [40, 20, 48], Vértes and Sahani [44] introduced the DDC for inference on generic hierarchical probabilistic generative models[1]. In the DDC framework, neurons with nonlinear tuning functions $\{\gamma(\boldsymbol{z})\}_{k=1}^{K_{\boldsymbol{\gamma}}}$, encode a random variable $\boldsymbol{Z} \sim q(\boldsymbol{z})$, as the expectation of these tuning functions:

$$r_{\boldsymbol{Z}}^{(k)} := \mathbb{E}_q[\gamma_k(\boldsymbol{Z})], k \in \{1, 2, ..., K_{\boldsymbol{\gamma}}\} \tag{2}$$

Under the maximum entropy principle, the DDC of $q(\boldsymbol{z})$ is the mean parameter of a generic exponential family distribution with $\boldsymbol{\gamma}(\boldsymbol{z})$ being the sufficient statistics.

Given the DDC $\boldsymbol{r}_{\boldsymbol{Z}}$ with associated tuning functions, the expectation of any function in the span of $\{\gamma_k(\cdot)\}$ can be read out from $\boldsymbol{r}$ with linear weights. More generally, the functions $\{\gamma(\boldsymbol{z})\}$ may be used as an approximating basis, so that

$$g(\boldsymbol{z}) \approx \sum_{k=1}^{K_{\boldsymbol{\gamma}}} \alpha_k \gamma_k(\boldsymbol{z}) = \boldsymbol{\alpha} \cdot \boldsymbol{\gamma}(\boldsymbol{z}) \quad \Rightarrow \quad \mathbb{E}_{q(\boldsymbol{z})}[g(\boldsymbol{z})] \approx \sum_{k}^{K_{\boldsymbol{\gamma}}} \alpha_k r_{\boldsymbol{Z}}^{(k)} \tag{3}$$

This holds by the linearity of expectation. Thus, as long as $\boldsymbol{\gamma}(\cdot)$ forms a rich enough set of basis functions, the mean rates $\boldsymbol{r}$ can be used to linearly approximate expectations of a large family of functions on $\boldsymbol{z}$, which can be useful for downstream computation. In the limit of an infinite number of neurons, $\boldsymbol{r}$ uniquely identifies any distribution given an appropriate $\boldsymbol{\gamma}(\cdot)$, and $\boldsymbol{r}$ is known as the kernel mean embedding [43].

### 2.3 DDC Recognition

Let the internal generative model of the brain have joint p.d.f. $p(\boldsymbol{z}, \boldsymbol{x})$, where $\boldsymbol{z}$ are latent and $\boldsymbol{x}$ are observed. Recognition is the process of finding the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$ for a given $\boldsymbol{x}$. The DDC for this conditional distribution is $\boldsymbol{r}_{\boldsymbol{Z}|\boldsymbol{x}} := \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{\gamma}(\boldsymbol{z})]$, a conditional expectation. This leads to the idea of training a recognition model to predict $\boldsymbol{\gamma}(\boldsymbol{z})$ from $\boldsymbol{x}$ using samples $\{\boldsymbol{z}^{(s)}, \boldsymbol{x}^{(s)}\} \sim p$. Let the recognition model take the form $\boldsymbol{W}\boldsymbol{\sigma}(\boldsymbol{x})$ with $\boldsymbol{\sigma}(\cdot)$ being a $K_{\boldsymbol{\sigma}}$-dimensional random but fixed nonlinear function. If the network is trained under the mean squared error (MSE)

$$\mathcal{L}(\boldsymbol{W}) = \mathbb{E}_{p(\boldsymbol{z},\boldsymbol{x})}\left[\|\boldsymbol{W}\boldsymbol{\sigma}(\boldsymbol{x}) - \boldsymbol{\gamma}(\boldsymbol{z})\|_2^2\right] \tag{4}$$

and provided that $\boldsymbol{\sigma}(\cdot)$ is rich enough for $\boldsymbol{W}$ to reach the minimum of (4), $\boldsymbol{W}\boldsymbol{\sigma}(\boldsymbol{x})$ produces the conditional expectation of the target variable $\boldsymbol{\gamma}(\boldsymbol{z})$, which is the DDC of $\boldsymbol{Z}|\boldsymbol{x}$ by definition.

$$\boldsymbol{r}_{\boldsymbol{Z}|\boldsymbol{x}} = \boldsymbol{W}^*\boldsymbol{\sigma}(\boldsymbol{x}), \quad \boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) \tag{5}$$

As argued in [44], minimizing (4) corresponds to minimizing of the average Kullback-Leibler divergence $KL[p(\boldsymbol{z}|\boldsymbol{x})\|q(\boldsymbol{z}|\boldsymbol{x})]$, where $q(\boldsymbol{z}|\boldsymbol{x})$ is in the exponential family with sufficient statistics $\boldsymbol{\gamma}(\boldsymbol{z})$. We interpret $\boldsymbol{\sigma}(\boldsymbol{x})$ as features of $\boldsymbol{x}$ extracted from upstream sensory areas, coding a belief about a deterministic $\boldsymbol{x}$ using DDC with $\boldsymbol{\sigma}(\boldsymbol{x})$ as the associated tuning functions.

What is attractive biologically about this approach is that $\boldsymbol{W}$ can be learned using the delta rule, as long as the brain is able to draw samples according to the correct internal model.

$$\boldsymbol{W} \leftarrow \epsilon \left[\boldsymbol{W}\boldsymbol{\sigma}(\boldsymbol{x}^{(s)}) - \boldsymbol{\gamma}(\boldsymbol{z}^{(s)})\right] \boldsymbol{\sigma}(\boldsymbol{x}^{(s)})^\intercal \quad \boldsymbol{z}^{(s)}, \boldsymbol{x}^{(s)} \sim p(\boldsymbol{z}, \boldsymbol{x}) \tag{6}$$

where $\epsilon$ is a small learning rate.

It is worth noting that the posterior *distribution function* $q(\boldsymbol{z}|\boldsymbol{x})$ is only implicitly specified by $\boldsymbol{r}_{\boldsymbol{Z}|\boldsymbol{x}}$ and the associated tuning functions $\boldsymbol{\gamma}(\cdot)$, which are sufficient for many computations that depend on posterior expectations (using (3)). This implied distribution is, however, only approximate because the minimum of (4) may not be reachable for any $\boldsymbol{W}$ given a finite dimensional $\boldsymbol{\sigma}(\cdot)$.

[1] Authors in [44] assumed that the generative model is in the exponential family for learning parameters.

## 3 DDC online recognition in dynamic environment

In this section we develop online recognition with the DDC, thereby extending the inference step from the deep hierarchical setting [44].

### 3.1 Temporally extended encoding functions

Models of online recognition usually seek to obtain the filtering marginal $p(\boldsymbol{z}_t|\boldsymbol{x}_{1:t})$ [11, 42] or the pairwise joint $p(\boldsymbol{z}_{t\text{-}1}, \boldsymbol{z}_t|\boldsymbol{x}_{1:t})$ [31]. We take a more extensive approach and formulate recognition as online updating of posterior beliefs about *all* the latent variables $\boldsymbol{z}_{1:t}$ given each new observation $\boldsymbol{x}_t$. To represent such distributions in DDC, we introduce neurons with temporally extended encoding functions $\boldsymbol{\psi}_t := \boldsymbol{\psi}(\boldsymbol{z}_{1:t})$, defined by a recurrence relationship encapsulated in a function $k$: $\boldsymbol{\psi}_t = k(\boldsymbol{\psi}_{t-1}, \boldsymbol{z}_t)$. In particular, we choose

$$\boldsymbol{\psi}_t = k(\boldsymbol{\psi}_{t-1}, \boldsymbol{z}_t) = \boldsymbol{U}\boldsymbol{\psi}_{t-1} + [\boldsymbol{\gamma}(\boldsymbol{z}_t); \boldsymbol{0}], \quad \|\boldsymbol{U}\|_2 < 1 \tag{7}$$

where $\boldsymbol{\gamma}(\boldsymbol{z}_t) \in \mathbb{R}^{K_{\boldsymbol{\gamma}}}$ is a static feature of $\boldsymbol{z}_t$ as in (2), and $\boldsymbol{U}$ is a $K_{\boldsymbol{\psi}} \times K_{\boldsymbol{\psi}}$ random projection matrix that has maximum singular value less than 1.0 to ensure stability. $\boldsymbol{\gamma}(\boldsymbol{z}_t)$ may be lower dimensional than $\boldsymbol{\psi}$ and only feeds into a subset of neurons. $\boldsymbol{\psi}_t$ is then capable of encoding a joint posterior of the history up to time $t$ through a DDC $\boldsymbol{r}_t := \mathbb{E}_{q(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})}[\boldsymbol{\psi}_t]$. If $\boldsymbol{\psi}_t$ depends only on $\boldsymbol{z}_t$ ($\boldsymbol{U} = \boldsymbol{0}$), then the corresponding DDC represents the conventional filtering distribution. Of course, the brain is unlikely to have perfect memory for the entire sequence, and for a finite number of encoding functions we expect the information about the past to decay with time. Trivially, with a permutation form of $\boldsymbol{U}$, the system can have perfect memory for at least $K_{\boldsymbol{\psi}}/K_{\boldsymbol{\gamma}}$

### 3.2 Learning to do inference on state space models

The goal of recognition is then to compute recursively in real time, combining $\boldsymbol{r}_{t\text{-}1}$ and a new observation $\boldsymbol{x}_t$. Using the idea introduced in Section 2, one can train a recognition network similar to (4) by supervised learning to compute this posterior mean. The cost function in this context is

$$\mathcal{L}^f(\boldsymbol{W}) = \mathbb{E}_{q(\boldsymbol{z}_{1:t}, \boldsymbol{x}_t|\boldsymbol{x}_{1:t\text{-}1})}\left[\|\boldsymbol{W}\boldsymbol{\sigma}(\boldsymbol{x}_t) - \boldsymbol{\psi}(\boldsymbol{z}_{1:t})\|_2^2\right] \tag{8}$$

However, the difficulty here compared to (4) is that the expectation is taken over the current posterior $q$ which is in general not easy to sample [27]. Using the expectation approximation property (3), we show in Appendix A that the optimal recognition parameter $\boldsymbol{W}$ depends on $\boldsymbol{r}_t$ in a complicated way. Instead, we consider optimizing a slightly different loss using a function $\boldsymbol{h}_{\boldsymbol{W}}(\boldsymbol{r}_{\boldsymbol{Z}_{1:t\text{-}1}|x_{1:t\text{-}1}}, \boldsymbol{x}_t)$ that has fixed recogntion parameters.

$$\tilde{\mathcal{L}^f}(\boldsymbol{W}) = \mathbb{E}_{q(\boldsymbol{z}_{1:t}, \boldsymbol{x}_t, \boldsymbol{x}_{1:t\text{-}1})}\left[\|\boldsymbol{h}_{\boldsymbol{W}}(\boldsymbol{r}_{t\text{-}1}, \boldsymbol{x}_t) - \boldsymbol{\psi}(\boldsymbol{z}_{1:t})\|_2^2\right] = \mathbb{E}_{p(\boldsymbol{x}_{1:t\text{-}1})}\left[\mathcal{L}^f(\boldsymbol{W})\right] \tag{9}$$

where $\boldsymbol{r}_{t\text{-}1}$ depends on $\boldsymbol{x}_{1:t\text{-}1}$ through recursive filtering. Instead of evaluating the MSE for a given past trajectory as in (8), $\tilde{\mathcal{L}^f}$ is the average MSE over all possible past trajectories. Given that there exists a minimum in (8), the minimum of (9) also exists for most well-behaved distributions on $\boldsymbol{x}_{1:t\text{-}1}$, and is attained if $\boldsymbol{W}$ minimizes (8) on all possible $\boldsymbol{x}_{1:t\text{-}1}$.

To allow biologically plausible learning of $\boldsymbol{W}$, we consider two simple forms of $\boldsymbol{h}_{\boldsymbol{W}}$:

$$\text{bilinear:} \quad \boldsymbol{h}^{bil}_{\boldsymbol{W}}(\boldsymbol{r}_{t\text{-}1}, \boldsymbol{x}_t) = \boldsymbol{W}(\boldsymbol{r}_{t\text{-}1} \otimes \boldsymbol{\sigma}(\boldsymbol{x}_t)) \tag{10}$$

$$\text{linear:} \quad \boldsymbol{h}^{lin}_{\boldsymbol{W}}(\boldsymbol{r}_{t\text{-}1}, \boldsymbol{x}_t) = \boldsymbol{W}[\boldsymbol{r}_{t\text{-}1}; \boldsymbol{\sigma}(\boldsymbol{x}_t)] \tag{11}$$

where $\otimes$ indicates the Kronecker product. That is, $\boldsymbol{h}^{bil}_{\boldsymbol{\mathsf{W}}}$ computes $\boldsymbol{r}_t$ linearly from the outer product of $\boldsymbol{r}_{t\text{-}1}$ and $\boldsymbol{\sigma}(\boldsymbol{x}_t)$, and $\boldsymbol{h}^{lin}_{\boldsymbol{W}}$ does so on the concatenation of the two (we discuss more about the bilinear form in Appendix B). These two choices allow $\boldsymbol{W}$ to be trained by the biologically plausible delta rule, using training examples of $\{\boldsymbol{r}_{1:t\text{-}1}, \boldsymbol{z}_t, \boldsymbol{x}_t\}$. These triplets can be obtained by simulating the generative model; training examples of $\boldsymbol{r}_{1:t\text{-}1}$ are bootstrapped by $\boldsymbol{h}_{\boldsymbol{W}}$ also on the simulated sequences, with the initial $\boldsymbol{r}_1 = \mathbb{E}_{q(\boldsymbol{z}_1|\boldsymbol{x}_0)}[\boldsymbol{\psi}(\boldsymbol{z}_1)]$ obtained using (6) and (5), but it has decaying influence on $\boldsymbol{W}$ due to the Markov property. If this process converges, recognition simply involves linear/bilinear operations on $\boldsymbol{r}_{t\text{-}1}$ and $\boldsymbol{\sigma}(\boldsymbol{x}_t)$.

Once we have $\boldsymbol{r}_t$, posterior information is accessible in the same way as before. Specifically, postdictive information that is an expectation of some function $l(\boldsymbol{z}_{t\text{-}\tau})$ can be read out in essentially the same way as Equation (3).

$$\mathbb{E}_{q(\boldsymbol{z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t})}[l(\boldsymbol{z}_{t\text{-}\tau})] \approx \boldsymbol{\alpha}^{\intercal}_{\boldsymbol{\psi}_t \to l} \cdot \boldsymbol{r}_t \quad \text{where} \quad \boldsymbol{\alpha}^{\intercal}_{\boldsymbol{\psi}_t \to l}\boldsymbol{\psi}(\boldsymbol{z}_{1:t}) \approx l(\boldsymbol{z}_{t\text{-}\tau}) \tag{12}$$

# 4 Experiments

In this section, we demonstrate the effectiveness of the proposed recognition method on biologically relevant simulations. For each experiments, we trained the DDC filter offline until it has learned the internal model, and ran recognition using fixed parameters $\boldsymbol{W}$ and read-out weights $\alpha$. Details of the experimental setup common to all experiments are described in Appendix C.

## 4.1 Auditory continuity illusions

In the auditory continuity illusion, the percept of a complex sound may be altered by subsequent acoustic signals. Two tone pulses separated by a silent gap are perceived to be discontinuous; however, when the gap is filled by sufficiently loud wide-band noise, listeners often report illusory continuation of the tone through the noise. This illusion is reduced if the second tone begins after a slight delay, even though the acoustic stimulus in the two cases is identical until noise offset [8, 32].

To model the essential elements of this phenomenon, we assume that the brain has a simple internal model for tone and noise stimuli described by (15) in Appendix C.1, with a binary Markov chain describing the onsets and offsets of tone and wide-band noise, and noisy observations of power in three frequency bands.

We run six different experiments after the DDC has learned to do inference on the internal model, and show in Figure 1 the marginal posterior distributions of the perceived tone level in the past at time $t$-$\tau$ for each stimulus presentation at time (t), using the present time DDC $\boldsymbol{r}_t$. In Figure 1A, when a clear tone is presented, the model perceives the correct level and localizes the tone period well until the end. Figure 1B and C show the continuity illusion. As the noise turns on, the belief about the tone's level decreases gradually and is uncertain about the two lower levels. When the noise turns off, an immediately following tone raises the inferred tone level during the noise period. By contrast, a gap between the noise and the second tone immediately reduces this perceived tone level.

We tested the model on three additional sound configurations and saw interesting behavior. In Figure 1D, the tone has a higher level than in Figure 1A-C. If the noise has slightly lower spectral density than the tone, the model believes that the tone might have been interrupted, but is uncertain. If this noise level is much lower (Figure 1E), no illusory tone is perceived. In the final experiment (Figure 1F), the model predicts that no continuity is perceived if the first tone is softer than the noise but second tone is louder.

## 4.2 Smoothing in the flash-lag effects with direction reversal

In the previous experiment, the internal model correctly describes the statistics of the stimuli. It is known that a mismatch of the internal model to the real world, such as human's slowness preference [41], can induce perceptual illusions. Here, we use our framework to model the flash-lag effect, although the same principle can also be used directly for the cutaneous rabbit effect in somatosensation [18].

In the flash-lag effect, an object A moves in a straight line and passes by another hidden object B, and B is briefly flashed at $t = 0$ when A is aligned with it; however, subjects perceive the flashed B to be lagging behind A [30, 34]. One early explanation is the extrapolation model [34]: viewers extrapolate the movement of A and report its predicted position when B is flashed. On the other hand, the latency difference model [38] assumes that the perception of flash is delayed by $t_0$ compared to the perception of A. However, neither explanation can account for another related finding: if the moving A suddenly switches direction and B is flashed at several offsets around the reversal position (but still aligned with A), the reported location of A for different flash locations of B form a smooth trajectory (Figure 2A, dots), instead of the broken line predicted by the extrapolation model, or the simple shift in time predicted by the latency difference model [46].

Rao, Eagleman, and Sejnowski [39] suggested that the lag itself might be due to the signal propagation delay as in the latency difference model, but the smoothing effect could be caused by an additional processing delay. After perceiving the flash at $t_0$, the brain takes an additional time $\tau$ to process the flash and to estimate the location of A. Crucially, subjects integrate the visible trajectory of A in this period to *postdict* the position of A at $t_0$, the perceived time of flash. This integration is based on an

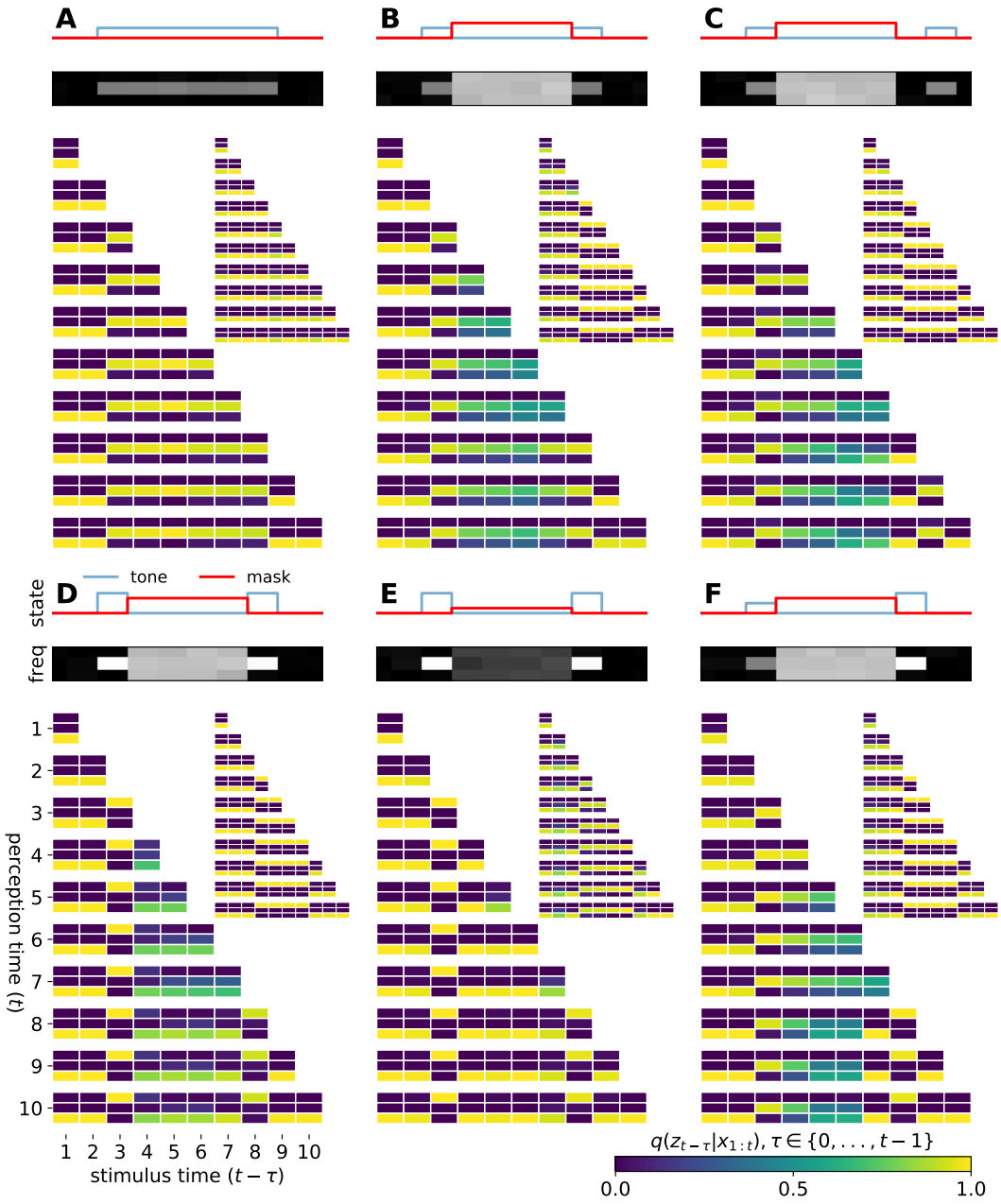


Figure 1: Modelling the auditory continuity illusion, showing decoded marginal distribution for perceived over tone level in the past ($t$-$\tau$) at each stimulus presentation time ($t$). There are six separate experiments marked form A to F. For each experiment, the top panel shows the true levels of the tone and mask; middle panel shows the spectrogram observation. In the lower panel, we show real-time posterior marginal probabilities of the tone $q(\boldsymbol{z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}), \tau \in \{0, \ldots, t\text{-}1\}$ at each stimulus presentation time ($t$), shown as horizontal "buses". At each perception time, color of each rectangle shows marginal posterior probabilities decoded from DDC using maximum entropy. The three levels indicates the tone level. Each inset shows the marginal probabilities for the mask levels.

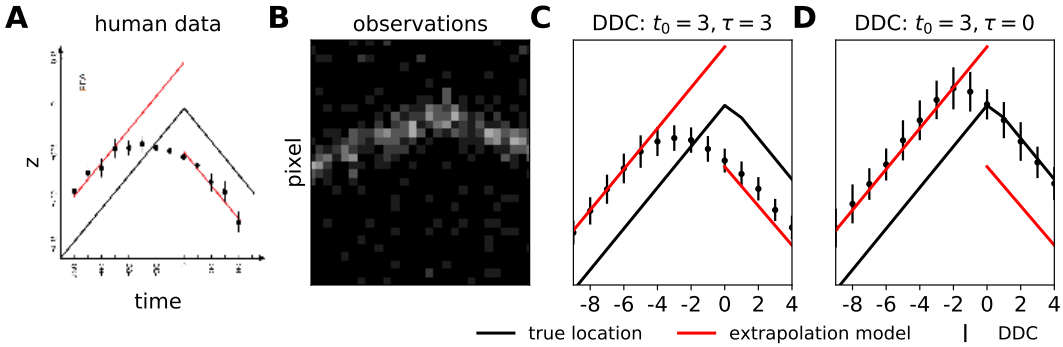


Figure 2: Modelling the flash-lag effect. black line shows the true trajectory of the moving object. red line shows the prediction of the extrapolation model. gray line shows the true posterior mean without delay. A, human data from [46]. B, the observation used in our simulation. C, DDC recognition using $\tau = 3$ future observations to postdict position at $t_0 = 3$ after the time of flash D, DDC recognition without posdition. (C,D), black dot with errorbar shows the mean and std of posterior mean estimates from 10 runs.

internal model preferring slow movements. The authors then used the Kalman smoother to reproduce the behavioral results.

Here, we apply the same idea but use a slightly more realistic observation. Details of the internal model are described in Appendix C.2. In short, the unobserved true dynamics is linear Gaussian with additive Gaussian noise, and the observation is an 1-D image of the true position with Poisson noise (Figure 2B). When the noise in the dynamics is small, establishing preference for slow movements, DDC recognition reproduces human data. Specifically, the shape of the curve is captured by taking future observations into account. Without postdiction ($\tau = 0$), the reported location tends to overshoot as also noted in [39].

### 4.3 Noisy and occluded tracking

In the auditory illusion example, if the tone frequency progresses with some pattern (e.g. ramping up), the illusory tone is usually heard to continue the same pattern during the noise. This is similar to tracking under noisy and occluded observations. It is possible to integrate the target's trajectory back in time to refine the perceived position during its disappearance; particular visitations in space may be important for planning and control, especially in multi-agent environments [2].

To test the recognition model, we create a stochastic oscillatory dynamics (Figure 3A) observed through a 1-D image with additive Gaussian noise and occlusion (details in Appendix C.3). An example is shown in Figure 3(A). We ran a simple bootstrap particle filter (PF) as benchmark Figure 3(B).

The results of DDC recognition for a particular sequence of $\boldsymbol{x}_{1:t}$ are shown in Figure 3(C-F). The single-step marginal histograms are obtained by projecting $\boldsymbol{r}_t$ onto a set of bin functions using (12). (maximum entropy decoding is less smooth, see Figure 6 in Appendix C.3). The decoded histogram from $\boldsymbol{r}_t$ is expected to be more noisy due to the non-smoothness of the bin functions, but it shows interesting temporal integration. Using the $R^2$ for predicting the true latent location as a measure of performance, the purely forward ($\tau = 0$) posterior mean is comparable to that of the particle filter. As we increase $\tau$, the number of future observations, we see not only an increase in $R^2$, but important changes in distribution. In the occluded regions, the posterior mass becomes more concentrated as $\tau$ increases, particularly towards the end of occlusions, as the result of including future observations. In addition, bimodalality is observed during some occlusion intervals, reflecting the nonlinearity in the true latent process.

Posterior widths tend to increase slightly for larger $\tau$, implying increased uncertainty about the distant past. As we used $K_{\boldsymbol{\psi}}/K_{\boldsymbol{\gamma}} = 5$, distortion in the encoded distribution is expected for $\tau > 5$. How the encoded distribution of the past changes over time can depend on the form of $k$ and $\boldsymbol{\gamma}$ in (7) and properties of the internal model, which we shall investigate in the future.

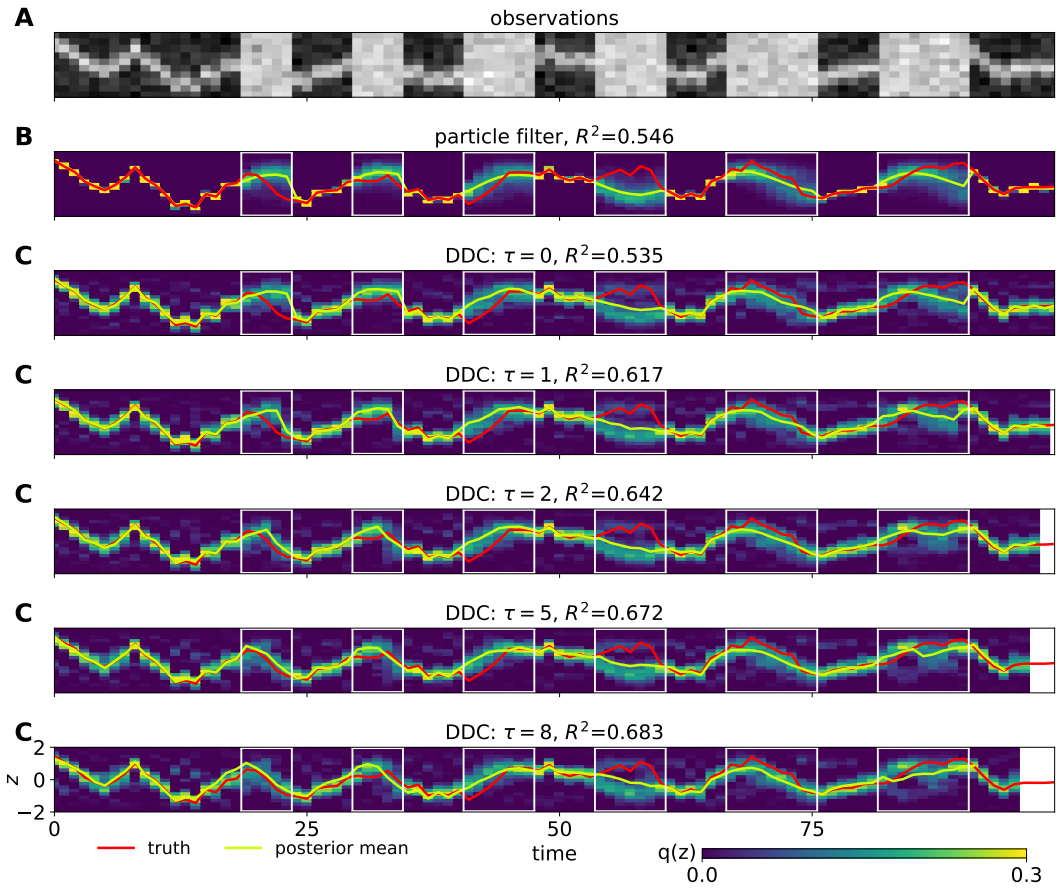


Figure 3: Noisy and occluded tracking. A, the 1-D image observation. B, posterior mean and mariginals estimated using a particle filter C-F, posterior marginals for the location at $t$-$\tau$ perceived at time $t$.

## 5 Conclusion and future work

We provided a biologically plausible online recognition framework based on the distributed distributional codes. It is able to represent and compute with a very rich class of distributions, makes weak assumptions on the internal model, and online inference is accurate and efficient. Using neurons with temporally extended encoding functions, new observations are used to update percepts of latent variables in the past without the need for explicit backward inference. The proposed method reproduces behavior results of perceptual illusions involving postdiction, a common form of inference in biological perception.

The key element in this framework is temporally extended encoding. The postdictive readout weights need to be trained while the history information is still available. But a good memory capacity may affect how well current information is held at the present time and also in the future. How this interplay affect perception needs further analysis.

The posterior of latent variable in the past refined by extra future observations better reflects of real dynamics, and can potentially help adjusting the internal model according to new statistics of observations; it would be interesting to see the impact of postdiction on adaptation.

## References

[1] David Alais and David Burr. “The ventriloquist effect results from near-optimal bimodal integration”. In: *Current biology* 14.3 (2004), pp. 257–262.

[2] Francesco Amigoni and Marco Somalvico. “Multiagent systems for environmental perception”. In: *Proceedings of the “Third AMS (American Meteorological Society) Conference on Artificial Intelligence Applications to Environmental Science”, Preprints CD-ROM and Abstract Volume*. 2003, p. 487.

[3] Peter W Battaglia, Robert A Jacobs, and Richard N Aslin. “Bayesian integration of visual and auditory signals for spatial localization”. In: *Josa a* 20.7 (2003), pp. 1391–1397.

[4] Jeff Beck, Alexandre Pouget, and Katherine A Heller. “Complex inference in neural circuits with probabilistic population codes and topic models”. In: *Advances in neural information processing systems*. 2012, pp. 3059–3067.

[5] Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. “Marginalization in neural circuits with divisive normalization”. In: *Journal of Neuroscience* 31.43 (2011), pp. 15310–15319.

[6] J Beck et al. “Probabilistic population codes and the exponential family of distributions”. In: *Progress in brain research* 165 (2007), pp. 509–519.

[7] Ulrik Beierholm et al. “Comparing Bayesian models for multisensory cue combination without mandatory integration”. In: *Advances in neural information processing systems*. 2008, pp. 81–88.

[8] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[9] Anne K Churchland et al. “Variance as a signature of neural computations during decision making”. In: *Neuron* 69.4 (2011), pp. 818–831.

[10] Mark M Churchland et al. “Stimulus onset quenches neural variability: a widespread cortical phenomenon”. In: *Nature neuroscience* 13.3 (2010), p. 369.

[11] Sophie Deneve, Jean-René Duhamel, and Alexandre Pouget. “Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of Kalman filters”. In: *Journal of Neuroscience* 27.21 (2007), pp. 5744–5756.

[12] David M. Eagleman and Terrence J. Sejnowski. “Motion integration and postdiction in visual awareness”. In: *Science* 287.5460 (2000), pp. 2036–2038.

[13] Chris Eliasmith and Charles H Anderson. *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press, 2004.

[14] A. Emin Orhan and Wei Ji Ma. “Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback”. In: *Nature Communications* 8 (2017), p. 138.

[15] Manfred Eppe and Mehul Bhatt. “Narrative based postdictive reasoning for cognitive robotics”. In: *arXiv preprint arXiv:1306.0665* (2013).

[16] Marc O Ernst and Martin S Banks. “Humans integrate visual and haptic information in a statistically optimal fashion”. In: *Nature* 415.6870 (2002), p. 429.

[17] Akihiro Funamizu, Bernd Kuhn, and Kenji Doya. “Neural substrate of dynamic Bayesian inference in the cerebral cortex”. In: *Nature neuroscience* 19.12 (2016), p. 1682.

[18] Frank A Geldard and Carl E Sherrick. “The cutaneous" rabbit": a perceptual illusion”. In: *Science* 178.4057 (1972), pp. 178–179.

[19] Steffen Grünewälder et al. “Conditional mean embeddings as regressors-supplementary”. In: *arXiv preprint arXiv:1205.4656* (2012).

[20] Geoffrey E. Hinton et al. “The "wake-sleep" algorithm for unsupervised neural networks”. In: *Science* 268.5214 (1995), pp. 1158–1161.

[21] hoon choi hoon and brian j. scholl brian j. “perceiving causality after the fact: postdiction in the temporal dynamics of causal perception”. In: *Perception* 35.3 (2006), pp. 385–399.

[22] Patrik O Hoyer and Aapo Hyvärinen. “Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior”. In: *NIPS*. Ed. by S Becker, S Thrun, and K Obermayer. 2003, pp. 293–300.

[23] Konrad P. Körding, Shih-pi Ku, and Daniel M. Wolpert. “Bayesian Integration in Force Estimation”. In: *Journal of Neurophysiology* 92.5 (2004), pp. 3161–3165. URL: `http://www.physiology.org/doi/10.1152/jn.00275.2004`.

[24] Konrad P Körding et al. “Causal Inference in Multisensory Perception”. In: *PLOS ONE* 2.9 (2007), pp. 1–10.

[25] Anna Kutschireiter et al. “Nonlinear Bayesian filtering and learning: A neuronal dynamics for perception”. In: *Scientific Reports* 7 (2017).

[26] Robert Legenstein and Wolfgang Maass. "Ensembles of Spiking Neurons with Noise Support Optimal Probabilistic Inference in a Dynamically Changing Environment". In: *PLoS Computational Biology* 10 (2014), e1003859.

[27] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. "Maximum entropy flow networks". In: *arXiv preprint arXiv:1701.03504* (2017).

[28] Wei Ji Ma et al. "Bayesian inference with probabilistic population codes". In: *Nature Neuroscience* 9.11 (2006), pp. 1432–1438. ISSN: 1097-6256.

[29] Wei Ji Ma et al. "Behavior and neural basis of near-optimal visual search." In: 14 (2011).

[30] Donald M Mackay. "Perceptual stability of a stroboscopically lit visual field containing self-luminous objects". In: *Nature* 181.4607 (1958), p. 507.

[31] Joseph G Makin, Benjamin K Dichter, and Philip N Sabes. "Learning to estimate dynamical state with probabilistic population codes". In: *PLoS computational biology* 11.11 (2015), e1004554.

[32] George A Miller and Joseph CR Licklider. "The intelligibility of interrupted speech". In: *The Journal of the Acoustical Society of America* 22.2 (1950), pp. 167–173.

[33] Yalda Mohsenzadeh, Suryadeep Dash, and J Douglas Crawford. "A state space model for spatial updating of remembered visual targets during eye movements". In: *Frontiers in systems neuroscience* 10 (2016), p. 39.

[34] Romi Nijhawan. "Motion extrapolation in catching." In: *Nature* (1994).

[35] Gergő Orbán and Daniel M Wolpert. "Representations of uncertainty in sensorimotor control". In: *Current opinion in neurobiology* 21.4 (2011), pp. 629–635.

[36] Gergő Orbán et al. "Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex". In: *Neuron* 92.2 (2016), pp. 530–543.

[37] Ivan V Oseledets. "Tensor-train decomposition". In: *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2295–2317.

[38] Gopathy Purushothaman et al. "Moving ahead through differential visual latency". In: *Nature* 396.6710 (1998), p. 424.

[39] Rajesh PN Rao, David M Eagleman, and Terrence J Sejnowski. "Optimal smoothing in visual motion perception". In: *Neural Computation* 13.6 (2001), pp. 1243–1253.

[40] Maneesh Sahani and Peter Dayan. "Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity". In: *Neural Computation* 15.10 (2003), pp. 2255–2279.

[41] Shinsuke Shimojo. "Postdiction: its implications on visual awareness, hindsight, and sense of agency". In: *Frontiers in psychology* 5 (2014), p. 196.

[42] Sacha Sokoloski. "Implementing a bayes filter in a neural circuit: The case of unknown stimulus dynamics". In: *Neural computation* 29.9 (2017), pp. 2450–2490.

[43] Le Song, Kenji Fukumizu, and Arthur Gretton. "Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models". In: *IEEE Signal Processing Magazine* 30.4 (2013), pp. 98–111.

[44] Eszter Vértes and Maneesh Sahani. "Flexible and accurate inference and learning for deep generative models". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 4166–4175.

[45] Louise Whiteley and Maneesh Sahani. "Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes". In: *Journal of Vision* 8.3 (2008), p. 2.

[46] David Whitney and Ikuya Murakami. "Latency difference, not spatial extrapolation". In: *Nature neuroscience* 1.8 (1998), p. 656.

[47] Jean-Jacques Orban de Xivry et al. "Kalman filtering naturally accounts for visually guided and predictive smooth pursuit dynamics". In: *Journal of Neuroscience* 33.44 (2013), pp. 17301–17313.

[48] Richard S. Zemel, Peter Dayan, and Alexandre Pouget. "Probabilistic Interpretation of Population Codes". In: *Neural Computation* 10.2 (1998), pp. 403–430.

# A neurally plausible model for online recognition and postdiction: Supplementary material

## A Formal solution to the filtering loss

We show the formal solution to minimizing (8) before discussing its biological implications.

**Proposition 1.** *Given a DDC of previous belief* $\boldsymbol{r}_{t\text{-}1}$*,* $\boldsymbol{W}_{\boldsymbol{r}_{t\text{-}1}}$ *below is the minimizer of* (8)

$$\mathcal{L}^f(\boldsymbol{W}) = \mathbb{E}_{q(\boldsymbol{z}_{1:t}, \boldsymbol{x}_t | \boldsymbol{x}_{1:t\text{-}1})} \left[ \| \boldsymbol{W} \boldsymbol{\sigma}(\boldsymbol{x}_t) - \boldsymbol{\psi}(\boldsymbol{z}_{1:t}) \|_2^2 \right] \tag{8 revisited}$$

$$\begin{aligned} \boldsymbol{W}_{\boldsymbol{r}_{t\text{-}1}} &= \boldsymbol{C}_{\boldsymbol{Z}_{1:t}, \boldsymbol{X}_t | \boldsymbol{x}_{1:t-1}} \boldsymbol{C}^{-1}_{\boldsymbol{X}_t, \boldsymbol{X}_t | \boldsymbol{x}_{1:t\text{-}1}} \\ \boldsymbol{C}_{\boldsymbol{Z}_{1:t}, \boldsymbol{X}_t | \boldsymbol{x}_{1:t\text{-}1}} &= \mathbf{C}_{\boldsymbol{Z}_{1:t}, \boldsymbol{X}_t | \boldsymbol{Z}_{t\text{-}1}} \boldsymbol{r}_{t\text{-}1} \quad \boldsymbol{C}_{\boldsymbol{X}_t, \boldsymbol{X}_t | \boldsymbol{x}_{1:t\text{-}1}} = \mathbf{C}_{\boldsymbol{X}_t, \boldsymbol{X}_t | \boldsymbol{Z}_{t\text{-}1}} \boldsymbol{r}_{t\text{-}1} \\ \mathbf{C}_{\boldsymbol{Z}_{1:t}, \boldsymbol{X}_t | \boldsymbol{Z}_{t\text{-}1}} &= \arg\min_{\mathbf{C}} \mathbb{E}_{p(\boldsymbol{z}_{t\text{-}1}, \boldsymbol{z}_t, \boldsymbol{x}_t)} \| \mathbf{C} \boldsymbol{\psi}_{t\text{-}1} - \boldsymbol{\psi}(\boldsymbol{z}_{1:t}) \boldsymbol{\sigma}(\boldsymbol{x}_t)^\intercal \|_2^2 \\ \mathbf{C}_{\boldsymbol{X}_t, \boldsymbol{X}_t | \boldsymbol{Z}_{t\text{-}1}} &= \arg\min_{\mathbf{C}} \mathbb{E}_{p(\boldsymbol{z}_{t\text{-}1}, \boldsymbol{x}_t)} \| \mathbf{C} \boldsymbol{\psi}_{t\text{-}1} - \boldsymbol{\sigma}(\boldsymbol{x}_t) \boldsymbol{\sigma}(\boldsymbol{x}_t)^\intercal \|_2^2 \end{aligned} \tag{13}$$

This is similar to the kernel Bayes rule [19]. The two minimization problem are essentially computing the readout weights used to be approximated the conditional covariance matrices $\boldsymbol{C}$. Filtering in this case involves solving these two problems before taking an inverse of a correlation matrix. If one interpret the two tensor $\mathbf{C}$'s as weights, the matrix $\boldsymbol{C}$'s are readout from $\boldsymbol{r}_{t\text{-}1}$, then it is not clear how the inverse and $\boldsymbol{W}_{\boldsymbol{r}_{t\text{-}1}}$ could be implemented by neural mechanisms.

## B Approximation solution for filtering

### B.1 The bilinear approximation and the tensor train decomposition

The bilinear approximation $\boldsymbol{h}_{\boldsymbol{W}}(\boldsymbol{r}_{t\text{-}1}, \boldsymbol{x}_t)$ (10) and the corresponding solution to minimizing the MSE (9) w.r.t. $\mathbf{W}$ is connected to the tensor train decomposition (TT) [37]. The MSE is

$$\tilde{\mathcal{L}}^f(\mathbf{W}) = \mathbb{E}_{q(\boldsymbol{z}_{1:t}, \boldsymbol{x}_t, \boldsymbol{x}_{1:t\text{-}1})} \left[ \| \mathbf{W} \cdot (\boldsymbol{r}_{t\text{-}1} \otimes \boldsymbol{x}_t) - \boldsymbol{\psi}(\boldsymbol{z}_{1:t}) \|_2^2 \right] \tag{14}$$

Denote the minimizer of (14) at each $t$ by $\mathbf{W}_t^*$. Consider the situation that, at each $t$, we would like to predict $\boldsymbol{\psi}_t$ using a sequence of observations $\boldsymbol{x}_{1:t}$. Let $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{K_{\boldsymbol{\sigma}}}$ be sufficiently rich so that there exists a linear operator $\mathbf{W}_t^{(p)}$ that maps from the product space of $\boldsymbol{\sigma}(\boldsymbol{x}_1) \otimes \cdots \otimes \boldsymbol{\sigma}(\boldsymbol{x}_t)$ to $\boldsymbol{r}_t := \mathbb{E}_q(\boldsymbol{z}_{1:t} | \boldsymbol{x}_{1:t})[\boldsymbol{\psi}(\boldsymbol{z}_{1:t})]$, then $\mathbf{W}_t^{(p)}$ is an order $t+1$ tensor which is expensive to estimate. Low rank approaches may alleviate the difficulty, such as TT. In fact, the sequence of minimizers to (14)

$\{\mathbf{W}_{t'}^*\}_{t'=1}^t$ form a TT of an order $t+1$ tensor $\mathbf{W}_t^{(f)}$ with the same shape as $\mathbf{W}_t^{(p)}$. For example:

$$\boldsymbol{W}_1^* = \arg\min_{\boldsymbol{W}_1} \mathbb{E}_p \sum_i \left( \sum_j W_{1,ji}\sigma_i(\boldsymbol{x}_1) - \psi_{1,j} \right)^2 \Rightarrow r_{1,j} = \sum_{ji} \boldsymbol{W}_{1,ji}^* \sigma_i(\boldsymbol{x}_1)$$

$$\begin{aligned}
\mathbf{W}_2^* &= \arg\min_{\mathbf{W}_2} \mathbb{E}_p \sum_l \left( \sum_{jk} W_{2,lkj} r_{1,j} \sigma_k(\boldsymbol{x}_2) - \psi_{2,l} \right)^2 \\
&= \arg\min_{\mathbf{W}_2} \mathbb{E}_p \sum_l \left( \sum_{jk} W_{2,lkj} \left[ \sum_{ij} W_{1,ji}^* \sigma_i(\boldsymbol{x}_1) \right] \sigma_k(\boldsymbol{x}_2) - \psi_{2,l} \right)^2 \\
&= \arg\min_{\mathbf{W}_2} \mathbb{E}_p \sum_l \left( \sum_{ik} \underbrace{\left[ \sum_j W_{2,lkj} W_{1,ji}^* \right]}_{W_{2,lki}^{(f)}} \sigma_i(\boldsymbol{x}_1)\sigma_k(\boldsymbol{x}_2) - \psi_{2,l} \right)^2
\end{aligned}$$

the summation in the square brackets is the TT of $\mathbf{W}_2^{(f)}$. Thus, the proposed optimization for (14) finds a tensor of the same shape as $\mathbf{W}_t^{(p)}$ in the TT space sequentially, predicting a new $\boldsymbol{\psi}_t$ by joining a new core tensor with $\mathbf{W}_{t\text{-}1}^{(f)}$, and only minimize the MSE in the space of the new core tensor to get $\mathbf{W}_t^*$.

# C Experimental details

In all simulations, we assume the brain can draw samples from the internal model, and the recognition weights $\boldsymbol{W}$ and readout weights $\boldsymbol{\alpha}$ have been trained on these samples for a long time and have converged. In our experiments, this condition was achieved by closed form regression in solving least square regressions, using 10,000-20,000 sequences from the generative model, and trained the recognition and readout parameters for around 100 time steps for the SSM to enter in the stationary regime. The learned parameters are then fixed for online inference. The base tuning functions $\boldsymbol{\gamma}(\cdot)$ in (7) and input feature map $\boldsymbol{\sigma}(\cdot)$ in (10) and (11) have $\tanh$ nonlinearity after random linear projection; the weights and biases in the projection are randomly drawn from a Gaussian with variance such that the these functions are relatively smooth for the inputs they receive. Code is available at `https://dereferer.me/?ohPJ8Bqepordr-3RDRj_5r0-4j7NdEB6mg8prxwp4vw8yb23kAkzEy8nN`

## C.1 Auditory continuity illusions

### C.1.1 Model setup

The internal model is described in (15). It has a 2-D binary latent dynamics for the tone ($z_{t,0}$) and mask ($z_{t,1}$), and a 3-D noisy observation $x_{t,i}, i \in \{0,1,2\}$ for three frequency bands.

$$\begin{aligned}
&c_{t,i} \sim \text{Bernoulli}(0.1) && i \in \{0,1\} \\
&l_{t,0} \sim \text{Uniform}(\{2,4\}) && l_{t,1} \sim \text{Uniform}(\{1,3\}) \\
&z_{t,i} = \begin{cases} z_{t\text{-}1,i} & \text{if } z_{t\text{-}1,i} \neq 0.0 \text{ and } c_{t,i} = 0 \\ c_{t,i} l_{t,i} & \text{if } z_{t\text{-}1,i} = 0.0 \\ 0 & \text{if } z_{t\text{-}1,i} \neq 0.0 \text{ and } c_{t,i} = 1 \end{cases} && i \in 0,2 \\
&x_{t,1} = \max\{z_{t,0}, z_{t,1}\} + \zeta_{t,i}, && \zeta_{t,1} \sim \mathcal{N}(0, 0.1^2) \\
&x_{t,i} = z_{t,1} + \zeta_{t,i}, && \zeta_{t,i} \sim \mathcal{N}(0, 0.1^2) && i \in 0,2
\end{aligned}$$

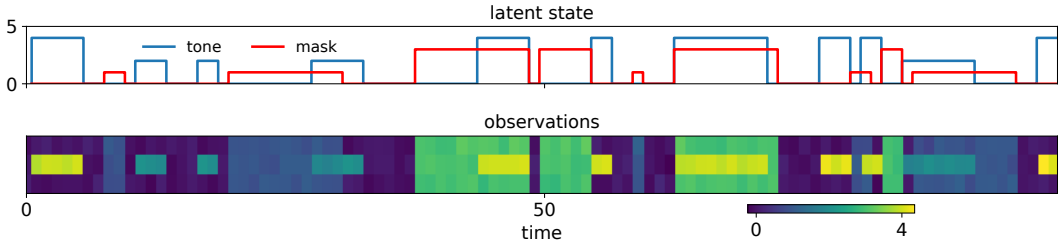


Figure 4: Example training data used for the auditory illusion experiment.

In words, the tone has energy levels $\{0, 2, 4\}$ and the mask has energy levels $\{0, 1, 3\}$. At each time step, the tone and the mask can turn on or fall off with probability 0.1. For each of the two, if it turns on, it takes one of the two non-zero levels with equal chance; but it can only fall down to 0. The middle band of reflects the greater level of the tone and the noise. The other two bands only contain the mask. All the three bands are contaminated by a small amount of i.i.d Gaussian noise. Example of the simulated data are shown in Figure 4.

In the DDC filter, we set the $K_{\boldsymbol{\psi}} = 200, K_{\boldsymbol{\gamma}} = 20$ and $K_{\boldsymbol{\sigma}} = 10$ and used the $\boldsymbol{h}^{bil}$ in (10).

### C.1.2 Additional methods and results

We showed in Figure 1 the marginal p.m.f of the inferred tone level given observations up to time the stimulus time $p(\boldsymbol{z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t})$ decoded from $\boldsymbol{r}_t$. This is done by first approximating its expectation over the static tuning function $\boldsymbol{\gamma}(\boldsymbol{z}_{t\text{-}\tau})$ (other choices are possible) using (12), obtaining $\boldsymbol{m}_{t\text{-}\tau} := \mathbb{E}_{q\boldsymbol{z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}}[\boldsymbol{Z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}]$, a DDC on $\boldsymbol{Z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}$. Using maximum entropy decoding, we can find the corresponding p.m.f. Let the discrete p.m.f be $p(\boldsymbol{z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}) = \prod_i^{|\mathcal{Z}|} p_i^{\delta(\boldsymbol{z}_{t\text{-}\tau}=z_i)}$, where $|\mathcal{Z}|$ is the cardinality of the support on $\boldsymbol{z}$ (9 in our case), and $\pi$ is the discrete probabilities that can be decoded from $\boldsymbol{r}$ and $\boldsymbol{\gamma}$ by solving the following optimization problem

$$\min_{\boldsymbol{p}} \sum_i^{|\mathcal{Z}|} p_i \log(p_i) \quad \text{s.t.} \sum_i^{|\mathcal{Z}|} p_i \gamma_j(z_i) = m_j, \sum_i^{|\mathcal{Z}|} p_i = 1, p_i \in [0, 1] \tag{16}$$

which is relatively simple for a 9-outcome (3 tone $\times$ 3 noise levels), discrete distribution.

However, the marginals do not show whether $\boldsymbol{r}_t$ has indeed encoded the *joint* distribution over tone and noise for the entire history, which is the actual distribution represented by encoding function $\boldsymbol{\psi}_t$. One can decoding the joint distribution using the same principle as (12), but this is only computationally tractable for small $t$, as the support of the joint grows exponentially with $t$. Instead, we examine samples from the joint distribution. As sampling from the mean representation directly is difficult [27], we develop a sequential sampling method suitable for our state-space model as follows.

Given a joint DDC $\boldsymbol{r}_t$, we first decode the marginal p.m.f. $p(\boldsymbol{z}_1|\boldsymbol{x}_{1:t})$ as done in the main text. A sample $\boldsymbol{z}_1^{(s)}|\boldsymbol{x}_{1:t}$ is then drawn. We then need to decode the marginal $p(\boldsymbol{z}_2|\boldsymbol{x}_{1:t}, \boldsymbol{z}_1^{(s)})$, which can be done in two ways:

1. Project $\boldsymbol{r}_t$ onto an encoding basis on $\boldsymbol{Z}_{1:2}$, and decode the maximum entropy distribution while fixing $\boldsymbol{Z}_1 = \boldsymbol{z}_1^{(s)}$ but allowing $\boldsymbol{z}_2$ to take on all possible values;
2. From $\boldsymbol{r}_t$ and $\boldsymbol{z}_1^{(s)}$, compute the DDC for $\boldsymbol{Z}_2|\boldsymbol{x}_{1:t}, \boldsymbol{z}_1^{(s)}$ associated with tuning functions $\boldsymbol{\gamma}$, then decode by maximum entropy as (16). This DDC can be found using the same method as in our filtering problem by minimizing a similar loss as (9)

$$\tilde{\mathcal{L}}^s(\boldsymbol{V}) = \mathbb{E}_{q(\boldsymbol{z}_2, \boldsymbol{z}_1, \boldsymbol{x}_{1:t})}\left[\|\boldsymbol{h}_{\boldsymbol{V}}(\boldsymbol{r}_t, \boldsymbol{z}_1) - \boldsymbol{\psi}(\boldsymbol{z}_2)\|_2^2\right] \tag{17}$$

   where $\boldsymbol{r}_t$ depends on $\boldsymbol{x}_{1:t}$, and $\boldsymbol{h}_{\boldsymbol{V}}$ can take any form not limited to (11) or (10).

Due to the Markov property, $\boldsymbol{Z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t\text{-}\tau\text{-}1}^{(s)} \overset{d}{=} \boldsymbol{Z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}, \boldsymbol{z}_{t\text{-}\tau\text{-}1}^{(s)}$. [2] We apply the second approach, draw samples sequentially and also evaluate the log likelihood for each joint sample. The results are

[2] We note that also $\boldsymbol{Z}_{t\text{-}\tau}|\boldsymbol{x}_{1:t}, \boldsymbol{z}_{t\text{-}\tau\text{-}1}^{(s)} = \boldsymbol{Z}_{t\text{-}\tau}|\boldsymbol{x}_{t\text{-}\tau:t}, \boldsymbol{z}_{t\text{-}\tau\text{-}1}^{(s)}$, suggesting a backward filtering approach in general. But here we are interested in decoding from $\boldsymbol{r}_t$

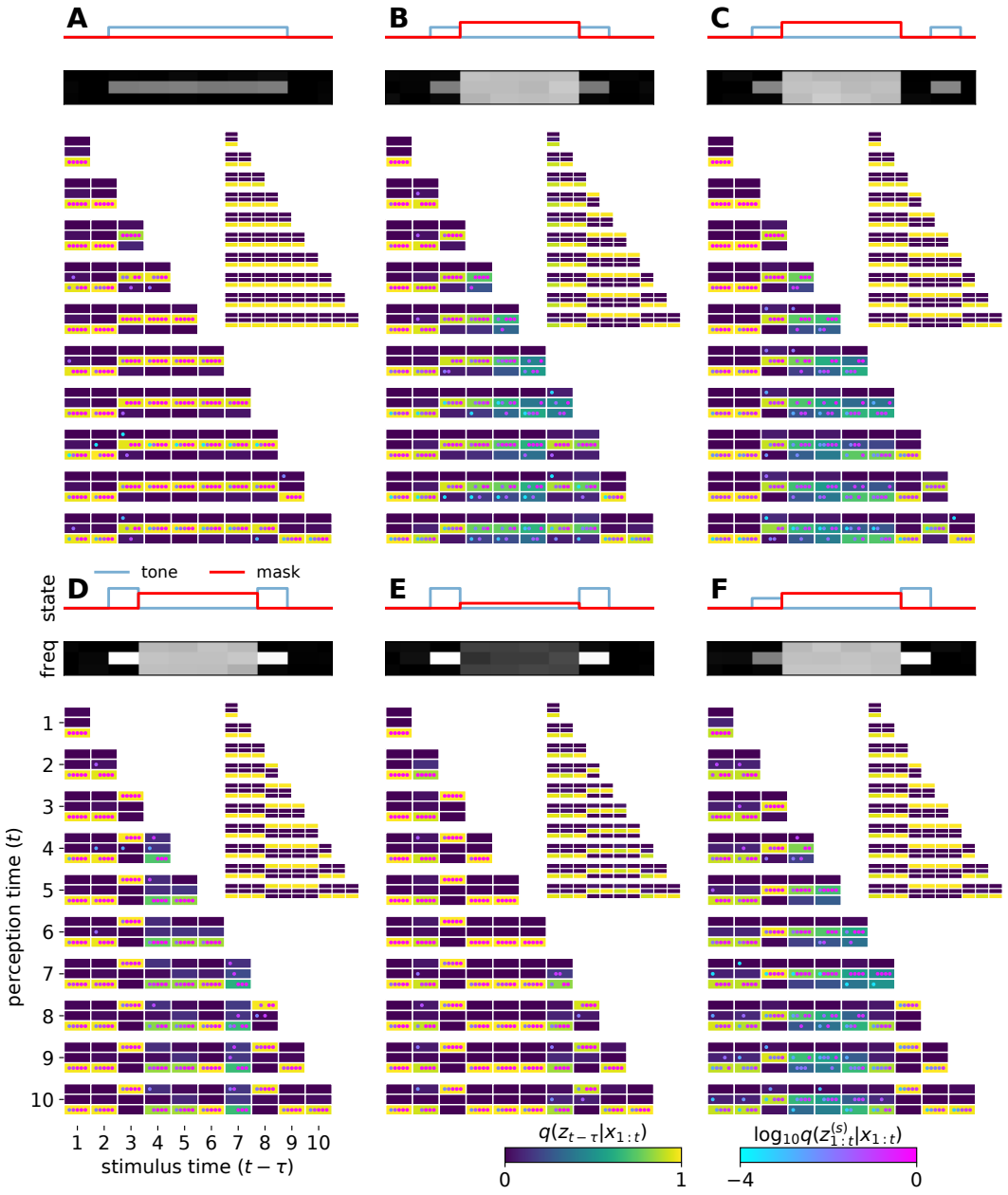


Figure 5: Posterior marginals as in Figure 1 and samples from the joint distribution $\boldsymbol{Z}_{1:t}|\boldsymbol{x}_{1:t}$ using sequential decoding. Samples in each rectangle are sorted according to the total joint log likelihood.

shown in Figure 5. We see more temporal correlation for each joint sample: during noise presentation, samples of the tone level are more likely to stay at the previous level or drift to a more probable level as uncertainty grows; towards the end of the noise, new observation becomes available and the majority of samples should switch to the new level suggested. Thus, $\boldsymbol{r}_t$ is capable of encoding correlations in the joint distribution of the whole history.

### C.2 Flash-lag effect

The internal model that reproduced the smoothing effect is

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t\text{-}1}) = \mathcal{N}([A\boldsymbol{z}_{t\text{-}1}]_+, [0.01^2, 0.002^2, 1e^{-15}]) \qquad A = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.8 \end{bmatrix} \tag{18}$$

$$p(x_{t,i}|\boldsymbol{z}_t) = \text{Poisson}\left(3\exp\left[-\frac{(\text{loc}(i) - z_{t,0})^2}{2 \times 1.5^2}\right]\right) \tag{19}$$

where $[]_+$ is a elastic bounding box at $\pm 1$. $loc$ is a linear transformation from pixel numbers to real values.

In the DDC filter, we set the dimensionalities $K_{\boldsymbol{\psi}} = 500, K_{\boldsymbol{\gamma}} = 100$ and $K_{\boldsymbol{\sigma}} = 100$ and used the $\boldsymbol{h}^{lin}(\cdot)$ in (11).

### C.3 Noisy and occluded tracking

The internal model has 3-D latent (2 continuous, 1 discrete) and 30-D observation.

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t\text{-}1}) = \mathcal{N}(f(\boldsymbol{z}_{t\text{-}1}), [0.1^2, 0.1^2]) \tag{20}$$

$$f(\boldsymbol{z}_t) = s_t \boldsymbol{A}\boldsymbol{z}_{t\text{-}1} \tag{21}$$

$$s_t = \frac{1}{\|\boldsymbol{z}_{t\text{-}1}\|_2 \exp(-4(\|\boldsymbol{z}_{t\text{-}1}\|_2 - 0.3) + 1)} \tag{22}$$

$$p(m_t|m_{t-1}) = (\text{Bernoulli}(0.1) + m_{t-1}) \mod 2 \tag{23}$$

$$p(z_{t,i}|\boldsymbol{z}_t, m_t) = \mathcal{N}\left(\max\left\{\exp\left[-\frac{(\text{loc}(i) - z_{t,0})^2}{2 \times 3^2}\right], m_t\right\}, I_{30} 0.1^2\right) \tag{24}$$

$$\tag{25}$$

where $loc$ is a linear transformation from pixel number to real values, and $\boldsymbol{A}$ is a rotation matrix by $\pi/8$. Due to the sigmoidal scaling, $\boldsymbol{z}_t$ stays around the unit circle most of the time, but can occasionally cross through the origin due to noise.

In the DDC filter, we set the dimensionalities $K_{\boldsymbol{\psi}} = 500, K_{\boldsymbol{\gamma}} = 100$ and $K_{\boldsymbol{\sigma}} = 200$ and used the $\boldsymbol{h}^{lin}_{\boldsymbol{W}}(\cdot)$ in (11).

The maximum entropy decoding of the posterior marginals are shown in Figure 6.

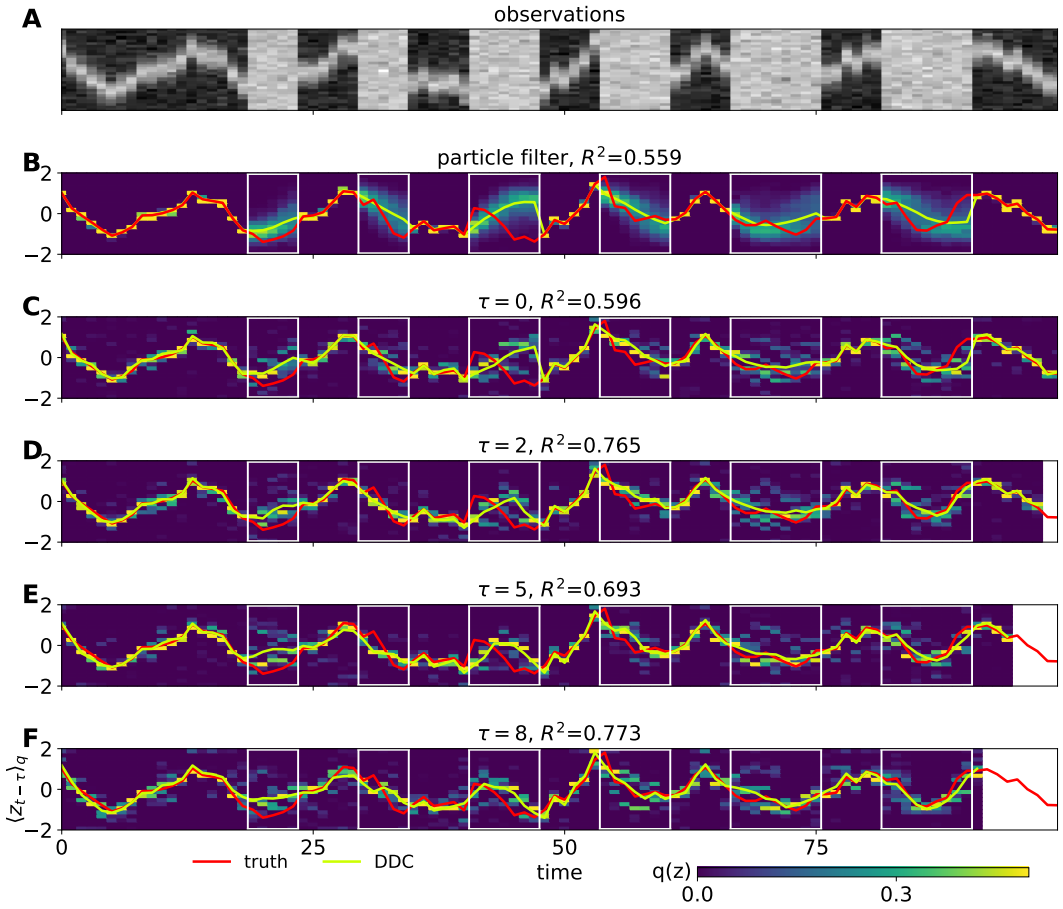


Figure 6: Maximum entropy decoding of the posterior marginals in the tracking experiment, compared with Figure 3 which is obtained by approximating expectation of bin functions.