

# Bayesian Filtering in Spiking Neural Networks: Noise, Adaptation, and Multisensory Integration

Omer Bobrowski, Ron Meir and Yonina C. Eldar  
Department of Electrical Engineering  
Technion, Israel

September 2008

NEURAL COMPUTATION, IN PRESS

## Abstract

A key requirement facing organisms acting in uncertain dynamic environments is the real-time estimation and prediction of environmental states, based upon which effective actions can be selected. While it is becoming evident that organisms employ exact or approximate Bayesian statistical calculations for these purposes, it is far less clear how these putative computations are implemented by neural networks in a strictly dynamic setting. In this work we make use of rigorous mathematical results from the theory of continuous time point process filtering, and show how optimal real-time state estimation and prediction may be implemented in a general setting using simple recurrent neural networks. The framework is applicable to many situations of common interest, including noisy observations, non-Poisson spike trains (incorporating adaptation), multisensory integration and state prediction. The optimal network properties are shown to relate to the statistical structure of the environment, and the benefits of adaptation are studied and explicitly demonstrated. Finally, we recover several existing results as appropriate limits of our general setting.

## 1 Introduction

The selection of appropriate actions in the face of uncertainty is a formidable task faced by any organism attempting to actively survive in a hostile dynamic environment. This task is further exacerbated by the fact that the organism does not have direct access to the environment (or to its internal body state), but must assess these states through noisy sensors, often representing the world via random spike trains. It is becoming increasingly evident that in many cases organisms employ exact or approximate Bayesian statistical calculations (Averbeck, Latham, & Pouget, 2006; Deneve, Latham, & Pouget,

2001; Ma, Beck, Latham, & Pouget, 2006; Pouget, Deneve, & Duhamel, 2002; Doya, Ishii, Pouget, & Rao, 2007; Knill & Pouget, 2004) in order to continuously estimate the environmental (or bodily) state, integrate information from multiple sensory modalities, form predictions and choose actions. What is less clear is how these putative computations are implemented by neural networks in a dynamic setting. Moreover, given that the environment itself is uncertain, it would seem natural to capture this uncertainty by a distribution over states rather than a single state estimator (Zemel, Dayan, & Pouget, 1998). This full distribution can be later utilized differentially in various contexts, and, in particular, for the optimal combination of different information sources. Thus, the effective representation of full probability distributions by neural networks is also an important issue which needs to be resolved.

The problem of hidden state estimation based on multiple noisy spike trains has been receiving increasing attention over the past few years. Much emphasis has been laid on Bayesian approaches, which facilitate the natural incorporation of prior information, and which can often be guaranteed to yield optimal solutions. While many naturally occurring problems are dynamic in nature, a large fraction of the work to-date has focused on static stimuli (e.g., (Averbeck et al., 2006; Deneve et al., 2001; Ma et al., 2006; Pouget et al., 2002; Pouget, Dayan, & Zemel, 2003; Zemel et al., 1998)). More recently attention has shifted to dynamic phenomena and online estimation (e.g., (Barbieri et al., 2004; Beck & Pouget, 2007; Deneve, 2008; Eden, Frank, Solo, & Brown, 2004; Huys, Zemel, Natarajan, & Dayan, 2007; Pitkow, Sompolinsky, & Meister, 2007)). Our work, formulated within the rigorous theory of real-time nonlinear filtering, and applied to dynamic spike train decoding, offers several advantages over previous work, as described in more detail in Section 7.3. In fact, our results indicate that optimal real-time state estimation based on point process observations is achievable by relatively simple neural architectures. As opposed to much previous work, there is no need for time discretization and input process smoothing, which may lead to loss of information. These results suggest a solid theoretical foundation for dynamic neural decoding and computation, and recover many previous results in appropriate limits. A particularly useful feature of the present framework is the demonstration that the computation of the posterior distribution can be achieved in real-time by a bilinear neural network. Ultimately, however, the merit of a model is based not only on its mathematical elegance, but on its power to explain existing experiments, and to predict novel behaviors. While some limits of our formulation, e.g., the static limit, lead to results which have already been experimentally verified, the main advantage of the general framework is in setting the stage for mathematically precise, yet experimentally verifiable, predictions for future experiments dealing directly with dynamic phenomena.

Consider the following generic situation. An agent observes the environment through a set of noisy (possibly multimodal) sensory neurons. Based on these observations the agent needs to estimate the state of the environment (more generally, the state distribution) with the highest accuracy possible. It is well known that if the stochastic dynamics of the environment and the observation process are fully known, then the state distribution can be optimally recovered through the so-called *Bayes filter* (Jazwinsky, 1970; Thrun, Burgard, & Fox, 2005) based on an exact calculation of the posterior state

distribution. For example, if both the environmental and observational processes are linear, and are corrupted by Gaussian noise, the optimal filter reduces to the classic Kalman filter (Anderson & Moore, 2005). For the state estimation procedure to be effective in a biological context, it must be possible to implement it robustly in real time by a neural network. In a biological setting, the agent observes the environment through a set of sensory neurons, each of which emits spikes at a rate which depends on the current state of the environment according to the neuron’s fixed response function (a.k.a. tuning curve). These spike trains are continually sent to a neural network which computes a probability distribution over environmental states.

Surprisingly it turns out that under well-defined mathematical conditions (hidden Markov process and Poisson spiking activity; see Section 2 for precise definitions) the solution to the problem of spike train decoding has been known for many years see, for example, (Brémaud, 1981), and the historical survey provided therein. However, the mathematical derivation in (Brémaud, 1981) is highly intricate, relying on sophisticated mathematical techniques from the theory of stochastic processes, which may not be widely available. This abstruseness may be one reason for the fact that this exact and rigorous body of theory has rarely been used by the computational neuroscience community (see, for example, (Twum-Danso & Brockett, 2001) for a notable exception). In fact, some of the results presented over the past few years in the context of hidden state estimation and neural decoding can be viewed as special cases of the general theory developed in (Boel & Benes, 1980) and (Brémaud, 1981). Because of the intricate nature of the derivation in (Brémaud, 1981), we present a simplified derivation available as an online appendix<sup>1</sup>. This online appendix will enable readers who are unfamiliar with the advanced theory of martingales to follow the derivation using simple techniques. Within this framework, the optimal posterior distribution over environmental states (the Bayes filter), given the sensory spike trains, is exactly computed in real-time by a bilinear recurrent neural network. It is essential to note that the posterior distribution is based on the *exact* spike times, so that no temporal information is lost (as is often the case when time is discretized or other approximations are made). A preliminary version of the results appears in (Bobrowski, Meir, Shoham, & Eldar, 2007). Next, we summarize the main contributions of this work.

**The main contributions of this work:** (1) Incorporation of environmental noise in the general filtering framework. Within this setting we establish the existence of an optimal width for the tuning function of the sensory cells. This width depends on the noise level, suggesting that for optimal performance the system must adapt to the specific environmental conditions. (2) Application of the framework to the multisensory integration of signals. Our results provide novel predictions in the dynamic setting, and recover previous results in the static limit (e.g., (Deneve et al., 2001)); see section 4 for details. Furthermore, they provide succinct explanations for several experimentally observed phenomena related to enhanced response and inverse effectiveness. (3) Development of a framework for history dependent spike trains, and a consideration of the effect of adaptation on system performance. Interestingly, we can show that adaptation can benefit neural computation. More specifically, when the system is subjected to

---

<sup>1</sup>See <http://www.technion.ac.il/~rmeir/BobMeiEld-appendix.pdf>

energy constraints (e.g., limits on the overall number of spikes fired per unit time), adaptation leads to near optimal performance. (4) Showing how a simply modified system addresses prediction of future states, rather than estimating the current state).

The remainder of this paper is organized as follows: Section 2 describes the precise problem formulation, and the basic filtering equation based on (Brémaud, 1981). We show how this equation can be implemented by a simple recurrent neural network, followed by several simulations demonstrating the system’s performance. Section 3 incorporates environmental noise and studies its effect. Section 4 considers multisensory integration, and Section 5 discusses in detail the case where the stimulus (or world state) is static. Section 6 presents an extension to a larger class of point processes, and demonstrates how phenomena such as adaptation can easily be incorporated. Section 7 briefly describes additional extensions (prediction, log-posterior computation) as well as a detailed comparison to previous work.

## 2 Filtering a Markov Process from Poisson Measurements

Consider a dynamic process  $X_t$  representing the state of the world (e.g., the location of an object, its shape, orientation, velocity, etc.). We assume that  $X_t$  is a continuous time finite state Markov process, with a finite state-space  $\mathcal{S} = \{s_1, \dots, s_N\}$  and an infinitesimal generator matrix  $Q = [q_{ij}]$ . This implies that the transition probabilities are given by

$$\mathbf{P}_{ij}^{(\tau)} \triangleq \mathbb{P}(X_{t+\tau} = s_j | X_t = s_i) = \begin{cases} q_{ij}\tau + o(\tau) & i \neq j \\ 1 + q_{ii}\tau + o(\tau) & i = j, \end{cases} \quad (2.1)$$

with  $q_{ii} = -\sum_{j \neq i} q_{ij} < 0$  (see (Grimmett & Stirzaker, 2001) for more details).

The state  $X_t$  is not directly observed, but is processed through a set of  $M$  sensory cells, each of which produces a spike train, associated with a counting process  $N_t^{(m)}$ . At this point we take the spikes to be generated by an inhomogeneous Poisson process, where the process rate depends on the current environmental state (such process are referred to as *doubly-stochastic Poisson processes*, see (Snyder & Miller, 1991)). We denote the rate of the process generated by the  $m$ -th cell by  $\lambda_m(X_t)$ , where  $\lambda_m(\cdot)$  represents the *tuning curve* of the  $m$ -th cell. The firing events of the different sensory cells are assumed to be independent given the state. Our goal is to compute the posterior probabilities

$$p_i(t) \triangleq \mathbb{P}\left(X_t = s_i \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right),$$

where  $N_{[0,t]}^{(m)} = \left\{N_s^{(m)}\right\}_{s=0}^t$  is the full history of the process  $N_t$ . More specifically, we are looking for an online computation method that can be carried out by a neural network.

In the remainder of this section we present the solution derived in (Brémaud, 1981), and discuss the interpretation of this solution as a neural network, followed by simulations. For completeness we present a full, albeit simplified, derivation of these results

in an online appendix. We note that simplified derivations of special cases of the filtering equations in (Brémaud, 1981), based on time discretization followed by a limit process, have been recently presented in (Deneve, 2008) and (Pitkow et al., 2007); see also (Twum-Danso & Brockett, 2001) for an intuitive explanation of the results from (Brémaud, 1981) in a continuous time context. We discuss this work in a comparative setting in Section 7.

## 2.1 The Filtering Equation

There is increasing interest in providing an answer to the problem presented above in different neuroscience contexts. Interestingly, as stated in Section 1, this mathematical filtering problem was addressed in the 1970s and rigorous solutions, under well defined mathematical conditions, exist since then. In a historical context we note that a mathematically rigorous approach to point process filtering in continuous time was developed during the early 1970s following the seminal work of Wonham (Wonham, 1965) for finite state Markov processes observed in Gaussian noise, and of Kushner (Kushner, 1967) and Zakai (Zakai, 1969) for diffusion processes. One of the first papers presenting a mathematically rigorous approach to nonlinear filtering in continuous time based on point process observations was Snyder (Snyder, 1972), extended later by Segall et al. (Segall, Davis, & Kailath, 1975). We comment that this paper considers only the case of a finite state space. The formalism for continuous state spaces is also available in some cases (e.g, (Boel & Benes, 1980)), but will not be pursued in this work.

The solution presented in (Brémaud, 1981) introduces a new set of non-negative and non-normalized functions  $\rho_i(\cdot)$ , related to  $p_i(t)$  by

$$p_i(t) = \frac{\rho_i(t)}{\sum_{j=1}^N \rho_j(t)}. \quad (2.2)$$

It is shown in Section VI.4 of (Brémaud, 1981) (see also the online appendix) that  $\{\rho_i(t)\}_{i=1}^N$  obey the following set of  $N$  differential equations,

$$\dot{\rho}_i(t) = \sum_{k=1}^N q_{ki} \rho_k(t) + \left( \sum_{m=1}^M (\lambda_m(s_i) - 1) \nu_m(t) \right) \rho_i(t) - \lambda(s_i) \rho_i(t), \quad i = 1, \dots, N, \quad (2.3)$$

where

$$\lambda(s_i) = \sum_{m=1}^M \lambda_m(s_i) \quad ; \quad \nu_m(t) = \sum_n \delta(t - t_n^{(m)}),$$

and  $\{t_n^{(m)}\}_{n=1}^{\infty}$  denote the spiking times of the  $m$ -th sensory cell. In other words, the function  $\nu_m(t)$  represents the spike train of the  $m$ -th sensory cell. The parameters  $q_{ki}$  are elements of the generator matrix  $Q$ . This set of equations can be written in vector form as

$$\dot{\boldsymbol{\rho}}(t) = Q^T \boldsymbol{\rho}(t) + \left( \sum_{m=1}^M (\Lambda_m - \mathbf{I}) \nu_m(t) \right) \boldsymbol{\rho}(t) - \Lambda \boldsymbol{\rho}(t), \quad (2.4)$$

where  $\mathbf{I}$  is the identity matrix, and

$$\boldsymbol{\rho}(t) = (\rho_1(t), \dots, \rho_N(t))^\top \quad ; \quad \Lambda_m = \text{diag}(\lambda_m(s_1), \dots, \lambda_m(s_N)) \quad ; \quad \Lambda = \sum_{m=1}^M \Lambda_m.$$

In appendix A we present a full closed form solution to (2.4). Examining the solution given in (A.3), we can analyze this network’s activity pattern. Between spikes,  $\boldsymbol{\rho}(t)$  depends exponentially and smoothly on time, varying at a time scale which depends on the eigenvalues of  $\mathbf{Q}$  and on the maximal firing rates of the sensory cells (through the tuning curves). Upon the arrival of a spike from the  $m$ -th sensory cell at time  $t$ ,  $\boldsymbol{\rho}(t)$  is updated according to  $\boldsymbol{\rho}(t^+) = \Lambda_m \boldsymbol{\rho}(t^-)$ . In other words, the variable  $\rho_i(t)$  evolves on a slow time scale between spikes and on a fast time scale upon the arrival of a spike.

### Network interpretation

The variables  $\{\rho_1, \dots, \rho_N\}$  in (2.3) can be interpreted as representing the activity of a set of  $N$  neurons in a recurrent (posterior) neural network. The second term in (2.3) represents the effect of the sensory inputs on each such posterior neuron. Each sensory neuron emits a Poisson spike train<sup>2</sup> based on the current input and its receptive field, and affects the posterior cell through its impulse train  $\nu_m(t)$ . In addition, the first term in (2.3) shows that each posterior neuron receives inputs from other posterior neurons based on the weights  $\{q_{ij}\}$ . The third term represents a simple, state-dependent, decay of the non normalized distribution. In many cases the variables  $\lambda(s_i)$  are near constant, in which case they affect the posterior variables  $\rho_i$  only the overall normalization, and can be dropped. A graphical display of the network and its interpretation can be found in Figure 1. Note that the first term in (2.3) represents a ‘mapping of the world’ onto the recurrent decoding network, establishing a representation of the environment through the synaptic weights; this can be interpreted as a Bayesian prior in a Bayesian setting.

One possible implementational problem with (2.3) is that the solution may explode exponentially. Since the physical variable of interest is the posterior probability, obtained by renormalizing  $\rho_i(t)$  as in (2.2), the overall normalization is irrelevant. Note, however, that in some cases (e.g., computing the minimum mean squared error estimator) the normalized distribution is required. In any event, as long as we are interested in the largest posterior probability, it is clear that the normalization has no effect. In principle, one can add to (2.3) an operation which periodically renormalizes the variables so that  $\sum_{i=1}^N \rho_i(t) = 1$ . Alternatively, one can add a term to the equation which guarantees that this normalization be automatically obeyed at each step, as was done, for example, in equation (2.13) of (Beck & Pouget, 2007). In the numerical demonstrations presented in the sequel we have renormalized  $\rho_i(t)$  periodically in order to prevent explosive solutions.

The network described in (2.3) can be viewed as a formal neural network. While it seems to be somewhat removed from a direct physiological interpretation, we believe that a physiologically plausible network can be constructed based on its principles. We discuss

---

<sup>2</sup>The Poisson assumption will be relaxed in Section 6.

the main interpretational issues here, but defer a full physiological implementation of these ideas to future work.

First, we note that while the sensory neurons produce spike trains represented through the variables  $\nu_m(t)$ , the posterior neurons are described by a *continuous* variable  $\rho_i(t)$ . Within a biological implementation of (2.4) one may view  $\rho_i(t)$  as the probability of spiking (see also (Pitkow et al., 2007)). This interpretation may seem to pose difficulties in a biological context since the probability of spiking cannot be directly communicated between neurons. However, an easy remedy for this would be to simply replicate each posterior neuron many times, and allow each such replicated neuron to fire Poisson spike trains at a rate consistent with (2.3). This would correspond to the well-studied linear Poisson spiking network, e.g., (Gerstner & Kistler, 2002). Since spikes can be directly communicated between neurons, such an implementation would be biologically feasible. A second difficulty with the physiological interpretation relates to the multiplicative gain, the second term in (2.4). This term requires that the activity of the posterior neuron  $i$  be modulated by a multiplicative term based on the activity of the sensory neurons. While not entirely standard, there is an increasing evidence for this type of multiplicative gain in biological neural networks in both the visual and somatosensory cortices (e.g., (C. J. McAdams & Maunsell, 1999; C. McAdams & Reid, 2005; Sripathi & Johnson, 2006)), and such interactions are thought to play an important role in neural computation (Salinas & Thier, 2000). Moreover, specific biophysical mechanisms and computational models have been proposed for these phenomena (Murphy & Miller, 2003; Sripathi & Johnson, 2006). A third issue which can be raised against the plausibility of the network proposed is that posterior neurons are affected differentially based on the feedforward input sensory neurons (multiplicative gain) and additively through the posterior network recurrent weights. However, given the very different nature of the two types of inputs, and their effects on the postsynaptic target through different distributions of receptors, it is not implausible that the two types of interactions lead to very different effects; see, for example, (Rivadulla, Sharma, & Sur, 2001). In fact, there is solid evidence for the existence of multiplicative interactions based on the special properties of the NMDA receptor (Rivadulla et al., 2001), which is ubiquitous in cortical circuits and is widely believed to lead to coincidence detection (Tsien, 2000). Next, we comment on the renormalization issue alluded to above. This can be addressed within a physiological context using the well documented phenomenon of divisive inhibition (e.g., (Cavanaugh, Bair, & Movshon, 2002)); see, for example, (Beck & Pouget, 2007) for a simple implementation of divisive inhibition, in a context related to (2.4), leading to normalization.

Finally, we comment briefly on the possible implementation of the formal network (2.3) in the brain. In a visual context we view the sensory layer in the model as corresponding to the inputs from the retina via the LGN to the cortex (a similar interpretation, *mutatis mutandis*, would hold for other sensory modalities). The recurrent connections  $\{Q_{ij}\}$  within the posterior network would then correspond to the lateral connections between cortical pyramidal neurons. The latter are well known to play an essential role in cortical processing, overwhelming the thalamic inputs by a wide margin. Interestingly, within our model, stronger connections  $Q_{ij}$  exist between neurons which represent

similar states. For example, in a dynamic context, the matrix elements between similar states are larger, corresponding to higher transition probabilities between such states. This observation is consistent with the larger observed functional connectivity between cells of similar orientation selectivity (e.g., (Ts'o, Gilbert, & Wiesel, 1986)). Moreover, the competitive dynamics of our model's posterior network is also consistent with the soft winner-take-all view of the lateral interactions between cortical neurons; see (Douglas & Martin, 2004) for a physiological motivation and demonstration. Experimental tests of our proposed model could consist of differentially interfering with the feedforward multiplicative interactions (possibly through NMDA receptor antagonists) and the lateral additive interactions suggested by our model, thereby comparing the different spatial and temporal effects of the two information streams. For example, we would expect that disrupting lateral connections ('prior knowledge') would lead to particularly significant performance degradation when the sensory input is sparse. We note that an equation similar to (2.4) has been derived recently in (Pitkow et al., 2007) for a two dimensional random walk Markov process. The latter paper provides further support to the idea that the visual area V1 may naturally implement this type of network.

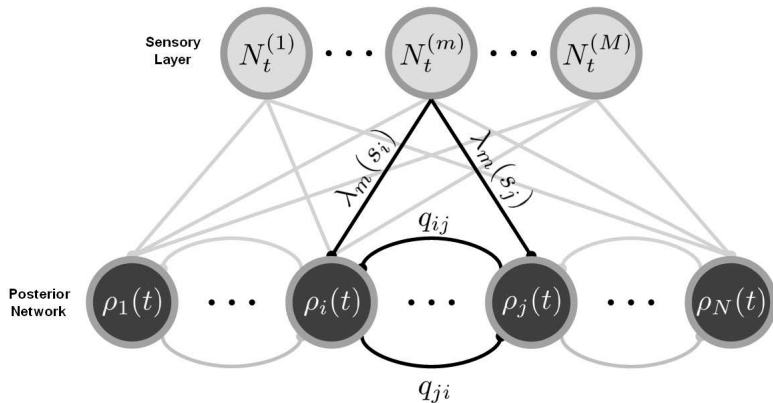


Figure 1: The decoding network structure. The sensory cells respond to a stimulus with spike trains  $N_t^{(m)}$ . The connection strength between the  $m$ th sensory cell and the  $i$ th posterior network cell is  $\lambda_m(s_i)$  (as in (2.3)). This connection weight multiplies the sensory cell activity and is passed as input to the second recurrent layer. This layer computes the posterior probabilities  $\rho(t)$  based on (2.3). The recurrent synaptic weights in the posterior network are controlled by the prior transition parameters  $q_{ji}$ .

### 2.1.1 Numerical demonstrations of the filtering equations

Next, we examine the system behavior by numerically solving (2.4) and its extensions<sup>3</sup>. The numerical solution corresponds to an actual implementation of the abstract neural network described in (2.3,2.4). These results are aimed at demonstrating performance, and can be viewed as a simple implementation of the experimental setup considered in (Warland, Reinagel, & Meister, 1997) in the context of retinal decoding.

<sup>3</sup>A closed form solution to (2.4) can be found in Appendix A

We consider a simple setting where the decoding system attempts to track a moving particle; this basic setup, with modifications, will serve for all the numerical demonstrations in this work. Consider a small object moving vertically on a line, jumping between a set of discrete states  $\{s_i\}_{i=1}^N$ , each representing the position of the object. The object is observed by a retina consisting of  $M$  sensory cells, where each sensory cell  $m$  generates a Poisson spike train with rate  $\lambda_m(X_t)$  where  $X_t$  is the world state at time  $t$ . The tuning curve of the  $m$ th sensory cell is taken to be a Gaussian<sup>4</sup> centered at  $c_m$ , of width  $\alpha$ , height  $\lambda_{\max}$ , and baseline level  $\lambda_{\text{base}}$ , namely  $\lambda_m(s) = \lambda_{\text{base}} + \lambda_{\max} \exp(-(s - c_m)^2/2\alpha^2)$ . The tuning-curve centers  $c_m$  are uniformly spread over the input domain. The same experimental protocol is used throughout the paper with slight variations required by the extended settings described in sections 3, 4 and 6.

For the simulations presented throughout this paper we use different  $Q$  matrices to represent the world state dynamics. The  $Q$  matrices are constructed in such a way that the most likely transitions are from any state to one of its neighbors, where a neighbor is defined by the Euclidean distance between the physical states. The general structure of the  $Q$  matrix is as follows:

$$q_{ij} = \begin{cases} c_i \exp\left(\frac{(i-j)^2}{2\beta^2}\right) & i \neq j \\ -\mu & i = j, \end{cases} \quad (2.5)$$

where  $\beta$  and  $\mu$  are positive real numbers, and  $c_i = \mu / \sum_{j \neq i} \left(\exp\left(\frac{(i-j)^2}{2\beta^2}\right)\right)$ , so that  $\sum_j q_{ij} = 0$ . The average number of transitions per unit time is  $\mu$ .

Figure 2 displays the motion of the particle between  $N = 250$  different states, for a choice of  $Q$  matrix as described above. The spiking activity of  $M = 125$  position sensitive sensory cells, and the tracked posterior distribution. In Figure 2(a)-2(c) the tuning curve parameters are chosen to produce only a few spikes. In this case we can see how the level of uncertainty (posterior variance) increases between spike arrivals, when no information is provided. In Figure 2(d)-2(f), the spiking activity is much more intense, leading to more accurate results (lower posterior variance).

Next, we consider a naturally discrete state discrimination task. This example also demonstrates how to achieve improved performance by enriching the state space of the model. Specifically, we augment the state representation by adding the movement direction as well as the visibility mode of the stimulus. Define a new set of states  $\tilde{s}_{ijk} = (s_i, d_j, v_k)$ , where  $s_i$  represent the object's locations,  $d_j$  denotes the current movement direction (in this case  $d_1 = \text{up}$ ,  $d_2 = \text{down}$ ), and  $v_k$  represents whether or not the stimulus is visible to the system ( $v_1 = \text{visible}$ ,  $v_2 = \text{invisible}$ ). The tuning curves  $\lambda_{ijk}$ ,  $i = 1, 2, \dots, N$ ,  $j, k = 1, 2$ , are constructed as follows. For states where the stimulus is visible, the tuning curves are Gaussian functions of the location as before. However, for the states where the stimulus is invisible, the cells cannot differentiate between different locations and hence they all respond with the same spontaneous rate  $\lambda_{\text{spon}}$ . Note that the movement direction is *not* encoded in the firing rate of the sensory cells.

---

<sup>4</sup>The results are demonstrated for a Gaussian tuning curve, but the theory applies to arbitrary tuning functions.

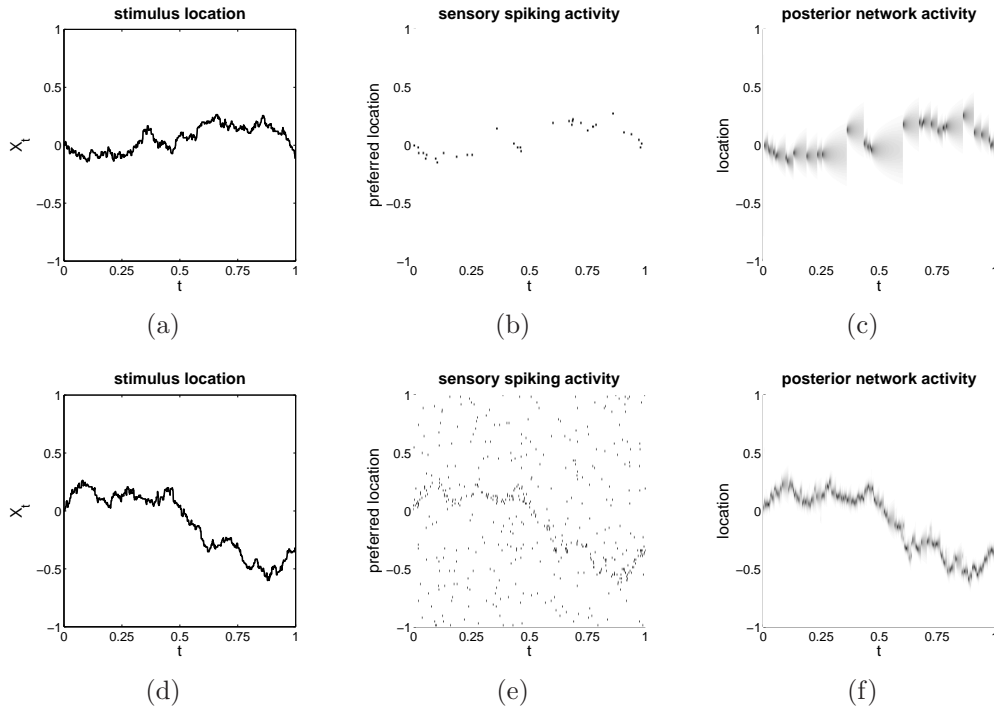


Figure 2: Tracking the motion of a single object in 1D. (a) The trajectory of the object’s movement. (b) Sensory activity. A dot represents a spike arriving from a sensory cell, where the y-axis represents the cell’s preferred location. In this simulation the firing rates are extremely low ( $\lambda_{\max} = 15, \lambda_{\text{base}} = 0$ ). (c) The activity of the posterior network. The y-axis represents the location represented by each cell, and the black intensity represents the probability  $\mathbb{P}(X_t | \text{spiking activity})$ , ranging from 0=white to 1=black. (d)-(f) Same setup, with  $\lambda_{\max} = 75, \lambda_{\text{base}} = 2.5$  leading to an intense sensory activity. In both simulations  $N = 250, M = 125, \alpha = 0.016, \beta = 2$  and  $\mu = 500$ .

The basic dynamic setup is as in Figure 2, except that the states are augmented as described in the previous paragraph. Denoting the non-normalized probabilities of this 3D state by  $\rho_{ijk}(t)$  we can retrieve the distribution of the location alone, by using the marginal distribution  $\rho_i^s(t) = \sum_{j,k} \rho_{ijk}(t)$ . The results are presented in Figure 3(c). Alternatively, by pooling the probabilities over all possible locations and visibility modes, we get  $\rho_j^d(t) = \sum_{i,k} \rho_{ijk}(t)$  - the discrete probabilities discriminating between the two possible movement directions. The results presented in Figure 3(d) demonstrate the high quality of the binary decision possible by classifying the direction of motion based on the maximal posterior probability for the (up,down) directions. Similarly, by pooling the probabilities over all possible locations and directions we get  $\rho_k^v(t) = \sum_{i,j} \rho_{ijk}(t)$  - the discrete probabilities discriminating between stimulus visibility and non-visibility. The results are presented in Figure 3(e).

The Q matrix used in this simulation is different than the previous one, and somewhat more complicated, as it has to include the transitions between directions and visibility modes. As there are  $N \times 2 \times 2 = 4N$  possible states, Q is of size  $4N \times 4N$ . The construction of the matrix Q in this example is provided for completeness in appendix B.

Tracking the movement direction naturally lends further robustness to the position estimation. As can be seen in Figure 3(c), when the input of the sensory cells is blocked (and the sensory cells fire spontaneously) the system estimates a movement that continues in the same direction. When the blockade is removed, the system is re-synchronized with the input. It can be seen that even during periods here sensory input is absent, the general trend is well predicted, even though the estimated uncertainty increases.

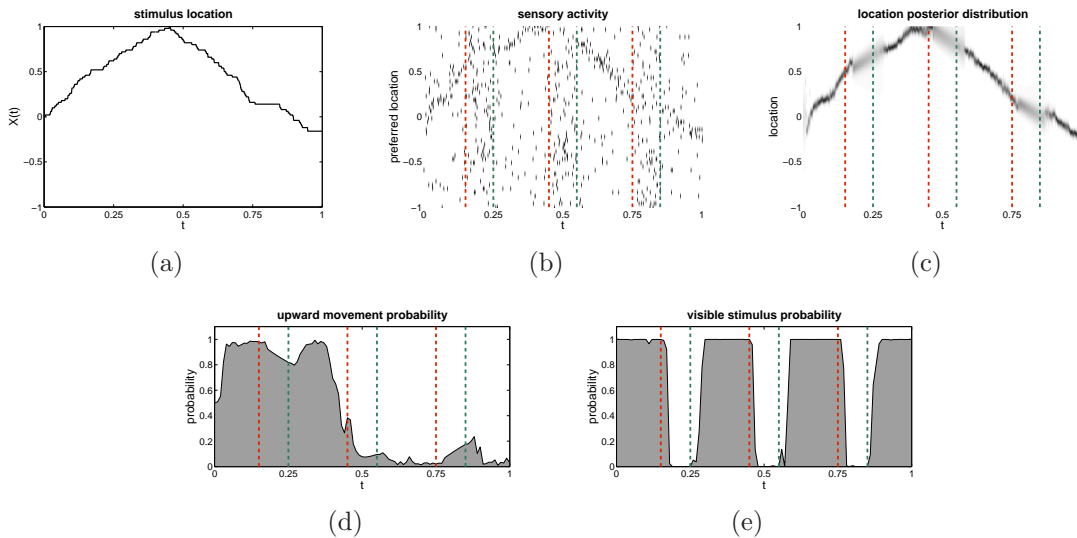


Figure 3: Tracking based on position and direction sensors. (a) The object’s trajectory. (b) The activity of sensory cells. The red bars mark points in time where the input was blocked (i.e., the transition  $(s_i, d_j, v_1) \rightarrow (s_i, d_j, v_2)$ ), and the green bars mark the times when the blockade was removed. (c) The posterior evolution based on place and direction sensory input. (d) The direction discriminating distribution. We present the probability of upward movement  $p_{\text{up}}$ , the downward movement probability is simply  $p_{\text{down}} = 1 - p_{\text{up}}$ . (e) The visibility discriminating distribution. We present the probability that the stimulus is visible  $p_{\text{visible}}$ , the invisible stimulus probability is  $p_{\text{invisible}} = 1 - p_{\text{visible}}$ . The simulation parameters are:  $N = 101$ ,  $M = 50$ ,  $\lambda_{\text{max}} = 75$ ,  $\lambda_{\text{base}} = 5$ ,  $\lambda_{\text{spn}} = 18.75$ .

### 3 Noisy Environment

The assumption so far (as in previous work) is that the tuning curve  $\lambda_m(X_t)$  is a *direct* function of the state, and that the uncertainty in the system arises from the Poisson spike trains only. In other words, we assume that the sensory cells *have access* to the state of the world, and the uncertainty is due to their own noisy activity. However, noise is likely to appear even before the spike trains are generated. For example, consider the task of tracking an object moving through haze. In this case the image perceived by the retina is blurred and unclear, therefore the neural activity cannot represent the object’s location directly, but rather noisy information about its location.

We consider the model in which a hidden state process  $X_t$  (defined similarly to the

previous section) passes through a noisy channel. This channel introduces an interference process  $W_t$ . As a result, the input arriving to the sensory system is  $F(X_t, W_t)$  rather than  $X_t$  itself, implying that the tuning curves are now  $\lambda_m(F(X_t, W_t))$  rather than  $\lambda_m(X_t)$  (see Figure 4). For example, assuming an additive-noise model we take  $F(x, w) = x + w$ . However the model presented here applies to any general function  $F$ . To simplify the notation we define  $\tilde{X}_t = (X_t, W_t)$ , and  $\lambda_m(\tilde{X}_t) \triangleq \lambda_m(F(X_t, W_t))$ .

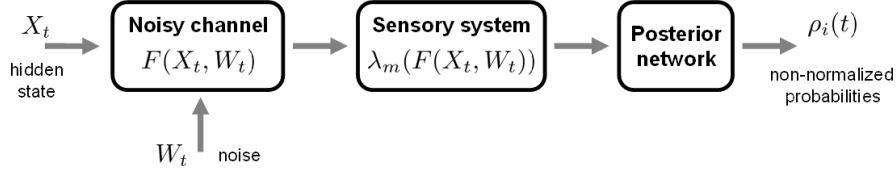


Figure 4: Environmental noise model.

Assuming that  $W_t \in \{w_1, \dots, w_L\}$  is a continuous time finite state Markov process (CFMP), independent of  $X_t$ , the combined process  $\tilde{X}_t$  is also a CFMP (with  $N \times L$  different states). Thus, using (2.4) we can compute the non-normalized probabilities  $\tilde{\rho}_{i,j}(t) \propto \mathbb{P}(\tilde{X}_t = (s_i, w_j) | N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)})$ . To obtain  $X_t$ 's non-normalized posterior distribution, we compute the marginal non-normalized probabilities -  $\rho_i(t) = \sum_j \tilde{\rho}_{i,j}(t)$  (see Figure 5). In order to avoid the computation of the nuisance noise distribution (as in Figure 5), we present in Appendix C the derivation of the following set of equations computing the state distribution directly

$$\dot{\rho}_i(t) = \sum_{k=1}^N q_{ki} \rho_k(t) + \left( \sum_{m=1}^M (\eta_m(s_i, t) - 1) \nu_m(t) \right) \rho_i(t) - \eta(s_i, t) \rho_i(t), \quad (3.1)$$

where

$$\eta_m(s_i, t) = \mathbb{E} \left[ \lambda_m(X_t, W_t) | X_t = s_i, N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)} \right] \quad ; \quad \eta(s_i, t) = \sum_{m=1}^M \eta_m(s_i, t).$$

Equation (3.1) is similar to (2.3), except that instead of  $\lambda_m(s_i)$  we require a *time-varying* synaptic weight  $\eta_m(s_i, t)$  which is the average sensory response (with respect to the noise).

In principle, computing the expectation required to estimate  $\eta_m(s_i, t)$  requires conditioning on the spiking history. Assuming that  $W_t$  changes sufficiently rapidly relative to the spiking activity, and that all its moments are finite, we may remove the spiking history from the conditional expectation implying the following approximate relationship

$$\eta_m(s_i, t) \approx \mathbb{E} [\lambda_m(X_t, W_t) | X_t = s_i] \triangleq \eta_m(s_i),$$

which yields the equation

$$\dot{\rho}_i(t) = \sum_k q_{ki} \rho_k(t) + \left( \sum_{m=1}^M (\eta_m(s_i) - 1) \nu_m(t) \right) \rho_i(t) - \eta(s_i) \rho_i(t) \quad (i = 1, 2, \dots, N), \quad (3.2)$$

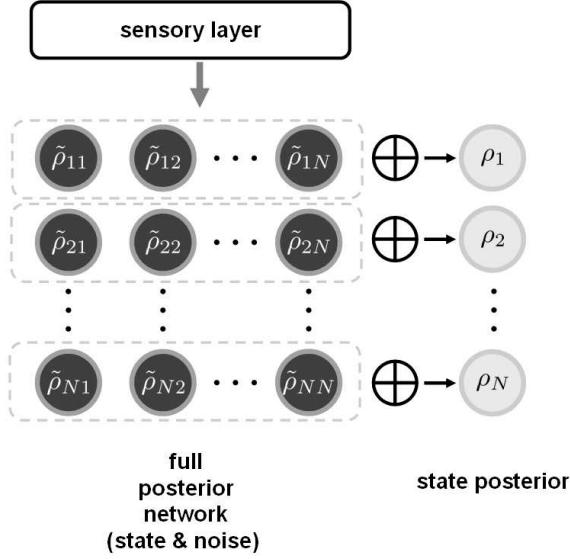


Figure 5: Computing the posterior distribution in the presence of noise. The full posterior network computes the posterior distribution of the combined state  $\tilde{X}_t = (X_t, W_t)$ . By a simple summation we get the posterior distribution of the state  $X_t$  alone.

where  $\eta(s_i) = \sum_{m=1}^M \eta_m(s_i)$ . The equation set (3.2) calculates the posterior distribution of the state process  $X_t$  alone, using the average responses of the sensory cells, with respect to the noise process  $W_t$ .

Similarly to the noiseless case, we can represent (3.2) in a vector form, as

$$\dot{\rho}(t) = Q^\top \rho(t) + \left( \sum_{m=1}^M (\Phi_m - \mathbf{I}) \nu_m(t) \right) \rho(t) - \Phi \rho(t), \quad (3.3)$$

where

$$\Phi_m = \text{diag}(\eta_m(s_1), \dots, \eta_m(s_N)) \quad ; \quad \Phi = \sum_{m=1}^M \Phi_m.$$

Figure 6 presents the performance of the system in the presence of noise. Figure 6(a)-6(c) present a single trial of tracking a stimulus moving through  $N = 250$  states, observed by  $M = 125$  sensory cells. The state model is similar to the one used in Figure 2. The noise state-space is uniformly distributed in the range  $[w_1, w_L]$  with  $w_1 = -1/2$ ,  $w_L = 1/2$ ,  $L = 1000$ . The noise distribution is of the form  $c \exp(-w_j^2/2\sigma_w^2)$  with  $\sigma_w^2 = 0.01$ , and we assume that it contributes additively to the input, i.e.,  $F(X_t, W_t) = X_t + W_t$ . It can be observed that the posterior network filters out the noise significantly. In Figure 6(d) we simulate different levels of noise, and compare the performance of the original filtering equation (2.4) with its noisy version (3.3). The comparison is based on the empirical mean squared error (MSE) of the minimum MSE (MMSE) optimal estimator calculated from the posterior distribution represented by the network, i.e.,

$$\hat{X}_t = \sum_i s_i \mathbb{P} \left( X_t = s_i \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)} \right).$$

Obviously, taking the noise statistics into account, significantly improves the system’s accuracy.

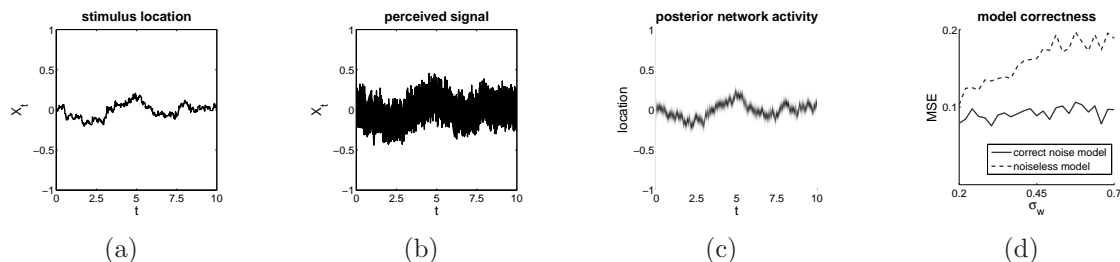


Figure 6: Noisy input simulation. (a) The original stimulus ( $N=250$ ). (b) The noisy version perceived by the sensory cells ( $M = 125, \sigma^2 = 0.01, \lambda_{\max} = 75, \lambda_{\text{base}} = 5, \alpha = 0.008, \beta = 3$  and  $\mu = 50$ ). (c) The posterior network response, filtering out the noise. (d) Comparing the MSE of the optimal estimator in the presence of noise. The solid line represents the MSE of the correct model that takes the noise into account. The dashed line represents the original model, where noise is ignored.

## 4 Multisensory Integration

Organisms usually observe the environment through multiple sensory modalities, integrating these different modalities in order to obtain a more reliable percept. Clearly a significant benefit may result from multisensory integration in situations where noise disrupts one, or both, of the modalities. Interestingly, it turns out that multisensory integration is much more prevalent in sensory processing than was once believed (Ghazanfar & Schroeder, 2006). In this section, we discuss how the framework described in Section 2 may be applied in this context to derive optimal multisensory integration of signals. In this work we only deal with the issue of the *integration* of the multimodal signals. As pointed out in (Deneve & Pouget, 2004) the problem is more involved than simple averaging of sensory modalities since the different modalities often use different coordinate frames, and some mechanism must be proposed in order to dynamically translate the signals to a common frame of reference. The present section deals with the general dynamic setting introduced in Section 2. The restriction to the static case is discussed in Section 5.

Consider the case of multi-modal inputs, where a subset of the sensory inputs arises from one modality (e.g., visual) while the remaining inputs arise from a different sensory modality (e.g., auditory). These modalities may differ in the shapes of their tuning curves, their response latencies and the information they provide about the stimulus. In the sequel we will use the symbols ‘V’ and ‘A’ to refer to any two modalities, and they should not be interpreted literally as visual and auditory. While we present the formulation for two sensory modalities, it can be easily extended to any number of modalities.

We briefly summarize our main contributions concerning multimodal integration. As far as we are aware our results provide the first derivation of optimal multimodal sensory

state estimation in a dynamic setting. Even though they follow directly from the general formulation in Section 3, they provide some specific insight into multisensory integration. First, we note that while it is clear that multisensory information is essential in providing information when one of the sensory sources disappears or is occluded, we show in Section 4.2 that it is essential also in standard noisy situations where multisensory information sources exist simultaneously. More specifically, given a fixed number of sensory cells, we show that splitting the information gathering between two sensory modalities leads to superior performance. As we show in Section 4.2 this occurs due to the independence of the noise processes contaminating each observation. Second, we provide a simple mechanistic explanation for the widely observed phenomenon of inverse effectiveness. Finally, we show in Section 5.1 that the dynamic extension to multisensory integration offered in this section, yields well known results that had been derived previously only in the static limit (Witten & Knudsen, 2005; Deneve & Pouget, 2004).

## 4.1 Multimodal Equations

We start with the simpler case, where no environmental noise is present. Consider the following case - we have a state process  $X_t$  that is observed via a set of  $M_v$  visual sensory cells with spiking activities -  $\{N_t^{v,(1)}, \dots, N_t^{v,(M_v)}\}$  firing at rates  $\{\lambda_1^v(X_t), \dots, \lambda_{M_v}^v(X_t)\}$ , and a set of  $M_a$  auditory sensory cells with spiking activities  $\{N_t^{a,(1)}, \dots, N_t^{a,(M_a)}\}$  firing at rates  $\{\lambda_1^a(X_t), \dots, \lambda_{M_a}^a(X_t)\}$ . We are interested in calculating the posterior probabilities

$$p_i(t) = \mathbb{P} \left( X_t = s_i \mid N_{[0,t]}^{v,(1)}, \dots, N_{[0,t]}^{v,(M_v)}, N_{[0,t]}^{a,(1)}, \dots, N_{[0,t]}^{a,(M_a)} \right).$$

Extending the unimodal case, it is easy to show that the non-normalized probabilities in this case satisfy

$$\dot{\boldsymbol{\rho}}(t) = \mathbf{Q}^\top \boldsymbol{\rho}(t) + \left( \sum_{m=1}^{M_v} (\Lambda_m^v - \mathbf{I}) \nu_m^v(t) + \sum_{m=1}^{M_a} (\Lambda_m^a - \mathbf{I}) \nu_m^a(t) \right) \boldsymbol{\rho}(t) - \Lambda \boldsymbol{\rho}(t), \quad (4.1)$$

where

$$\begin{aligned} \nu_m^v(t) &= \sum_n \delta \left( t - t_n^{v,(m)} \right) & \Lambda_m^v &= \text{diag}(\lambda_m^v(s_1), \lambda_m^v(s_2), \dots, \lambda_m^v(s_N)) \\ \nu_m^a(t) &= \sum_n \delta \left( t - t_n^{a,(m)} \right) & \Lambda_m^a &= \text{diag}(\lambda_m^a(s_1), \lambda_m^a(s_2), \dots, \lambda_m^a(s_N)) \\ & & \Lambda &= \sum_{m=1}^{M_v} \Lambda_m^v + \sum_{m=1}^{M_a} \Lambda_m^a \end{aligned}$$

Equation (4.1) represents the optimal multisensory computation in this case.

The formulation in the presence of environmental noise is as follows. Consider a state process  $X_t$  that has two projections -  $V_t = F_v(X_t, W_t^v)$  and  $A_t = F_a(X_t, W_t^a)$ . Each of these projections contains partial and noisy information about  $X_t$ . We assume that  $V_t$  and  $A_t$  are independent given  $X_t$ . The input to the visual system is  $V_t$  (instead of  $X_t$ ), namely the visual tuning curves are  $\{\lambda_1^v(V_t), \dots, \lambda_{M_v}^v(V_t)\}$ . Similarly, the input to the auditory system is  $A_t$ , so the auditory tuning curves are  $\{\lambda_1^a(A_t), \dots, \lambda_{M_a}^a(A_t)\}$ . In

other words, we introduce different sources of noise to each of the modalities. Extending the derivation in Section 3 it is easy to show that the filtering equation in this case is,

$$\dot{\boldsymbol{\rho}}(t) = \mathbf{Q}^\top \boldsymbol{\rho}(t) + \left( \sum_{m=1}^{M_v} (\Phi_m^v - \mathbf{I}) \nu_m^v(t) + \sum_{m=1}^{M_a} (\Phi_m^a - \mathbf{I}) \nu_m^a(t) \right) \boldsymbol{\rho}(t) - \Phi \boldsymbol{\rho}(t), \quad (4.2)$$

where

$$\begin{aligned} \eta_m^v(s_i) &= \mathbb{E}[\lambda_m^v(V_t) | X_t = s_i] & \eta_m^a(s_i) &= \mathbb{E}[\lambda_m^a(A_t) | X_t = s_i] \\ \Phi_m^v &= \text{diag}(\eta_m^v(s_1), \eta_m^v(s_2), \dots, \eta_m^v(s_N)) & \Phi_m^a &= \text{diag}(\eta_m^a(s_1), \eta_m^a(s_2), \dots, \eta_m^a(s_N)) \\ \Phi &= \sum_{m=1}^{M_v} \Phi_m^v + \sum_{m=1}^{M_a} \Phi_m^a \end{aligned}$$

Note that this representation is very general. The only assumption made is that  $V_t, A_t$  have *some statistical relationship with  $X_t$* . This allows, for example, for each of the inputs  $V_t$  and  $A_t$  to convey different pieces of information about the state  $X_t$ , with different levels of uncertainty and noise.

## 4.2 Multimodal observations provide more information

The following discussion is qualitative, and aims to provide some intuition for the benefit gained by multimodal processing, especially in the noisy setting. Looking at (4.2) it may seem that having  $M_v$  visual cells and  $M_a$  auditory cells yields the same results as a system with  $M_v + M_a$  sensory cells of the same modality. This claim, however, is incorrect.

Consider two sensory modalities denoted by ‘V’ and ‘A’, each of which observes *noisy* state processes  $\{V_t\}$  and  $\{A_t\}$ , respectively (see Section 3 for a definition). The most accurate information that can be extracted about the state from the sensory spike trains of each modality alone are the probabilities  $\mathbb{P}(X_t = s_i | V_{[0,t]})$  and  $\mathbb{P}(X_t = s_i | A_{[0,t]})$ . In the multisensory case, the ideal state reconstruction is given by  $\mathbb{P}(X_t = s_i | V_{[0,t]}, A_{[0,t]})$ , which is never worse than the reconstruction offered based on a single modality. This occurs because all inputs of the same modality are driven by the same partial noisy information, and therefore the accuracy level of such a unimodal system is restricted. However, adding inputs from a different modality provides a second observation on the same data, and increases the system’s accuracy significantly. We demonstrate this effect in Figure 7, by showing that for a fixed number of sensory cells it is advantageous to split the resources between two sensory modalities rather than using a single modality with the same number of sensory cells<sup>5</sup>.

Consider a tracking task where the number of states is  $N = 51$ , and we use  $M_v = 25$  sensory cells of the V modality. We now add to those sensory cells another group of  $M_a$  cells of a second modality, for varying values of  $M_a$ . The state and noise setup here are similar to those of Fig.6, only now we solve (4.2) instead of (3.3). The solid line represents the empirical MSE of the MMSE optimal estimator in a multimodal network receiving  $M_v$  inputs from the first modality and  $M_a$  inputs from a second modality.

<sup>5</sup>The example provided of this enhancement applies to a specific setup. Establishing general conditions for it to hold is an interesting open question.

Recall that each sensory modality is driven by a different noise process. The dashed line represents the MSE in a unimodal network receiving  $M_v + M_a$  inputs from a single modality. Following the discussion above, since the second modality is driven by a different noise process, it provides a second observation on the process, which for this choice of parameters, improves the system’s accuracy.

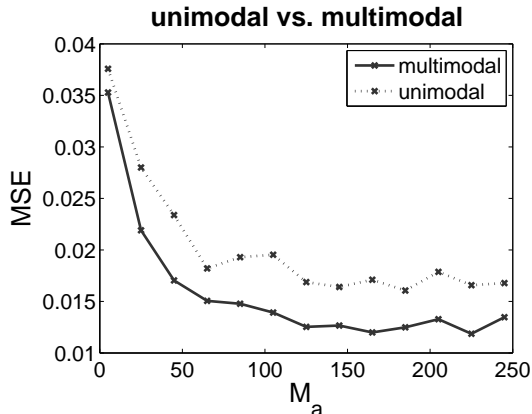


Figure 7: Increasing accuracy by using a second modality. Solid line - Taking  $M_v$  inputs from one modality and  $M_a$  inputs from the other modality. Dashed line - Using  $M_v + M_a$  inputs from a single modality. The noise variance for both modalities was equal to 0.2. We plot the empirical MSE of the optimal estimator for different values of  $M_a$ . The remaining the simulation parameters are  $N = 51$ ,  $M_v = 25$ ,  $\lambda_{\max} = 50$ ,  $\lambda_{\text{base}} = 1$  and  $\alpha = 0.02$ ; results are averaged over five trials.

## 5 The Static Case

In this section we examine the case of a static (yet random) stimulus, in order to gain further insight into the system’s behavior. As stated in Section 1, much earlier work has dealt with this limit, and we show how many previous results are recovered with this setting. Moreover, we present some explicit experimental predictions in this context.

The assumption that the process  $X_t \equiv X$  is constant in time, implies that  $Q = 0$ . Using (A.3) is it easy to see that

$$\rho_i(t) = \rho_i(0) \exp \left( -t \sum_{m=1}^M \lambda_m(s_i) \right) \prod_{m=1}^M (\lambda_m(s_i))^{N_t^{(m)}}, \quad (5.1)$$

where  $N_t^{(m)}$  is the number of spikes arriving from the  $m$ -th input cell during the interval  $[0, t]$ . This result has already been presented in previous work, e.g., (Sanger, 1996).

### 5.1 The Gaussian Case

We examine the static stimulus results, assuming Gaussian tuning functions, and a Gaussian prior. In the unimodal case we show that the optimal estimator can be ap-

proximated by the population vector method (see (Georgopoulos, Kalaska, Caminiti, & Massey, 1982)). We can also show that the optimal multisensory estimator can be approximated by a weighted average of the optimal unimodal estimators. Similar results have been described in previous work (e.g., (Deneve & Pouget, 2004; Witten & Knudsen, 2005)), and are supported by experimental data.

We start by examining the model without noise, and extend the results in the noise model later.

**Unimodal case** Assume that all the tuning functions are shifted versions of the same prototype Gaussian, namely

$$\lambda_m(s) = \lambda_{\max} \exp\left(-\frac{(s - c_m)^2}{2\alpha^2}\right). \quad (5.2)$$

The parameter  $\lambda_{\max}$  represents the maximal firing rate of the cells and will remain constant throughout this discussion. The cell's preferred location is represented by  $c_m$ , where we assume that all the  $c_m$ s are uniformly spread over a given range (i.e.  $c_m = m\Delta c$ ). The tuning function's width is represented by  $\alpha$ . The assumption on the prior is that  $\rho_i(0)$  is a 'discrete Gaussian' of the form

$$\rho_i(0) \propto \exp(-s_i^2/2\sigma_x^2). \quad (5.3)$$

Assuming that the state space spans a wide range, and that  $|s_i - s_{i-1}| \rightarrow 0$ , we can regard  $\sigma_x^2$  as the prior's variance. Applying (5.2) and (5.3) to (5.1) yields

$$\rho_i(t) = c \exp(-s_i^2/2\sigma_x^2) \exp\left(-t \sum_{m=1}^M \lambda_m(s_i)\right) \prod_{m=1}^M \left(\lambda_{\max} \exp\left(-\frac{(s_i - c_m)^2}{2\alpha^2}\right)\right)^{N_t^{(m)}}. \quad (5.4)$$

When the tuning functions are dense enough, the sum in the second exponential is a constant (independent of  $i$ ), and by combining all the other exponentials we obtain a posterior distribution that is still discrete, but its expression is the same as a Gaussian distribution with the following mean and variance

$$\mu = \frac{\sum_{m=1}^M c_m N_t^{(m)}}{\frac{\alpha^2}{\sigma_x^2} + \sum_{m=1}^M N_t^{(m)}} \quad ; \quad \sigma^2 = \left(\frac{1}{\sigma_x^2} + \frac{\sum_{m=1}^M N_t^{(m)}}{\alpha^2}\right)^{-1}. \quad (5.5)$$

Thus, if we consider sufficiently many cells in the network (dense enough  $s_i$ 's), then the mean and variance of the posterior distribution calculated by the system are those in (5.5). Note that both the minimum mean squared error (MMSE) and the maximum a-posteriori (MAP) estimators in this case equal to  $\mu$ . Interestingly, when we take the prior to be flat (i.e.  $\sigma_x \rightarrow \infty$ ) the posterior mean is given by a average of the receptive field centers, weighted by the spiking activity of the corresponding sensory cells, leading to the well known population vector estimator (Georgopoulos et al., 1982). However, as is clear from our analysis, the population vector is optimal only under very restrictive conditions.

**Multimodal case** The Bayesian approach is widely used in the framework of multisensory integration (e.g., (Deneve & Pouget, 2004; Witten & Knudsen, 2005)). Assume we

have a random variable  $X$  observed via two abstract measurements  $V$  and  $A$ , and that given the value of  $X$  the measurements  $V$  and  $A$  are independent. In this case, using Bayes theorem

$$p(X|V, A) = \frac{p(V, A|X)p(X)}{p(V, A)} = \frac{p(V|X)p(A|X)p(X)}{p(V, A)} \propto \frac{p(X|V)p(X|A)}{p(X)}, \quad (5.6)$$

and when the prior over  $X$  is flat we get

$$p(X|V, A) \propto p(X|V)p(X|A). \quad (5.7)$$

Equations (5.6) and (5.7) are extensively used in the multisensory integration literature, and are supported by experimental evidence (see (Witten & Knudsen, 2005)). Using our framework, it is easy to show that in the static case

$$\rho_i^{v,a}(t) = \frac{\rho_i^v(t)\rho_i^a(t)}{\rho_i(0)}, \quad (5.8)$$

analogously to (5.6). Now, assume that the different modalities have different tuning-curve widths, denoted by  $\alpha_v$  and  $\alpha_a$ . Decoding using each modality separately, according to (5.5) we get approximately the following mean and variance for the calculated posterior distributions

$$\begin{aligned} \mu_v &= \frac{\sum_{m=1}^{M_v} c_m N_t^{v,(m)}}{\frac{\alpha_v^2}{\sigma_x^2} + \sum_{m=1}^{M_v} N_t^{v,(m)}} \quad ; \quad \sigma_v^2 = \left( \frac{1}{\sigma_x^2} + \frac{\sum_{m=1}^{M_v} N_t^{v,(m)}}{\alpha_v^2} \right)^{-1} \\ \mu_a &= \frac{\sum_{m=1}^{M_a} c_m N_t^{a,(m)}}{\frac{\alpha_a^2}{\sigma_x^2} + \sum_{m=1}^{M_a} N_t^{a,(m)}} \quad ; \quad \sigma_a^2 = \left( \frac{1}{\sigma_x^2} + \frac{\sum_{m=1}^{M_a} N_t^{a,(m)}}{\alpha_a^2} \right)^{-1}. \end{aligned}$$

Using (5.8), it is easy to show that the posterior distribution produced by the multimodal network is also a Gaussian with the following mean and variance

$$\begin{aligned} \mu_{v,a} &= \left( \frac{1/\sigma_v^2}{1/\sigma_v^2 + 1/\sigma_a^2 - 1/\sigma_x^2} \right) \mu_v + \left( \frac{1/\sigma_a^2}{1/\sigma_v^2 + 1/\sigma_a^2 - 1/\sigma_x^2} \right) \mu_a \\ \sigma_{v,a}^2 &= \frac{1}{1/\sigma_v^2 + 1/\sigma_a^2 - 1/\sigma_x^2}. \end{aligned}$$

This implies that the optimal MMSE (or MAP) estimator in this case is a linear combination of the unimodal optimal estimators  $(\mu_v, \mu_a)$ , where the weight applied to each modality is inversely proportional to its posterior variance. If we assume that the auditory input, for example, supplies no information about the stimulus then  $\sigma_a^2 = \sigma_x^2$ , which leads us back to the unimodal case. Also, taking  $\sigma_x \rightarrow \infty$  (a ‘flat’ prior), yields

$$\mu_{v,a} = \left( \frac{1/\sigma_v^2}{1/\sigma_v^2 + 1/\sigma_a^2} \right) \mu_v + \left( \frac{1/\sigma_a^2}{1/\sigma_v^2 + 1/\sigma_a^2} \right) \mu_a \quad ; \quad \sigma_{v,a}^2 = \frac{1}{1/\sigma_v^2 + 1/\sigma_a^2}. \quad (5.9)$$

The posterior mean estimate based on a weighted mixture of the single modality responses has been experimentally observed (e.g., (Ernst & Banks, 2002; Deneve & Pouget,

2004; Witten & Knudsen, 2005)). One possible application of these ideas concerns the situation where different contrast levels distinguish between the two modalities. In this case the modality with lower contrast will lead to reduced firing activity, and increased variance (see equations above for the dependence of the variance on the spiking activity), thereby reducing its relative contribution.

We now turn to extend the results above to the noisy setting. Note that the only difference between the recursive equations (2.4),(3.3) is the replacement of  $\Lambda_m$  by  $\Phi_m$ . Assuming that the noise follows a ‘discrete Gaussian’ distribution, it is not hard to show that in this case,  $\eta_m(s_i) \propto \lambda_{\max} \exp\left(-\frac{(s_i - c_m)^2}{2(\alpha^2 + \sigma_w^2)}\right)$ . Following the steps described in the unimodal setting above results in a ‘discrete Gaussian’ posterior distribution with the following mean and variance

$$\mu = \frac{\sum_{m=1}^M c_m N_t^{(m)}}{\frac{\alpha^2 + \sigma_w^2}{\sigma_x^2} + \sum_{m=1}^M N_t^{(m)}} \quad ; \quad \sigma^2 = \left( \frac{1}{\sigma_x^2} + \frac{\sum_{m=1}^M N_t^{(m)}}{\alpha^2 + \sigma_w^2} \right)^{-1}. \quad (5.10)$$

Note, that by taking  $\sigma_w = 0$  (no external noise), we get the same result as in (5.5). The results in the multimodal case are similar.

The multi-modal case offers specific predictions for the optimal weighting of different sensory modalities. From (5.9) we see that the optimal posterior mean estimate is given by a weighted average of the unimodal means, weighted by their inverse variances (this result has been established previously). Such a weighted combination seems to be a general feature of multisensory integration (Witten & Knudsen, 2005), and has been observed in both the visual-auditory case (Deneve & Pouget, 2004) and in a visual-haptic setup (Ernst & Banks, 2002). Another interesting feature of our solution relates to the strength of the response of multisensory neurons, as compared to the responses to individual modalities. It has been observed (Stanford, Quessy, & Stein, 2005) that multisensory neurons in the mammalian superior colliculus exhibit an enhanced response when a bimodal stimulus is presented. The enhancement can be super-additive, additive or sub-additive. Moreover, the phenomenon of inverse effectiveness has been observed, whereby neurons which respond weakly to either of two sensory modalities respond super-additively when receiving bimodal input. In the case where both modalities alone respond vigorously, no such enhancement is observed. In Figure 8 we use (5.9) to present typical bimodal responses in two cases, which agree qualitatively with these observations related to cross-modal enhancement; see figure captions for details. These results provide a mechanistic explanation for the phenomenon of cross-modal enhancement, an explanation which has hitherto been lacking (Stanford et al., 2005). Finally, we comment that the relation established between the optimal tuning curve width to the environmental noise level leads to a clear and testable prediction.

## 5.2 Optimal Tuning-Curve Width

In this section we demonstrate the existence of an optimal value for the tuning curve width, which depends on the environment. The intuition behind this is simple. Consider

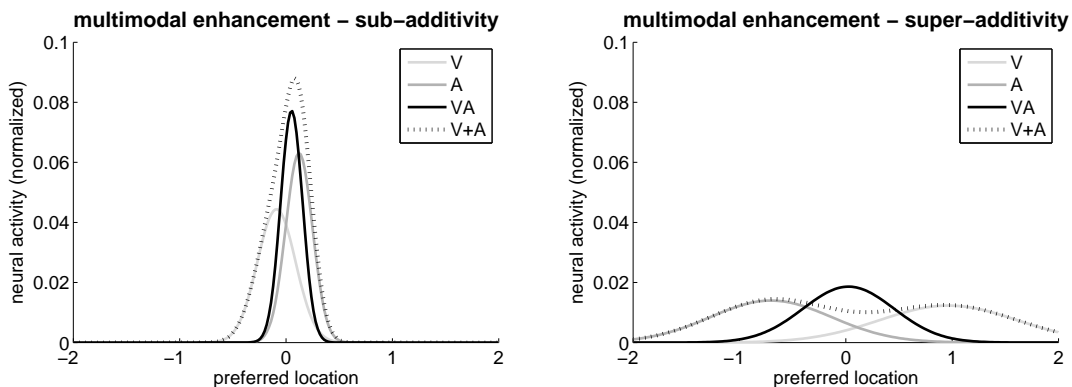


Figure 8: Multimodal Enhancement based on (5.9). (a) Typical Gaussian responses (after normalization) to strong and coherent stimuli. Here, we observe that neurons responding to both modalities exhibit a sub-additive enhancement (i.e., the response to the multimodal stimulus is smaller than the sum of the responses to the unimodal stimuli). (b) Typical Gaussian responses (after normalization) to weak stimuli. Here, we observe that neurons responding to both modalities exhibit a super-additive enhancement (i.e., the response to the multimodal stimulus is larger than the sum of the responses to the unimodal stimuli).

a fixed number of tuning curves covering some finite domain. Narrow tuning curves lead to a low number of spikes, but to good localization of the source once a spike is detected. When the tuning curves are wide, a large number of spikes is detected, while only poor localization can be achieved. Interestingly, we find that the optimal width of the tuning curve is proportional to the noise level.

Similarly to (5.5), we can show that in the presence of additive noise, where the tuning-functions, the prior distribution, and the noise distribution are Gaussians, the posterior distribution computed by the network is a ‘discrete Gaussian’ with the following mean and variance

$$\mu = \frac{\sum_{m=1}^M c_m N_t^{(m)}}{\frac{\alpha^2 + \sigma_w^2}{\sigma_x^2} + \sum_{m=1}^M N_t^{(m)}} \quad ; \quad \sigma^2 = \left( \frac{1}{\sigma_x^2} + \frac{\sum_{m=1}^M N_t^{(m)}}{\alpha^2 + \sigma_w^2} \right)^{-1}. \quad (5.11)$$

where  $\alpha$  is the tuning-curve width,  $\sigma_x^2$  is the prior variance, and  $\sigma_w^2$  is the noise variance. Notice the ambivalent effect of the tuning curve’s width on the posterior variance (which is strongly related to the MSE, as will be shown soon). In the variance expression given by (5.5) we can see that as the width of the tuning-curve increases (larger  $\alpha$ ),  $\sigma^2$  increases, which makes the input less reliable. On the other hand wider tuning-curves cause an increase in the total activity of the cells (since they respond to a wider range of states), thus decreasing the value of  $\sigma^2$ . In this section we explore the system’s performance as a function of the tuning curve width. The time parameter  $t$  will remain *constant* throughout this discussion.

The optimal estimator of  $X$  in the MSE sense, given the observations  $N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}$  is the conditional expectation  $\hat{X} = \mathbb{E} \left[ X \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)} \right]$ . Recalling that both  $\hat{X}$  and  $\sigma^2$  are random variables (as they are functions of the stochastic spike counts), it is easy

to show that

$$\mathbb{E} \left[ \left( X - \hat{X} \right)^2 \right] = \mathbb{E} [\sigma^2].$$

This means that the expectation of  $\sigma^2$  is actually the MSE of the optimal estimator. Thus, it is desirable to choose a value of  $\alpha$  that minimizes  $\mathbb{E}[\sigma^2]$ .

Defining the random variable  $Y = \sum_{m=1}^M N_t^{(m)}$ , we can write the MSE as

$$MSE(\alpha) = \mathbb{E} [\sigma^2] = \mathbb{E} \left[ \left( \frac{1}{\sigma_x^2} + \frac{Y}{\alpha^2 + \sigma_w^2} \right)^{-1} \right]. \quad (5.12)$$

Given the value of  $X$  and the full trajectory of the noise  $\{W_s; 0 \leq s \leq t\}$ , we know that  $N_t^{(m)} \sim \text{Poisson} \left( \int_0^t \lambda_m(X, W_s) ds \right)$ , and that  $N_t^{(1)}, \dots, N_t^{(M)}$  are independent. Therefore,  $Y$  is also a Poisson random variable, with rate parameter which can be shown (assuming a large value of  $M$  and a small value of  $\Delta c$ ) to be given by

$$\lambda = \sum_m \int_0^t \lambda_m(X, W_s) ds \approx \lambda_{\max} t \frac{\sqrt{2\pi}\alpha}{\Delta c}.$$

Note that this value is independent of the specific realization of  $X$  and  $W_{[0,t]}$ , and therefore for the current discussion we shall treat  $Y$  as being a Poisson random variable, with a constant parameter  $\lambda = \lambda_{\max} t \frac{\sqrt{2\pi}\alpha}{\Delta c}$ .

Evaluating the expression in (5.12) analytically is complicated, however we can approximate it numerically. In Figure 9 we plot the MSE as a function of the width  $\alpha$  for different values of environmental noise level ( $\sigma_w$ ). As we can see in Figure 9(b)-9(c), there is indeed an optimal value for the width  $\alpha$ , which increases monotonically with the noise level. In other words, as the noise level increases, the tuning curve must adapt and increase its width accordingly. Such a result is of ecological significance, as it relates the properties of the environment (noise level) to the optimal system properties (tuning curve width).

## 6 History Dependent Point Processes and Sensory Adaptation

So far we assumed that the input spike trains emitted by the sensory cells are Poisson spike trains, with state-dependent rate functions  $\lambda_m(X_t)$ . Poisson spike trains serve as a convenient model that is often used for mathematical tractability. However, this assumption falls short from providing an adequate model for many well known biophysical phenomena such as refractoriness and adaptation. In this section we introduce a larger family of processes, which, similarly to Poisson processes, are characterized by a rate function, except that this rate function depends on the history of the process itself. This class of processes is referred to as Self Exciting Point Processes in (Snyder & Miller,

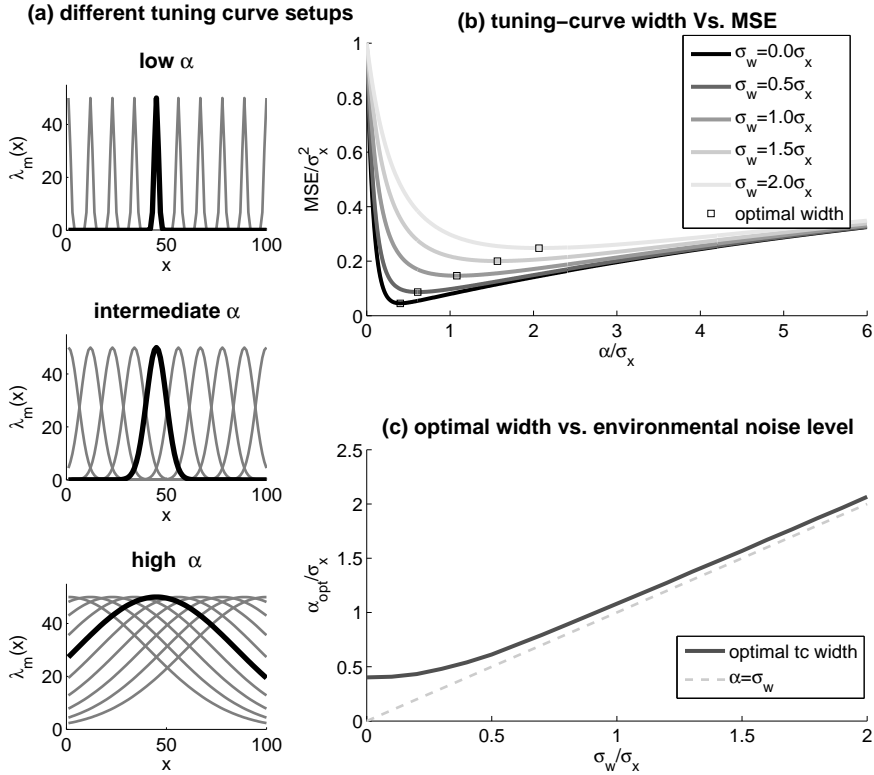


Figure 9: Determining the optimal tuning curve. (a) Sample tuning curves with different widths covering the input domain. (b) The MSE as a function of the tuning curve width, computed using (5.12) (normalized by the prior standard deviation  $\sigma_x$ ), for increasing levels of environmental noise, with  $\Delta c = 10^{-3}$ ,  $\lambda_{\max} = 50$ ,  $t = 10^{-3}$  (c) The optimal value of  $\alpha$ , as a function of the environmental noise level  $\sigma_w^2$ .

1991); however, in order to conform to the nomenclature in the neuroscience literature (e.g., (Eden et al., 2004)) we use the term *history-dependent point processes*, which we abbreviate as HDPP. This family contains the Poisson, general renewal, and other more complex processes. Using such processes we can model complex biophysical phenomena. Surprisingly, assuming general history dependent point processes as inputs instead of Poisson inputs, yields similar results to the ones that were presented throughout this paper. history dependent point processes have been used previously in (Barbieri et al., 2004; Eden et al., 2004) in the context of a discrete time approximation to optimal filtering.

An important motivation for weakening the assumption of Poisson firing relates to adaptation. Adaptation is a ubiquitous phenomenon in sensory systems, whereby a system changes its response properties as a function of the environmental stimuli (see (Wark, Lundstrom, & Fairhall, 2007) for a recent review). In some cases adaptation is shown to improve performance and reduce ambiguity (e.g., (Fairhall, Lewen, Bialek, & Ruyter Van Steveninck, 2001; Sharpee et al., 2006)). We have already seen an example of adaptation in Section 3, where we showed that improved performance in the face of

increasing noise can be obtained by modifying the properties of the cells' tuning curves. However, we did not consider dynamic mechanisms for achieving adaptation. In this section we allow the sensory cells' responses to change dynamically depending on their past behavior. We show that adaptation indeed leads to improved performance when resources are limited. More specifically, we show that given energetic constraints (e.g., a limit on the number of spikes fired within a given interval), adaptation outperforms a naive approach which does not use adaptation.

The precise mathematical definition of HDPPs which can be characterized by a *conditional rate function* can be found in (Snyder & Miller, 1991). Such processes extend Poisson processes in allowing the rate function to depend on the history of the process. The conditional rate function (a.k.a. intensity process) is given by

$$\lambda(t, N_{[0,t]}) \triangleq \lim_{\delta \rightarrow 0} \frac{1}{\delta} \mathbb{P}(N_{(t,t+\delta]} = 1 | N_{[0,t]}). \quad (6.1)$$

In analogy to the definition of doubly stochastic Poisson processes, we define the *doubly stochastic history dependent point process* as a HDPP for which the intensity has some stochastic element (other than the history).

## 6.1 Filtering a CFMP from HDPP Observations

We aim here to weaken the assumption about the sensory activity. Instead of assuming that each spike train is a DSPP with rate function  $\lambda_m(X_t)$ , we assume now that each spike train is a doubly stochastic HDPP with state-dependent intensity process  $\lambda_m(t, N_{[0,t]}, X_t)$ . To simplify the notation we will use the abbreviation  $\lambda_m(t, X_t)$ . In this case, similar derivation to the Poisson case (see appendix D) leads to the following set of equations computing the posterior distribution

$$\dot{\rho}_i(t) = \sum_k q_{ki} \rho_k(t) + \left( \sum_{m=1}^M (\lambda_m(t, s_i) - 1) \nu_m(t) \right) \rho_i(t) - \lambda(t, s_i) \rho_i(t), \quad (6.2)$$

where  $\lambda(t, s_i) = \sum_{m=1}^M \lambda_m(t, s_i)$ . This equation is similar to (2.3), except that here the efficacy of the input depends on its history in addition to the current state. Next, we show how adaptation can be captured within this general framework.

## 6.2 Application to Sensory Adaptation

As in the Poisson case, we interpret the set of equations in (6.2) as representing the activity of a recurrent neural network, where the synaptic weights are represented by the values of  $q_{ki}$ . In contrast to the Poisson setting, here the efficacy of the input is time-dependent rather than constant. A spike arriving from the  $m$ -th sensory cell at time  $t_n^{(m)}$  affects the network according to the time-dependent tuning-curve  $\lambda_m(t_n^{(m)}, s_i)$ .

The present framework significantly expands the class of processes that can be handled by the model, and the types of phenomena that can be examined. As an example

we demonstrate how a simple adaptation mechanism can be implemented within this framework. Using this model we can show not only that adaptation can be used to reduce the total number of spikes emitted by the sensory cells (energy saving) without degrading performance, but also that in some cases adaptation helps in improving the system's precision.

To model adaptation we use the rate function

$$\lambda_m(t, s_i) = \mu_m(t)\phi_m(s_i) \quad (6.3)$$

where  $\phi_m(s_i) = \lambda_{\max} \exp(-(s_i - c_m)/2\alpha^2)$  is a deterministic Gaussian state response, and  $\mu_m(t)$  is the adaptation factor (a similar model was used in (Berry & Meister, 1998)). The variable  $\mu_m(t)$  obeys the following dynamics:

- When no spikes are emitted by the  $m$ -th cell,

$$\tau \dot{\mu}_m(t) = 1 - \mu_m(t).$$

This leads to an exponential recovery to 1 with a time-scale of  $\tau$ .

- When the  $m$ -th cell emits a spike,  $\mu_m(t)$  is updated according to

$$\mu_m(t^+) = [\mu_m(t^-) - \Delta]_+,$$

where  $x_+ = \max(0, x)$ . In other words, each spike reduces the firing potential of a cell, a phenomenon referred to as spike rate adaptation.

This adaptation scheme decreases the firing-rate of sensory cells that fire most intensively, while the others are hardly affected. The parameters  $\tau$  and  $\Delta$  control the speed and strength of the adaptation process.

Note, that upon the arrival of the  $n$ -th spike from the  $m$ -th sensory cell,  $\rho_i(t)$  is multiplied by  $\lambda_m(t_n^{(m)}, s_i) = \mu_m(t_n^{(m)})\phi_m(s_i)$ . Since the term  $\mu_m(t_n^{(m)})$  is independent of  $i$ , and constant multiplication does not affect the normalized distribution, we can write the recursive equation in this case as

$$\dot{\rho}_i(t) = \sum_k q_{ki} \rho_k(t) + \left( \sum_{m=1}^M (\phi_m(s_i) - 1) \nu_m(t) \right) \rho_i(t) - \lambda(t, s_i) \rho_i(t), \quad (6.4)$$

where  $\lambda(t, s_i) = \sum_{m=1}^M \lambda_m(t, s_i)$ .

Next, we demonstrate that a non-adapting system with a *fixed* rate leads to inferior performance with respect to an adaptive system, which fires far fewer spikes. Figure 10 compares the performance of the system with and without adaptation, for different values of  $\Delta$  and  $\tau$ . We examine two parameters - the *spike-count ratio* and the *MSE ratio*, defined by

$$\begin{aligned} \text{spike-count ratio} &= \frac{\text{total spike-count } \{\Delta, \tau\}}{\text{total spike-count } \{\text{no adaptation}\}} \\ \text{MSE ratio} &= \frac{\text{MSE } \{\Delta, \tau\}}{\text{MSE } \{\text{no adaptation}\}}. \end{aligned}$$

In Figure 10(a) we see that increasing the level of adaptation (by increasing either  $\tau$  or  $\Delta$ ) reduces the total amount of spikes emitted by the system. In Figure 10(b) we observe that in spite of the lower spiking activity, the error in the adapting system does not increase, on the contrary - in most cases it even decreases. Figures 10(a) and 10(b) represent a temporal average over a window of 10 seconds, averaged over 25 different realizations. This phenomenon can be explained by examining the self-inhibition term represented by  $\lambda(t, s_i)$ . The dashed line in Figure 10(c) represents  $\lambda(t, s_i)$  (for a constant value of  $t$ ) without adaptation. In this case,  $\lambda$  is nearly constant, slightly decreasing at the edges. This implies that the variables  $\rho_i$  corresponding to states near the edge of the domain exhibit weaker self inhibition. Thus, when no spikes arrive these posterior cells dominate the others, implying that low sensory spiking activity strengthens the belief that the stimulus is outside the network’s coverage area. However, when an adaptation mechanism is introduced (the solid line in Figure 10(c)),  $\lambda(t, s_i)$  displays a local minimum in the neighborhood of the true state, which implies that the posterior cells in this neighborhood will exhibit weaker self inhibition. This effect helps to maintain higher probability for the true state even though the sensory activity decays.

## 7 Extensions and Comparisons to Previous Work

### 7.1 Prediction

The network discussed so far deals with estimating the current state of the world. However, in order to perform real-time tasks and act in a dynamic environment, an organism must be able to make predictions about future world states. In this section we show how our framework can be easily adjusted to predict future states.

In appendix E we present a simple extension of our framework to compute the non-normalized future probability vector denoted by  $\boldsymbol{\rho}^\tau(t)$ , which represents the posterior probability distribution for the state at time  $t + \tau$  based on the sensory data available up to time  $t$ . The prediction equation in this case is

$$\dot{\boldsymbol{\rho}}^\tau(t) = \mathbf{Q}^\top \boldsymbol{\rho}^\tau(t) + \left( \sum_{m=1}^M (\Lambda_m^\tau - \mathbf{I}) \nu_m(t) \right) \boldsymbol{\rho}^\tau(t) - \Lambda^\tau \boldsymbol{\rho}^\tau(t), \quad (7.1)$$

where

$$\Lambda_m^\tau = (\mathbf{P}^{(\tau)})^\top \Lambda_m (\mathbf{P}^{(\tau)})^{-\top} \quad ; \quad \Lambda^\tau = (\mathbf{P}^{(\tau)})^\top \Lambda (\mathbf{P}^{(\tau)})^{-\top}$$

and  $(\cdot)^{-\top}$  represents the inverse of the transpose. Note that (7.1) is identical to (2.4) except for a change in the tuning curve matrices  $\Lambda_m$ . The underlying connectivity structure of the posterior network, given by the matrix  $\mathbf{Q}$ , is unchanged.

### 7.2 Computing the Log-Posterior Distribution

Another approach often used in the field of neural decoding is based on computing the logarithm of the posterior probabilities, rather than the probabilities themselves (e.g.,

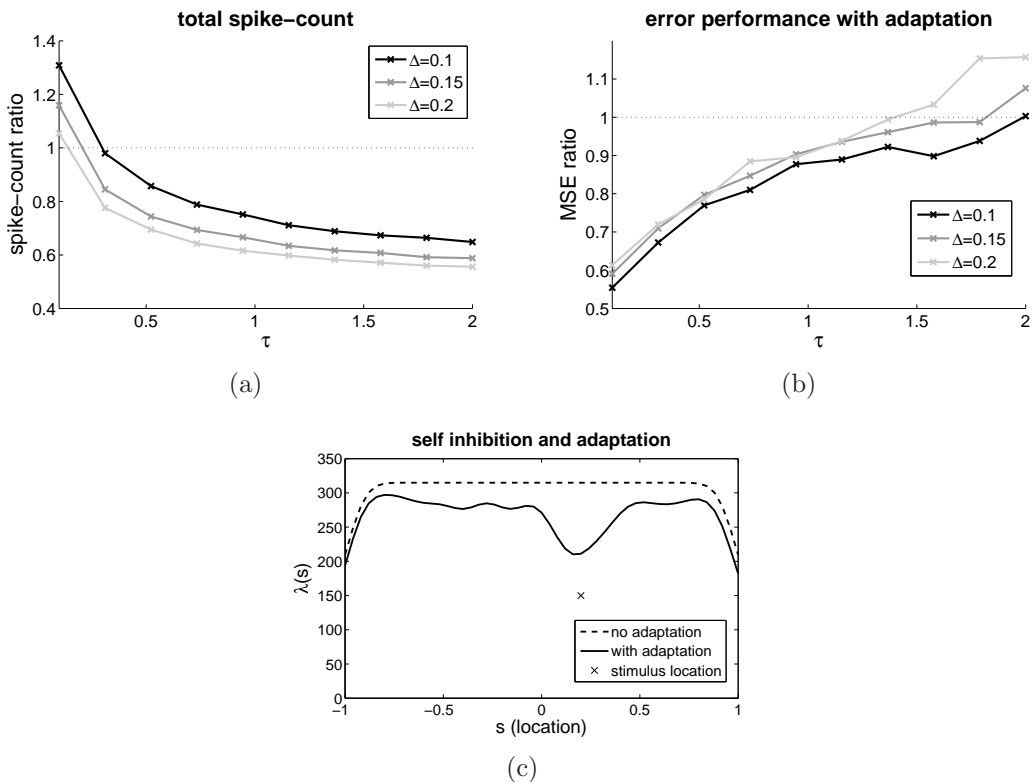


Figure 10: Comparing the performance of the system with and without adaptation, for different values of  $\tau$  and  $\Delta$ . (a) Comparing the total spiking activity, by examining the spike-count ratio. Clearly, for most values of  $\tau$ , the overall number of spikes in the adaptation model is significantly lower than the non-adaptive model. The simulation parameters are  $N = 101$ ,  $M = 50$ ,  $\lambda_{\text{base}} = 0$ ,  $\alpha = 0.04$ ,  $\beta = 5$  and  $\mu = 5$ . In the non-adapting model we use  $\lambda_{\text{max}} = 40$ , while in the adapting model  $\lambda_{\text{max}} = 75$ . (b) Comparing the MSE performance, by examining the MSE ratio. Evidently, for most values of  $\Delta$  and  $\tau$ , the MSE ratio is less than 1, implying that the adapting system is more accurate (even though less spikes are emitted). (c) A sample of the self inhibition term  $\lambda(t, s_i) = \sum_m \lambda_m(t, s_i)$ . Solid line - the self inhibition term in the adaptation model. One can observe a local minimum where the stimulus is located. Dashed line - the self inhibition term without adaptation. Here  $\lambda$  is nearly constant, with no preference to the current state.

(Rao, 2004, 2006) ). Using the framework suggested in this paper, it is easy to derive the filtering equations computing the log of the non-normalized probabilities denoted by  $r_i(t) = \log \rho_i(t)$ . The derivation, presented appendix F, yields the following set of equations

$$\dot{r}_i(t) = \sum_{k=1}^N q_{ki} \exp(r_k(t) - r_i(t)) + \sum_{m=1}^M \log(\lambda_m(s_i)) \nu_m(t) - \lambda(s_i). \quad (7.2)$$

One advantage of the log representation is that the input arriving from the sensory cells contributes *linearly* to the evolution of  $r_i$  (instead of multiplicatively as in (2.3)). However, the recurrent interaction between the different elements is *non-linear* (as the  $r_k$  variables appears in an exponent). Note also that the periodic normalization required

to retain stability in (2.3) renders the equations nonlinear.

Considering a binary process  $X_t \in \{0, 1\}$ , and denoting

$$L_t = \log \frac{\mathbb{P}\left(X_t = 1 \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right)}{\mathbb{P}\left(X_t = 0 \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right)},$$

then  $L_t = r_1(t) - r_0(t)$ , and thus, using (7.2), it is easy to show that

$$\dot{L}_t = q_{01} (1 + e^{-L_t}) - q_{10} (1 + e^{L_t}) + \sum_{m=1}^M \log \left( \frac{\lambda_m(s_1)}{\lambda_m(s_0)} \right) \nu_m(t) - (\lambda(s_1) - \lambda(s_0)).$$

This equation was introduced in a recent paper (Deneve, 2008), where it was suggested that a single-cell computes this log-ratio.

### 7.3 Comparison to Previous Work

In order to place our contribution in context we briefly review existing work, and then stress the novelty of the approach taken here. We would like to stress that our main motivation has been on the real time implementation of state estimation by a neural network in continuous time. While a fair amount of work has been devoted to developing effective discrete time numerical schemes for filtering in the context of neural decoding (e.g., (Barbieri et al., 2004; Eden et al., 2004; Shoham et al., 2005)), our aim differs in suggesting a scheme which can be implemented by a neural system. Moreover, we have provided significant extensions of this framework (summarized in Section 1), which, to the best of our knowledge, have not appeared previously in the literature.

As stated in the first part of the paper, a large fraction of the work to date has focused on static stimuli. The work most closely related to ours, and dealing with time-dependent phenomena, appears in (Deneve et al., 2001; Rao, 2004, 2006; Barbieri et al., 2004; Eden et al., 2004; Huys et al., 2007). Rao (Rao, 2004) proposed a mechanism for representing time-varying probability distributions in the neurons' activity patterns, where the network's connectivity structure and intrinsic dynamics are responsible for performing the required computation. Rao's networks (Rao, 2004) use linear dynamics and discrete time to approximately compute the log-posterior distributions from noisy continuous inputs (rather than actual spike trains). However, the domain of applicability of the approximations suggested in (Rao, 2004) is unclear (see (Beck & Pouget, 2007) for further discussion). The work of (Rao, 2006) suggests a discrete time formulation for state estimation in the log probability domain, and proposes several physiological mechanisms for implementing these algorithms. The work suggest performing exact inference in nonlinear networks, and approximate inference in linear networks. Barbieri et al. (Barbieri et al., 2004) consider a continuous state linear dynamical system in continuous time, observed through an inhomogeneous point process model, and develop a recursive decoding algorithm. This derivation requires time discretization and the approximation of the posterior distribution by a Gaussian. No neural implementation

is offered for their algorithm. Beck and Pouget (Beck & Pouget, 2007), probably closest in spirit to our work, introduced networks in which the neurons directly represent and compute approximate posterior probabilities (rather than their logarithms as in (Rao, 2004)) from discrete-time approximate firing rate inputs, using non-linear mechanisms such as multiplicative interactions and divisive normalization. The main distinctions with the work presented here are: (i) The state estimation is approximate; (ii) a time discretization process is used in the derivation, thereby limiting the class of observation processes for which the results hold (see Eq. (2.7) in (Beck & Pouget, 2007)). Huys et al. (Huys et al., 2007) have recently suggested an exact solution to spike train decoding of real-valued states evolving in continuous time, and observed through Poisson spike trains. This work assumes a Gaussian process prior over input trajectories and Gaussian tuning curves, facilitating an exact calculation of the posterior state distribution. The optimal state estimator derived is nonlinear and cannot be implemented in real time. Moreover, no neural architecture is offered for state estimation (but, see (Zemel, Huys, Natarajan, & Dayan, 2005) for an approximate neural solution). However, this work suggested a very interesting approach to dynamic spike train decoding through population codes, based on different underlying assumptions than those used here. Finally, we comment on two recent papers suggesting a solution to the problem of dynamic spike train decoding. The work of Deneve (Deneve, 2008) considers a binary state process and derives the differential equation for the logarithm of the ratio of the posterior probabilities. As shown in Section 7 equation (2.4) can be manipulated to establish Deneve’s results. Another derivation of the filtering equation appeared recently in (Pitkow et al., 2007), where the authors suggested a mechanism by which the visual system may identify objects in spite of retinal jitter. The retinal motion was modeled as a two-dimensional random walk, which is a special case of the hidden Markov model in the present formulation. We note that the object in (Pitkow et al., 2007) was static, while the dynamics was due to the random jitter of the retina. In fact, it is possible to use the formulation presented in Section 3 to extend this work to the detection of moving objects.

In summary, we have developed an approach which leads to exact dynamic spike train decoding using neural networks operating in real time. The framework is sufficiently general to deal effectively with noise, multimodality, adaptation and prediction within a single unified setting. However, in the context of continuous state spaces our approach is approximate since, in this case, it would be based on discretization. Since sensory systems possess a finite resolution, we do not think this is a major restriction, but this is clearly an issue which warrants further detailed study. As far as we are aware none of the work in the literature, which deals with continuous state spaces, has suggested exact continuous time online neural solutions. We view this issue as a significant open question for future research in neural decoding.

## 8 Discussion

We have presented a mathematically rigorous and biologically plausible approach to state estimation and prediction based on computation by spiking neurons. The estima-

tion is performed by a recurrent neural network incorporating prior information through its synaptic weights, and receiving a continuous stream of environmental data through spiking sensory cells. The proved optimality of the solution under rather general conditions, and its implementation through a neural network, provide a solid foundation for future extensions. The approach extends very naturally to noisy inputs, to multimodal observations, and to non-Poisson spiking processes. In the static limit we recover well known and experimentally verified solutions.

We view this work as providing only a first step in a long road towards understanding the means by which a neural system can use spiking sensory activity to estimate and predict environmental states in real time. A recent example (Pitkow et al., 2007) presents an application of this framework to the decoding of retinal signals. Many extensions can be envisaged, of which we mention only three. First, it is important to provide detailed biophysical mechanisms by which the decoding network dynamics can be implemented by real biological neurons and synapses, rather than the more abstract neurons described in this work. Second, we have assumed prior information is available through the transition matrix  $Q$ . Ideally, this matrix, forming the synaptic connectivity in (2.3) should be learned on a slow time scale through interacting with the environment, and the utilization of biologically plausible synaptic plasticity rules. Finally, our current formulation leads to a representation in which each neuron corresponds to a single environmental state (grandmother cell representation). An open question relates to the derivation of a distributed state representation, which is optimal in terms of state estimation, similarly to the current network, while offering further robustness and error-correction abilities characteristic of distributed representations.

**Acknowledgment** The work of RM was partially supported by a grant from the Center for Complexity Science.

# Appendix

## A Closed-Form Solution

In this section we present a closed-form solution to the filtering equation (2.4). The solution will be later used to analyze the system's behavior.

Denote by  $t_i$  the time of arrival of the  $i$ -th event in the combined group of processes  $\{N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\}$ , and by  $m(t_i)$  the index of the input process emitting the event at  $t_i$ . For every  $t$  in  $[t_i, t_{i+1})$  (during which no events occur) we have

$$\dot{\boldsymbol{\rho}}(t) = (\mathbf{Q}^\top - \Lambda) \boldsymbol{\rho}(t).$$

The solution to this equation is

$$\boldsymbol{\rho}(t) = e^{(\mathbf{Q}^\top - \Lambda)(t-t_i)} \boldsymbol{\rho}(t_i) \quad t_i \leq t < t_{i+1}. \quad (\text{A.1})$$

When an event is emitted at time  $t_{i+1}$  by input  $m(t_{i+1})$  the vector  $\boldsymbol{\rho}$  is instantly updated according to

$$\boldsymbol{\rho}(t_{i+1}) = \boldsymbol{\rho}(t_{i+1}^-) + (\Lambda_{m(t_{i+1})} - \mathbf{I}) \boldsymbol{\rho}(t_{i+1}^-) = \Lambda_{m(t_{i+1})} \boldsymbol{\rho}(t_{i+1}^-), \quad (\text{A.2})$$

Combining (A.1) and (A.2) yields the closed-form solution to (2.4),

$$\boldsymbol{\rho}(t) = e^{(\mathbf{Q}^\top - \Lambda)(t-t_n)} \left( \prod_{i=1}^n \Lambda_{m(t_i)} e^{(\mathbf{Q}^\top - \Lambda)(t_i-t_{i-1})} \right) \boldsymbol{\rho}(0), \quad (\text{A.3})$$

where  $t_n$  is the last event in the interval  $[0, t]$ , and  $t_0 = 0$ .

## B Construction of the transition matrix in Section 2.1.1

We describe the construction of the matrix  $\mathbf{Q}$  for the example in section 2.1.1 where the 'direction' and 'visibility' modes are contained in the state definition (see Figure 3).

Figure 11 demonstrates the  $\mathbf{Q}$  matrix for 5 possible locations. As we have 2 possible directions and 2 possible visibility modes, there are  $5 \times 2 \times 2 = 20$  states overall. We number the composite states as follows: states 1–5: {visible, up}; states 6–10: {visible, down}; states 11–15: {invisible, up}; states 16–20: {invisible, down}.

The dark blue squares along the diagonal represent the average time spent in each state (as in any generator matrix). The orange squares describe a transition to the next location (depending on the movement direction) without changing the direction or the visibility mode. For example, state #2 = (2, up, visible) to state #3 = (3, up, visible). The yellow squares describe changing the movement direction, and the location

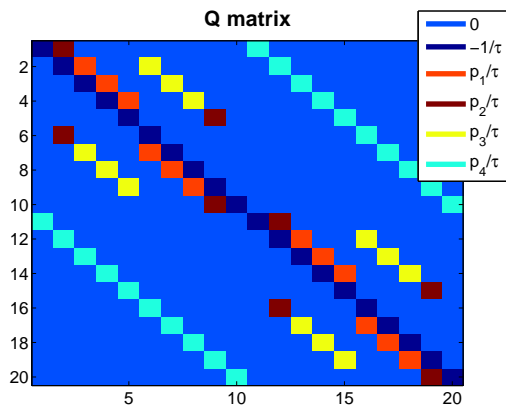


Figure 11: The matrix  $Q$  chosen for the example of Figure 3

accordingly. The visibility is unchanged. For example: state #2 = (2, up, visible) to state #6 = (1, down, visible). The brown squares describe a transition from the edge locations (1 or 5). For example, state #1 = (1, up, visible) to #2 = (2, up, visible). The probability represented in brown is the sum of the probabilities from the yellow and orange squares. Finally, the light-blue squares describe visibility changes. The location and direction remains unchanged. For example, state #2 = (2, up, visible) to state #12 = (2, up, invisible).

## C Noise Model

We present the derivation of (3.1) for the non-normalized posterior distribution in the case of noisy observations. Consider a CFMP  $X_t$  passing through a noisy channel. This noise channel introduces an interference process  $W_t$ . We model the sensory tuning curves in this case as a function of the combined process  $\tilde{X}_t = (X_t, W_t)$ , namely  $\lambda_m(\tilde{X}_t)$ . We assume that  $W_t$  is a CFMP with state-space  $\mathcal{W} = \{w_1, \dots, w_L\}$ .

**Notation:** Throughout this section, terms related to the combined process  $\tilde{X}_t$  are annotated with the tilde symbol  $\sim$ .

The derivation in the noisy case relies on the observation that the combined process  $\tilde{X}_t$  is also a CFMP (with  $N \times L$  states). Therefore, we can apply equation (2.4) to compute the non-normalized probabilities  $\tilde{\rho}_{\tilde{i}}(t) \propto \mathbb{P}\left(\tilde{X}_t = \tilde{s}_{\tilde{i}} \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right)$ , where  $\tilde{i} = (i_x, i_w) \in \mathbb{N}^2$  is a double-entry index, and  $\tilde{s}_{\tilde{i}} = (s_{i_x}, w_{i_w})$  represents a combined state. To obtain  $X_t$ 's non-normalized posterior distribution, we compute the marginal non-normalized probabilities -  $\rho_i(t) = \sum_{i_w} \tilde{\rho}_{i, i_w}(t)$ .

## C.1 Filtering the Combined Process (State & Noise)

As mentioned previously, the process  $\tilde{X}_t = (X_t, W_t)$  is a CFMP with  $N \times L$  different states. Thus, using (2.3) we have

$$\frac{d}{dt} \tilde{\rho}_i(t) = \sum_{k=1}^N \tilde{q}_{\tilde{k}, \tilde{i}} \tilde{\rho}_{\tilde{k}}(t) + \left( \sum_{m=1}^M (\lambda_m(\tilde{s}_i) - 1) \nu_m(t) \right) \tilde{\rho}_i(t) - \lambda(\tilde{s}_i) \tilde{\rho}_i(t). \quad (\text{C.1})$$

where  $\tilde{q}_{\tilde{k}, \tilde{i}}$  is an element in  $\tilde{X}_t$ 's generator matrix  $\tilde{Q} \in \mathbb{R}^{NL \times NL}$ .

In order to simplify the equations, we explore the structure of  $\tilde{Q}$ . Denoting the transition probabilities of the process  $\tilde{X}_t$  by  $\tilde{\mathbf{P}}_{\tilde{k}\tilde{i}}^{(\tau)} = \mathbb{P} \left( \tilde{X}_{t+\tau} = (s_{i_x}, w_{i_w}) \mid \tilde{X}_t = (s_{k_x}, w_{k_w}) \right)$ , yields

$$\tilde{q}_{\tilde{k}, \tilde{i}} = \left. \frac{d}{d\tau} \tilde{\mathbf{P}}_{\tilde{k}\tilde{i}}^{(\tau)} \right|_{\tau=0}.$$

However, recalling that  $X_t$  and  $W_t$  are independent, clearly  $\tilde{\mathbf{P}}_{\tilde{k}\tilde{i}}^{(\tau)} = \mathbf{P}_{k_x i_x}^{(\tau)} \mathbf{P}_{k_w i_w}^{w, (\tau)}$ , where  $\mathbf{P}_{k_w i_w}^{w, (\tau)}$  is the transition probability of the noise process  $W_t$ . Therefore

$$\begin{aligned} \tilde{q}_{\tilde{k}, \tilde{i}} &= \left. \frac{d}{d\tau} \left( \mathbf{P}_{k_x i_x}^{(\tau)} \mathbf{P}_{k_w i_w}^{w, (\tau)} \right) \right|_{\tau=0} \\ &= \left( \frac{d}{d\tau} \mathbf{P}_{k_x i_x}^{(\tau)} \mathbf{P}_{k_w i_w}^{w, (\tau)} + \mathbf{P}_{k_x i_x}^{(\tau)} \frac{d}{d\tau} \mathbf{P}_{k_w i_w}^{w, (\tau)} \right) \Big|_{\tau=0} \\ &= q_{k_x, i_x} \mathbf{P}_{k_w i_w}^{w, (0)} + \mathbf{P}_{k_x i_x}^{(0)} q_{k_w, i_w}^w \\ &= q_{k_x, i_x} \delta_{k_w, i_w} + \delta_{k_x, i_x} q_{k_w, i_w}^w. \end{aligned} \quad (\text{C.2})$$

Substituting (C.2) into (C.1) yields

$$\begin{aligned} \frac{d}{dt} \tilde{\rho}_i(t) &= \sum_{\tilde{k}} (q_{k_x, i_x} \delta_{k_w, i_w} + \delta_{k_x, i_x} q_{k_w, i_w}^w) \tilde{\rho}_{\tilde{k}}(t) + \left( \sum_{m=1}^M (\lambda_m(\tilde{s}_i) - 1) \nu_m(t) \right) \tilde{\rho}_i(t) - \lambda(\tilde{s}_i) \tilde{\rho}_i(t) \\ &= \sum_{k_x} q_{k_x, i_x} \tilde{\rho}_{k_x, i_w}(t) + \sum_{k_w} q_{k_w, i_w}^w \tilde{\rho}_{i_x, k_w}(t) + \left( \sum_{m=1}^M (\lambda_m(\tilde{s}_i) - 1) \nu_m(t) \right) \tilde{\rho}_i(t) - \lambda(\tilde{s}_i) \tilde{\rho}_i(t). \end{aligned} \quad (\text{C.3})$$

The recursive set of equations in (C.3) computes the posterior distribution of the joint (state and noise) process  $\tilde{X}_t$ . However, since the system is required to estimate the state, rather than the noise, we show how to reduce this set of equations so that the posterior state distribution is obtained.

## C.2 Filtering the State Process

Recalling that  $\rho_i(t) = \sum_{i_w} \tilde{\rho}_{i,i_w}(t)$  we can use (C.3) to derive a differential equation for  $X_t$ 's non-normalized posterior distribution,

$$\begin{aligned}
\dot{\rho}_i(t) &= \sum_{i_w} \frac{d}{dt} \tilde{\rho}_{i,i_w}(t) \\
&= \underbrace{\sum_{k_x, i_w} q_{k_x, i} \tilde{\rho}_{k_x, i_w}(t)}_{S_1} + \underbrace{\sum_{k_w, i_w} q_{k_w, i_w}^w \tilde{\rho}_{i, k_w}(t)}_{S_2} \\
&\quad + \sum_{m=1}^M (\nu_m(t) - 1) \underbrace{\sum_{i_w} (\lambda_m(s_i, w_{i_w}) - 1) \tilde{\rho}_{i, i_w}(t)}_{S_3} - \underbrace{\sum_{i_w} \lambda(s_i, w_{i_w}) \tilde{\rho}_{i, i_w}(t)}_{S_4}
\end{aligned} \tag{C.4}$$

Examining the first two sums in (C.4) yields

$$S_1 = \sum_{k_x} q_{k_x, i} \sum_{i_w} \tilde{\rho}_{k_x, i_w}(t) = \sum_{k_x} q_{k_x, i} \rho_{k_x}(t), \tag{C.5}$$

$$S_2 = \sum_{k_w} \tilde{\rho}_{i, k_w}(t) \underbrace{\sum_{i_w} q_{k_w, i_w}^w}_{=0} = 0. \tag{C.6}$$

In order to simplify  $S_3$  recall that

$$\mathbb{P}\left(W_t = w_{i_w} \mid X_t = s_i, N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right) = \frac{\mathbb{P}\left(W_t = w_{i_w}, X_t = s_i \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right)}{\mathbb{P}\left(X_t = s_i \mid N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right)} = \frac{\tilde{\rho}_{i, i_w}(t)}{\rho_i(t)},$$

or alternatively

$$\tilde{\rho}_{i, i_w}(t) = \rho_i(t) \mathbb{P}\left(W_t = w_{i_w} \mid X_t = s_i, N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right).$$

Therefore,

$$\begin{aligned}
S_3 &= \rho_i(t) \sum_{i_w} (\lambda_m(s_i, w_{i_w}) - 1) \mathbb{P}\left(W_t = w_{i_w} \mid X_t = s_i, N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right) \\
&= \rho_i(t) \left( \mathbb{E}\left[\lambda_m(X_t, W_t) \mid X_t = s_i, N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right] - 1 \right).
\end{aligned}$$

Denoting  $\eta_m(s_i, t) = \mathbb{E}\left[\lambda_m(X_t, W_t) \mid X_t = s_i, N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)}\right]$ , we can write

$$S_3 = (\eta_m(s_i, t) - 1) \rho_i(t). \tag{C.7}$$

Similarly, one can show that

$$S_4 = \eta(s_i, t) \rho_i(t), \tag{C.8}$$

where  $\eta(s_i, t) = \sum_{m=1}^M \eta_m(s_i, t)$ .

Substituting (C.5)-(C.8) into (C.4), yields

$$\dot{\rho}_i(t) = \sum_{k=1}^N q_{ki} \rho_k(t) + \left( \sum_{m=1}^M (\eta_m(s_i, t) - 1) \nu_m(t) \right) \rho_i(t) - \eta(s_i, t) \rho_i(t). \quad (\text{C.9})$$

The equation set (C.9) calculates the posterior distribution of the state process  $X_t$  only, using the average responses of the sensory cells, with respect to the noise process  $W_t$ . This set of equations is presented as (3.1) in the main text.

## D Filtering a CFMP from SEPP Observations

Consider a CFMP  $X_t$  observed via a set of SEPPs  $N_t^{(1)}, \dots, N_t^{(M)}$ , with history dependent rate functions denoted by  $\lambda_1(t, X_t), \dots, \lambda_m(t, X_t)$ . The derivation in (Brémaud, 1981) (and also in the online appendix) for Poisson observations is based on the likelihood function (or sample-density function) of each of the Poisson processes. It can be proved (see section 6.2 in (Snyder & Miller, 1991)) that given the trajectory of  $X_t$ , the likelihood function for a realization of  $N_t^{(m)}$  is given by

$$f_{N^{(m)}} \left( N_{[0,t]}^{(m)} \right) = \left( \prod_{t_n \leq t} \lambda_m(t_n, X_{t_n}) \right) \exp \left( - \int_0^t \lambda_m(s, X_s) ds \right). \quad (\text{D.1})$$

where  $\{t_n\}_{n \in \mathbb{N}}$  are the event arrival times of the process  $N_t^{(m)}$ . This expression is similar in structure to the likelihood function of the Poisson process (see section 2.3 in (Snyder & Miller, 1991), and also VI.3 in (Brémaud, 1981)). Thus, it is possible to verify that all the steps in the Poisson derivation still hold, and therefore the filtering equation in this case is the one presented in (6.2).

## E Prediction

In this section we show how our framework can be easily adjusted to predict future states. Formally speaking, we define the prediction problem as calculating the posterior probabilities

$$p_i^\tau(t) = \Pr \left( X_{t+\tau} = s_i | N_{[0,t]}^{(1)}, \dots, N_{[0,t]}^{(M)} \right).$$

Denoting the non-normalized prediction probability vector by  $\boldsymbol{\rho}^\tau(t)$ , and recalling that  $X_t$  is a CFMP, it is easy to show that

$$\boldsymbol{\rho}^\tau(t) = (\mathbf{P}^{(\tau)})^\top \boldsymbol{\rho}(t), \quad (\text{E.1})$$

and thus, using (2.4) we find that

$$\dot{\boldsymbol{\rho}}^\tau(t) = (\mathbf{P}^{(\tau)})^\top \dot{\boldsymbol{\rho}}(t) = (\mathbf{P}^{(\tau)})^\top \left( \mathbf{Q}^\top + \sum_{m=1}^M (\Lambda_m - \mathbf{I}) \nu_m(t) - \Lambda \right) \boldsymbol{\rho}(t).$$

Assuming that  $\mathbf{P}^{(\tau)}$  is a non-singular matrix, according to (E.1) we can replace  $\boldsymbol{\rho}(t)$  with  $(\mathbf{P}^{(\tau)})^{-\top} \boldsymbol{\rho}^\tau(t)$ , where  $(\cdot)^{-\top}$  denotes the inverse of the transpose. This yields

$$\begin{aligned} \dot{\boldsymbol{\rho}}^\tau(t) &= (\mathbf{P}^{(\tau)})^\top \left( \mathbf{Q}^\top + \sum_{m=1}^M (\Lambda_m - \mathbf{I}) \nu_m(t) - \Lambda \right) (\mathbf{P}^{(\tau)})^{-\top} \boldsymbol{\rho}^\tau(t) \\ &= \mathbf{Q}^\top \boldsymbol{\rho}^\tau(t) + \sum_{m=1}^M (\Lambda_m^\tau - \mathbf{I}) \nu_m(t) \boldsymbol{\rho}^\tau(t) - \Lambda^\tau \boldsymbol{\rho}^\tau(t), \end{aligned}$$

where

$$\Lambda_m^\tau = (\mathbf{P}^{(\tau)})^\top \Lambda_m (\mathbf{P}^{(\tau)})^{-\top} \quad ; \quad \Lambda^\tau = (\mathbf{P}^{(\tau)})^\top \Lambda (\mathbf{P}^{(\tau)})^{-\top}$$

using the fact that  $(\mathbf{P}^{(\tau)})^\top \mathbf{Q} (\mathbf{P}^{(\tau)})^{-\top} = \mathbf{Q}$ . This equation is presented as (7.1) in the main text.

## F Computing the Log-Posterior Distribution

In this section we show how to compute the logarithm of the non-normalized probabilities denoted by  $r_i(t) = \log(\rho_i(t))$ . In order to find the differential equation computing  $r_i(t)$  we examine the original equation in two different cases.

### Between spike arrivals:

During the periods between spike arrivals,  $\rho_i(t)$  is a differentiable function. Therefore we can write

$$\dot{r}_i(t) = \frac{d}{dt} \log(\rho_i(t)) = \frac{\dot{\rho}_i(t)}{\rho_i(t)}. \quad (\text{F.1})$$

Using equation (2.3), during these periods,  $\rho_i(t)$  evolves according to

$$\dot{\rho}_i(t) = \sum_k q_{ki} \rho_k(t) - \lambda(s_i) \rho_i(t). \quad (\text{F.2})$$

Combining (F.1) and (F.2) yields

$$\dot{r}_i(t) = \sum_k q_{ki} \frac{\rho_k(t)}{\rho_i(t)} - \lambda(s_i) = \sum_{k=1}^N q_{ki} \frac{\rho_k(t)}{\rho_i(t)} - \lambda(s_i).$$

Recalling that  $\rho_i(t) = \exp(r_i(t))$  we get

$$\dot{r}_i(t) = \sum_{k=1}^N q_{ki} \exp(r_k(t) - r_i(t)) - \lambda(s_i). \quad (\text{F.3})$$

### At a spike arrival:

Recall, that when a spike arrives from the  $m$ -th sensory cell, the effect on  $\rho_i$  is (see (A.2))

$$\rho_i(t_n) = \lambda_m(s_i) \rho_i(t_n^-).$$

Therefore, the effect on  $r_i$  is

$$r_i(t_n) = \log(\rho_i(t_n)) = \log(\lambda_m(s_i)\rho_i(t_n^-)) = \log \lambda_m(s_i) + r_i(t_n^-),$$

implying that

$$r_i(t_n) - r_i(t_n^-) = \log \lambda_m(s_i). \quad (\text{F.4})$$

Combining (F.3),(F.4) leads to

$$\dot{r}_i(t) = \sum_{k=1}^N q_{ki} \exp(r_k(t) - r_i(t)) - \lambda(s_i) + \sum_{m=1}^M \log(\lambda_m(s_i)) \nu_m(t). \quad (\text{F.5})$$

This equation is presented as equation (7.2) in the main text.

## References

- Anderson, B., & Moore, J. (2005). *Optimal filtering*. Dover Books.
- Averbeck, B., Latham, P., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nat Rev Neurosci*, 7(5), 358–366.
- Barbieri, R., Frank, L., Nguyen, D., Quirk, M., Solo, V., Wilson, M., et al. (2004). Dynamic analyses of information encoding in neural ensembles. *Neural Comput*, 16(2), 277–307.
- Beck, J., & Pouget, A. (2007). Exact inferences in a neural implementation of a hidden markov model. *Neural Comput*, 19(5), 1344–1361.
- Berry, M., & Meister, M. (1998). Refractoriness and neural precision. *J Neurosci*, 18(6), 2200–2211.
- Bobrowski, O., Meir, R., Shoham, S., & Eldar, Y. (2007). A neural network implementing optimal state estimation based on dynamic spike train decoding. In D. Koller & Y. Singer (Eds.), *Neural information processing systems* (Vol. NIPS). MIT Press. (To appear)
- Boel, R., & Benes, V. (1980). Recursive nonlinear estimation of a diffusion acting as the rate of an observed Poisson process. *IEEE Tras Inf Theory*, 26(5), 561–575.
- Brémaud, P. (1981). *Point processes and queues: Martingale dynamics*. Springer, New York.
- Cavanaugh, J., Bair, W., & Movshon, J. A. (2002). Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *J Neurophysiol*, 88(5), 2530–2546.
- Deneve, S. (2008). Bayesian spiking neurons i: inference. *Neural Comput*, 20(1), 91–117.
- Deneve, S., Latham, P., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nat Neurosci*, 4(8), 826–831.
- Deneve, S., & Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *J Physiol Paris*, 98(1-3), 249–258.
- Douglas, R., & Martin, K. (2004). Neuronal circuits of the neocortex. *Annu Rev Neurosci*, 27, 419–51.

- Doya, K., Ishii, S., Pouget, A., & Rao, R. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT Press.
- Eden, U., Frank, L., Solo, V., & Brown, E. (2004). Dynamic analysis of neural encoding by point process adaptive filtering. *Neural Computation*, *16*, 971-998.
- Ernst, M., & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433.
- Fairhall, A., Lewen, G., Bialek, W., & Ruyter Van Steveninck, R. de. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*(6849), 787-92.
- Georgopoulos, A., Kalaska, J., Caminiti, R., & Massey, J. (1982, Nov). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J Neurosci*, *2*(11), 1527-1537.
- Gerstner, W., & Kistler, W. (2002). *Spiking neuron models*. Cambridge University Press.
- Ghazanfar, A., & Schroeder, C. (2006). Is neocortex essentially multisensory? *Trends Cogn Sci*, *10*(6), 278-285.
- Grimmett, G., & Stirzaker, D. (2001). *Probability and random processes* (Third ed.). Oxford University Press.
- Huys, Q., Zemel, R., Natarajan, R., & Dayan, P. (2007). Fast population coding. *Neural Comput*, *19*(2), 404-441.
- Jazwinsky, A. (1970). *Stochastic processes and filtering theory*. Academic Press.
- Knill, D., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*, *27*(12), 712-719.
- Kushner, H. (1967). Dynamical equations for optimal nonlinear filtering. *J. Differential Equations*, *3*, 179-190.
- Ma, W., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci*, *9*(11), 1432-1438.
- McAdams, C., & Reid, R. (2005). Attention modulates the responses of simple cells in monkey primary visual cortex. *J Neurosci*, *25*, 11023-11033.
- McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on the reliability of individual neurons in monkey visual cortex. *Neuron*, *23*(4), 765-773.
- Murphy, B. K., & Miller, K. D. (2003). Multiplicative gain changes are induced by excitation or inhibition alone. *J Neurosci*, *23*, 10040-10051.
- Pitkow, X., Sompolinsky, H., & Meister, M. (2007, Dec). A neural computation for visual acuity in the presence of eye movements. *PLoS Biol*, *5*(12), e331. Available from <http://dx.doi.org/10.1371/journal.pbio.0050331>
- Pouget, A., Dayan, P., & Zemel, R. (2003). Inference and computation with population codes. *Annu. Rev. Neurosci*, *26*, 381-410.
- Pouget, A., Deneve, S., & Duhamel, J. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nat Rev Neurosci*, *3*(9), 741-747.
- Rao, R. (2004). Bayesian computation in recurrent neural circuits. *Neural Comput*, *16*(1), 1-38.
- Rao, R. (2006). Neural models of Bayesian belief propagation. In K. Doya, S. Ishii, A. Pouget, & R. Rao (Eds.), *Bayesian brain* (chap. 11). MIT Press.
- Rivadulla, C., Sharma, J., & Sur, M. (2001). Specific roles of nmda and ampa recep-

- tors in direction-selective and spatial phase-selective responses in visual cortex. *J Neurosci*, *21*, 1710–1719.
- Salinas, E., & Thier, P. (2000). Gain modulation: a major computational principle of the central nervous system. *Neuron*, *27*, 15–21.
- Sanger, T. (1996). Probability density estimation for the interpretation of neural population codes. *J Neurophysiol*, *76*(4), 2790–2793.
- Segall, A., Davis, M., & Kailath, T. (1975). Nonlinear filtering with counting observations. *IEEE Tran. Information Theory*, *21*(2), 143–149.
- Sharpee, T., Sugihara, H., Kurgansky, A., Rebrik, S., Stryker, M., & Miller, K. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature*, *439*(7079), 936–942.
- Shoham, S., Paninski, L., Fellows, M., Hatsopoulos, N., Donoghue, J., & Norman, R. (2005). Statistical encoding model for a primary motor cortical brain-machine interface. *IEEE Trans Biomed Eng.*, *52*(7), 1312–22.
- Snyder, D. (1972). Filtering and detection for doubly stochastic Poisson processes. *IEEE Transactions on Information Theory*, *IT-18*, 91–102.
- Snyder, D., & Miller, M. (1991). *Random point processes in time and space* (Second ed.). New York: Springer.
- Sripati, A., & Johnson, K. (2006). Dynamic gain changes during attentional modulation. *Neural Comput*, *18*, 1847–1867.
- Stanford, T., Quessy, S., & Stein, B. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *J Neurosci*, *25*(28), 6499–6508.
- Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. MIT Press.
- Tsien, J. (2000). Linking Hebb’s coincidence-detection to memory formation. *Curr Opin Neurobiol*, *10*(2), 266–73.
- Ts’o, D. Y., Gilbert, C. D., & Wiesel, T. N. (1986). Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *J Neurosci*, *6*(4), 1160–1170.
- Twum-Danso, N., & Brockett, R. (2001). Trajectory estimation from place cell data. *Neural Netw*, *14*(6-7), 835–844.
- Wark, B., Lundstrom, B., & Fairhall, A. (2007). Sensory adaptation. *Curr Opin Neurobiol*, *17*(4), 423–429.
- Warland, D. K., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *J Neurophysiol*, *78*(5), 2336–2350. (n1079)
- Witten, I., & Knudsen, E. (2005, Nov). Why seeing is believing: merging auditory and visual worlds. *Neuron*, *48*(3), 489–496.
- Wonham, W. (1965). Some applications of stochastic differential equations to optimal nonlinear filtering. *J. SIAM Control*, *2*(3), 347–369.
- Zakai, M. (1969). On the optimal filtering of diffusion processes. *Z. Wahrscheinlichkeitstheorie verw Gebiete*, *11*, 230–243.
- Zemel, R., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Comput*, *10*(2), 403–430.
- Zemel, R., Huys, Q., Natarajan, R., & Dayan, P. (2005). Probabilistic computation in spiking populations. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (p. 1609–1616). Cambridge, MA: MIT Press.