LETTER Communicated by Alexandre Pouget

Fast Population Coding

Quentin J. M. Huys
qhuys@cantab.net
Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, U.K.

Richard S. Zemel
zemel@cs.toronto.edu
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3H5

Rama Natarajan
rama@cs.toronto.edu
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3H5

Peter Dayan
dayan@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, U.K.

Uncertainty coming from the noise in its neurons and the ill-posed nature of many tasks plagues neural computations. Maybe surprisingly, many studies show that the brain manipulates these forms of uncertainty in a probabilistically consistent and normative manner, and there is now a rich theoretical literature on the capabilities of populations of neurons to implement computations in the face of uncertainty. However, one major facet of uncertainty has received comparatively little attention: time. In a dynamic, rapidly changing world, data are only temporarily relevant. Here, we analyze the computational consequences of encoding stimulus trajectories in populations of neurons. For the most obvious, simple, instantaneous encoder, the correlations induced by natural, smooth stimuli engender a decoder that requires access to information that is nonlocal both in time and across neurons. This formally amounts to a ruinous representation. We show that there is an alternative encoder that is computationally and representationally powerful in which each spike contributes independent information; it is independently decodable, in other words. We suggest this as an appropriate foundation for understanding time-varying population codes. Furthermore, we show how adaptation to

 © 2007 Massachusetts Institute of Technology

temporal stimulus statistics emerges directly from the demands of simple decoding.

1 Introduction

From the earliest neurophysiological investigations in the cortex, it became apparent that sensory and motor information is represented in the joint activity of large populations of neurons (Barlow, 1953; Georgopoulos, Schwartz, & Kettner, 1983). There are by now substantial ideas and data about how these representations are formed (Rao, Olshausen, & Lewicki 2002), how information can be decoded from recordings of this activity (Paradiso, 1988; Snippe & Koenderinck, 1992; Seung & Sompolinsky, 1993), and how various sorts of computations, including uncertainty-sensitive, Bayesian optimal statistical processing can be performed through the medium of feedforward and recurrent connections among the populations (Pouget, Zhang, Deneve, & Latham, 1998; Deneve, Latham, & Pouget, 2001). Critical issues that have emerged from these analyses are the forms of correlations between neurons in the populations, whether these correlations are significant for decoding and computation, and what sorts of prior information are relevant to computations and can be incorporated by such networks.

However, many theoretical investigations into population coding have so far somewhat neglected a major dimension of coding: time. This is despite the beautiful and influential analyzes of circumstances in which individual spikes contribute importantly to the representation of rapidly varying stimuli (Bialek, Rieke, de Ruyter van Steveninck & Warland, 1991; Reinagel & Reid, 2000; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997; Johansson & Birznieks, 2004) and the importance accorded to fast-timescale spiking by some practical investigations into population coding (Wilson & McNaughton, 1993; Schwartz, 1994; Zhang, Ginzburg, McNaughton, & Sejnowski, 1998; Brown, Frank, Tang, Quirk, & Wilson, 1998). The assumption is often made that encoded objects do not vary quickly with time and that therefore spike counts in the population suffice. Even some approaches that consider fast decoding (Brunel & Nadal, 1998; Van Rullen & Thorpe, 2001) treat stimuli as being discrete and separate rather than as evolving along whole trajectories.

In this letter, we study the generic computational consequences of population coding in time. We analyze decoding in time as a proxy for computation in time as it is the most comprehensive computation that can be performed (accessing all information present). Decoding therefore constitutes a canonical test (Brown et al., 1998; Zhang et al., 1998). We consider a regime in which stimuli are not static and create sparse trains of spikes. Decoding trajectory information from these population spike trains is thoroughly ill posed, and prior information about what trajectories are likely

comes to play a critical role. We show that optimal decoding with ecological priors formally couples together the spikes, making trajectory inference computationally very hard. We thus consider the prospects for neural populations to recode the information about the trajectory into new sets of spikes that do support simple computations. Phenomena reminiscent of adaptation emerge as a by-product of the maintenance of a computationally advantageous code.

We analyze the extension of one of the simplest ideas about population codes for static stimuli (Snippe & Koenderinck, 1992) to the case of trajectories. This links a neurally plausible population encoding model with a naturally realistic gaussian process prior. Unlike some previous work on decoding in time (Brown et al., 1998; Zhang et al., 1998; Smith & Brown, 2003), we do not confine ourselves to recursively specifiable priors and can therefore treat smoother cases. It is these smooth priors that render decoding, and likely other computations, hard and inspire an energy-based (product of experts) recoding (Hinton, 1999; Zemel, Huys, Natarajan, & Dayan, 2005), which makes for readier probabilistic computation.

Section 2 starts with a simple encoding model. It introduces the need for priors, their shape, and analytical results for decoding in time. Section 3 shows how priors determine the form in which information is available to downstream neurons. We show that the decoder corresponding to the simple encoder can be extraordinarily complex, meaning that the encoded information is not readily available to downstream neurons. Finally, section 4 proposes a representation that has comparable power but for which decoding requires vastly less downstream computation.

2 A Gaussian Process Prior Approach

As a motivating example, consider tennis. The player returning a serve has to predict the position of the ball based on data acquired in fractions of seconds. Experts compensate for the extraordinarily sparse stimulus information with a very rich temporal prior over ball trajectories and thus make predictions that are accurate enough to guarantee many a winning return.

Figure 1 illustrates the setting of this article more formally. It shows an array of neurons with partially overlapping tuning functions that emit spikes in response to a stimulus that varies in time. These could be V1 neurons responding to a target (the tennis ball) as it moves through their receptive fields, or hippocampal neurons with place fields firing as a rat explores an environment. The task is to decode the spikes in time, that is, recover the trajectory of the stimulus (the ball's position, say) based on the spikes, a knowledge of the neuronal tuning functions (cf. Brown et al., 1998; Zhang et al., 1998, for hippocampal examples), and some knowledge about the temporal characteriztics of the stimulus (the prior). In Figure 1, the ordinate represents the stimulus space (here one-dimensional for illustrative

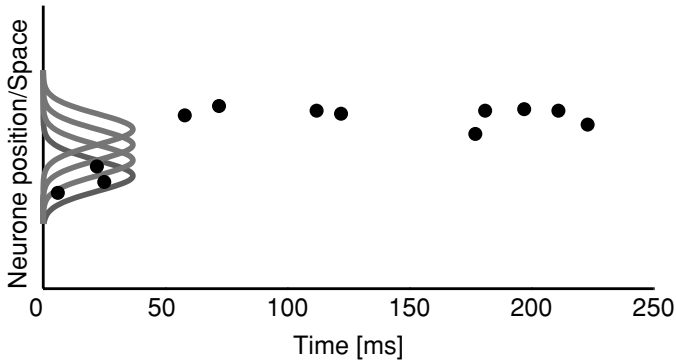

Figure 1: The problem: Reconstructing the stimulus as a function of time, given the spikes emitted by a population of neurons. When a neuron with preferred stimulus s_i emits a spike at time t, a black dot is plotted at (t, s_i). A few example tuning functions are shown in gray. The ordinate represents stimulus space, with each neuron being positioned according to its preferred stimulus s_i. The decoding problem is related to fitting a line through these points, which is achievable only if there is prior information about the line to be fitted (e.g., the order of a polynomial fit or the smoothness).

purposes) and the abscissa, time. Neuron i has preferred stimulus s_i. If it emits a spike ξ_t^i at time t, a dot is drawn at position (t, s_i). The dots in Figure 1 thus represent the spiking activity of the entire population of neurons over time. Our aim is to find, for each observation time T, a distribution over likely stimulus values s_T given all the spikes previous to T. This is related to fitting a line representing the trajectory of the stimulus through the points. It is a thoroughly ill-posed problem, for instance, because we are not given any information about the stimulus at all between the spikes.

To solve this ill-posed problem, we have to bring in additional knowledge in the form of a prior distribution about likely stimulus trajectories. The prior distribution specifies the temporal characteriztics of the trajectories (e.g., how smooth they are) and also whether they live within some constrained part of the stimulus space. Subjects are assumed to possess such prior information ahead of time—for instance, from previous exposures to trajectories (a good tennis player will have seen many serves).

To gain analytical insight into the structure of decoding in this temporally rich case, we consider a very simple spiking model $p(\xi_t^i|s_t)$ (c.f., Snippe & Koenderinck, 1992, for the static case), augmented with a simple prior over stimulus trajectories $p(\mathbf{s})$. We thereafter follow standard approaches (Zhang et al., 1998; Brown et al., 1998) by performing causal decoding and thus recovering $p(s_T|\boldsymbol{\xi})$ over the current stimulus s_T at time T given all the J spikes $\boldsymbol{\xi} \equiv \{\xi_{t_j}^i\}_{j=1}^J$ at times $0 < \{t_j\}_{j=1}^J < T$ in the observation period

([0, T)), emitted by the entire population. Here, $i = 1, \ldots, N$ designates the neuron that emitted the spike.

To state the problem in mathematical terms, we can write (at least for the case that there is no spike at time T itself)

$$p(s_T|\boldsymbol{\xi}) \propto p(s_T)p(\boldsymbol{\xi}|s_T) \tag{2.1}$$

$$= p(s_T)\int d\mathbf{s}_{\overline{T}}\; p(\boldsymbol{\xi}|\mathbf{s}_{\overline{T}})p(\mathbf{s}_{\overline{T}}|s_T), \tag{2.2}$$

where, being slightly notationally sloppy, we are integrating over stimulus trajectories $\mathbf{s}_{\overline{T}}$ up to, but not including, time T, but restricted to just those trajectories that end at s_T.

Equation 2.2 lays bare the two parts of the definition of the problem. One is the likelihood $p(\boldsymbol{\xi}|\mathbf{s}_{\overline{T}})$ of the spikes given the trajectory. This will be assumed to arise from a Poisson-gaussian spiking model. The other is the prior

$$p(s_T)p(\mathbf{s}_{\overline{T}}|s_T) = p(\mathbf{s}) \tag{2.3}$$

over the trajectories. This will be assumed to be a gaussian process.

2.1 Poisson-Gaussian Spiking Model. We first define the spiking model. Let $\phi_i(s)$ be the tuning function of neuron i and assume independent, inhomogeneous, and instantaneous Poisson neurons (Snippe & Koenderinck, 1992; Brown et al., 1998; Barbieri et al., 2004). Let j be an index running over all the spikes in the population, with $i(j)$ reporting the index of the neuron that spiked at time t_j. Then, from the basic definition of an inhomogeneous Poisson process, the likelihood of a particular population spike train $\boldsymbol{\xi}$ given the stimulus trajectory $\mathbf{s}_{\overline{T}}$ can be written as

$$p(\boldsymbol{\xi}|\mathbf{s}_{\overline{T}}) = \prod_j \phi_{i(j)}(s_{t_j}) \exp\left(-\sum_i \int_0^{\overline{T}} dt\phi_i(s_t)\right) \tag{2.4}$$

$$\propto \prod_j \phi_{i(j)}(s_{t_j}), \tag{2.5}$$

assuming that the trajectories are such that we can swap the order of the sum and the integral in the exp(·), that tuning functions are sufficiently dense that the sum spiking rate is constant independent of the location of the stimulus s_t, and that no two neurons ever fire together.

Finally, we assume squared-exponential (gaussian) tuning functions,

$$\phi_i(s_{t_j}) = \phi_{\max} \exp\left(-\frac{\left(s_{t_j} - s_i\right)^2}{2\sigma^2}\right),$$

where $\phi_{\max}$ is the maximal firing rate of a neuron and s_i the ith neuron's preferred stimulus. Combining this with our previous assumptions (see equation 2.5) and completing the square implies that

$$p(\boldsymbol{\xi}|\mathbf{s}_{\overline{T}}) \propto \phi_{\max} \exp\left(-\frac{(\mathbf{s}_\xi - \boldsymbol{\theta})^T(\mathbf{s}_\xi - \boldsymbol{\theta})}{2\sigma^2}\right), \tag{2.6}$$

where the spikes from the entire population have been ordered in time; the jth component of both $\mathbf{s}_\xi$ and $\boldsymbol{\theta}$ corresponds to the jth spike and is, respectively, the stimulus at that spike's time t_j and the preferred stimulus s_i of the neuron that produced it. Note that time is continuous here.

2.2 Gaussian Process Prior. The prior $p(\mathbf{s})$ defines a distribution over stimulus trajectories that are continuous in time. However, $p(\boldsymbol{\xi}|\mathbf{s}_{\overline{T}})$ in equation 2.6 depends on only the times t_j at which neurons in the population spike. Thus, in the integral in equation 2.2, we can formally marginalize or integrate out all the nonspiking times, making the key quantity to be defined by the prior to be $p(\mathbf{s}_\xi, s_T)$.

For a gaussian process (GP), this quantity is a multivariate gaussian, defined by its $(J+1)$-dimensional mean vector $\mathbf{m}$ and covariance matrix $\mathcal{C}$, which can in general depend on the times t_j. We write the distribution as

$$p(\mathbf{s}_\xi, s_T) \sim \mathcal{N}(\mathbf{m}, \mathcal{C}) \qquad \mathcal{C}_{t_j t_{j'}} = c \exp\left(-\alpha \|t_j - t_{j'}\|^\zeta\right). \tag{2.7}$$

The parameter $\zeta \geq 0$ dictates the smoothness and the correlation structure of the process. If $\zeta = 0$, then the stimulus is assumed to be constant (we sometimes call this the static case). Setting $\zeta = 1$ corresponds to assuming that the stimulus evolves as an Ornstein-Uhlenbeck (OU) or first-order autoregressive process. This is the generative model underlying Kalman filters (Twum-Danso & Brockett, 2001) and generates an autocorrelation with the Fourier spectrum $\sim 1/\omega^2$ often observed experimentally (Atick, 1992; Dong & Atick, 1995; Wang, Liu, Sanchez-Vives, & McCormick, 2003). This can be generalized to nth-order autoregressive processes. Setting $\zeta = 2$ leads to the opposite end of the spectrum, with smooth trajectories that are non-Markovian. The parameter α dictates the temporal extent of the correlations and c their overall size (c also parameterizes the scale of the overall process). Example trajectories drawn from these priors for $\zeta = \{1, 2\}$ are shown in Figure 2. For most of the letter, we will let $\mathbf{m} = \mathbf{0}$. Assuming

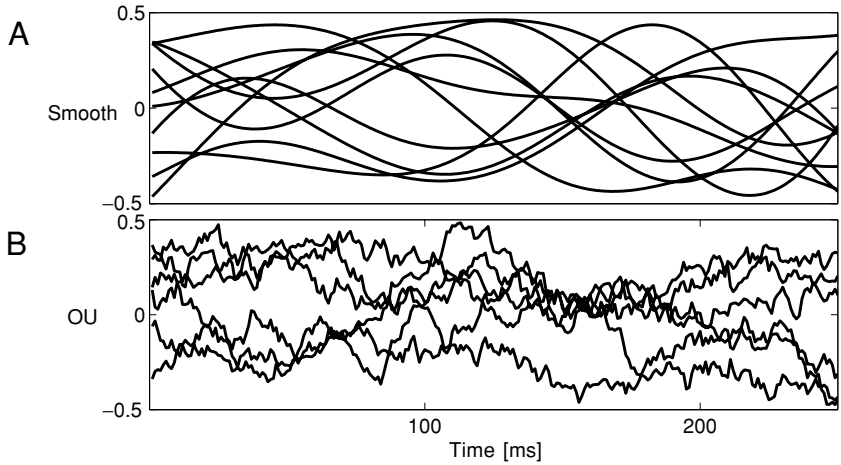

Figure 2: Example trajectories drawn from the prior distribution in equation 2.7. (A) Examples for the smooth covariance matrix with $\zeta = 2$. (B) The OU covariance matrix, $\zeta = 1$.

a GP prior with a particular covariance matrix is exactly equivalent to regularizing the autocorrelation of the trajectory.

2.3 Posterior. Making these assumptions, we can write down the posterior distribution $p(s_T|\boldsymbol{\xi})$ analytically by solving equation 2.2. It is a simple gaussian distribution with mean $\mu(T)$ and variance $\nu^2(T)$ given in terms of tuning function widths σ, the vector $\boldsymbol{\theta}$, and the covariance matrix $\mathcal{C}$.

All three terms in equation 2.2 are now defined. The conditional distribution $p(\mathbf{s}_\xi|s_T)$ is given in terms of the partitioned covariance matrix $\mathcal{C}$,

$$p(\mathbf{s}_\xi|s_T) = \mathcal{N}_{\mathbf{s}_\xi}\left(\mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}s_T, \left(\mathcal{C}_{\xi\xi} - \mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi}\right)\right),$$

where $\mathcal{C}_{\xi\xi}$ is the covariance matrix of the stimulus at all the spike times, $\mathcal{C}_{T\xi}$ and $\mathcal{C}_{\xi T}$ are vectors with the cross-covariances between the spike times and the observation time T, and $\mathcal{C}_{TT}$ is the marginal (static) stimulus prior at the observation time (constant for the stationary processes considered here). The corresponding partitioning of the matrix $\mathcal{C}$ is

$$\mathcal{C} = \left(\begin{array}{c|c} \mathcal{C}_{\xi\xi} & \mathcal{C}_{\xi T} \\ \hline \mathcal{C}_{T\xi} & \mathcal{C}_{TT} \end{array}\right). \tag{2.8}$$

The remaining two terms in equation 2.2 are given by $p(s_T) = \mathcal{N}_{s_T}(0, \mathcal{C}_{TT})$ and equation 2.6. As the integral in equation 2.2 is a convolution of two gaussians, the variances add, and the integral evaluates to

$$p(\boldsymbol{\xi}|s_T) = \mathcal{N}_{\boldsymbol{\theta}}\left(\mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}s_T, \left(\mathcal{C}_{\xi\xi} - \mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi}\right) + \mathbf{I}\sigma^2\right).$$

Finally, taking a product with $p(s_T)$, renormalizing, and applying the matrix inversion lemmas (see appendix A), we get

$$\mu(T) = \mathbf{k}(\boldsymbol{\xi}, T) \cdot \boldsymbol{\theta}(T) \tag{2.9}$$

$$\nu^2(T) = \mathcal{C}_{TT} - \mathbf{k}(\boldsymbol{\xi}, T) \cdot \mathcal{C}_{\xi T} \tag{2.10}$$

$$\mathbf{k}(\boldsymbol{\xi}, T) = \mathcal{C}_{T\xi}(\mathcal{C}_{\xi\xi} + \mathbf{I}\sigma^2)^{-1}. \tag{2.11}$$

The mean $\mu(T)$ of the gaussian posterior is thus a weighted sum of the preferred stimulus of those neurons that emitted particular spikes. The weights are given by what we term the temporal kernel $\mathbf{k}(\boldsymbol{\xi}, T)$. As we will see, the weight given to each spike will depend strongly on the time at which it occurred. A spike that occurred in the distant past will be given small weight. The posterior variance depends on only $\mathcal{C}$ and σ^2. Remember that $\mathcal{C}$ depends on only the times of spikes, not the identities of the neurons that fired them. The posterior variance ν^2, similar to a Kalman filter, depends on only *when* data are observed, not *what* data. This depends on the squared exponential nature of the tuning functions ϕ, and other tuning functions (e.g., with nonzero baselines) may not lead to this quality. However, it will not affect the conclusions reached below. This posterior distribution $p(s_T|\boldsymbol{\xi})$ is well known in the GP literature as the predictive distribution (MacKay, 2003).

2.4 Structure of the Code. The operations needed to evaluate the posterior $p(s_T|\boldsymbol{\xi})$ give us insight into the structure of the code and will be analyzed in section 3 for various priors. If the posterior is a function of combinations of spikes, postsynaptic neurons have to have simultaneous access to all those spikes. This point will be critical in temporal codes, as the spikes to which access is required are spread out in time. Only if spikes are interpretable independently can they be forgotten once they have been used for inference. All information the spikes contribute to some future time $T' > T$ is then contained within $p(s_T|\boldsymbol{\xi})$. If the posterior depends on combinations of spikes (as will be the case for ecological, smooth priors), information that can be extracted from a spike about times $T' > T$ is not entirely contained within $p(s_T|\boldsymbol{\xi})$. As a result, past spikes have to be stored and the posterior recomputed using them—an operation that is nonlocal in time. We will show that under ecological priors, the posterior depends on spike combinations and is thus complex. Decoding for the simple encoder (the spiking model) is thus hard. In section 4, we will illustrate the type

of computations ("recoding") a network has to perform to access all the information. This will be equivalent to finding a new, complex encoder in time for which decoding is simple.

3 Effect of the Prior

The effect of the prior manifests itself very clearly in the temporal kernels $\mathbf{k}(\boldsymbol{\xi}, T)$ from equation 2.11 and the independence structure of the code. We show this by analyzing a representative set of priors in terms of both the behavior of the temporal kernels and the structure of the code, including priors that generate constant, varyingly rough and entirely smooth trajectories. (MATLAB example code can be downloaded from http://www.gatsby.ucl.ac.uk/~qhuys/code.html.)

3.1 Constant stimulus prior $\zeta = 0$. We first show that our treatment of the time-varying case is an exact generalization of the case in which the stimulus is fixed (does not change relative to the mean $\mathbf{m}$), by rederiving classical results for static stimuli. Snippe and Koenderinck (1992) have shown that the posterior mean and variance (under a flat prior) is given by a weighted spike count,

$$\mu(T) = \frac{\sum_i n_i(T) s_i}{J(T)} \qquad \nu^2(T) = \frac{\sigma^2}{J(T)} \tag{3.1}$$

where $n_i(T) = \int_0^T dt\, \xi_t^i$ is the ith neuron's spike count and $J(T) = \sum_i n_i(T)$ is the total population spike count at time T.

If we let $\zeta = 0$, the matrix $\mathcal{C}_{\xi\xi} = c\mathbf{n}\mathbf{n}^T$ where $\mathbf{n}$ is a $J(T) \times 1$ vector of ones. Equations 2.9 to 2.11 can then be solved analytically:

$$\begin{aligned}
\left((\mathcal{C}_{\xi\xi} + \mathbf{I}\sigma^2)^{-1}\right)_{ij} &= \frac{(\sigma^2 + cJ(T))\delta_{ij} - c}{\sigma^2(\sigma^2 + cJ(T))} \\
\mathbf{k}(\boldsymbol{\xi}, T) &= \frac{c}{\sigma^2 + cJ(T)}\mathbf{n} \\
\mu(T) &= \frac{c\sum_i n_i(T) s_i}{\sigma^2 + cJ(T)} \\
\nu^2(T) &= \frac{c\sigma^2}{\sigma^2 + cJ(T)},
\end{aligned}$$

which is exactly analogous to equation 3.1 with an informative prior. The temporal kernel $\mathbf{k}(\boldsymbol{\xi}, T)$ does not decay but is flat, with a magnitude proportional to $1/J(T)$. The contribution of each neuron to the mean $\mu(T)$ is given by its spike count $n_i(T)$. Each spike is given the same weight,

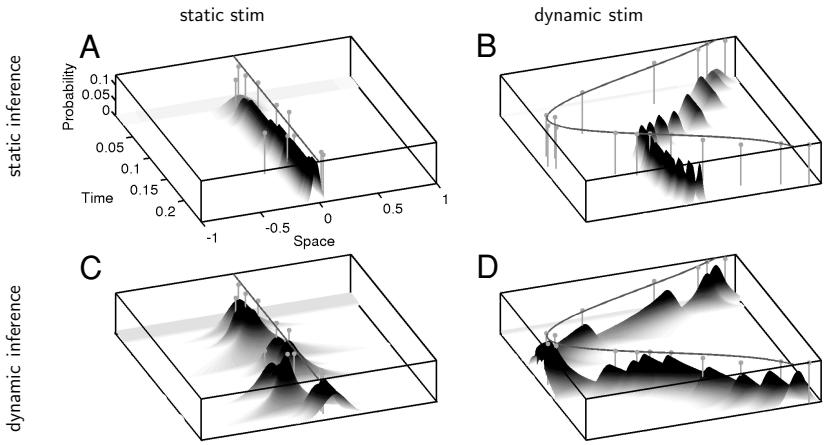

Figure 3: Comparison of static and dynamic inference. Throughout, the posterior density $p(s_T|\boldsymbol{\xi})$ is indicated by gray shading, the spikes are vertical (gray) lines with dots, and the true stimulus is the line at the top of each plot. (A) Static stimulus, constant temporal kernel. (B) Moving stimulus, constant temporal kernel. (C) Static stimulus, decaying temporal kernel. (D) Moving stimulus, decaying temporal kernel. *A* and *D* show that only a match between true stimulus statistics and prior allows the posterior to capture the stimulus well.

which is a sensible approach only if spikes are eternally informative about the stimulus. This is true only if the covariance matrix is flat, which itself implies that the only time-varying component of the stimulus is in the mean $\mathbf{m}$ and not the covariance C. If the stimulus is a varying function of time $s(t)$, spikes at time t' are informative only about the stimulus at times t close to t' and the influence of each spike on the posterior should fade away with time. This is illustrated in Figure 3. Figure 3A shows the present static case, where the stimulus does indeed not move; over time, the posterior $p(s_T|\boldsymbol{\xi})$ sharpens up around the true value. However, if the stimulus does move, the posterior ends up at the wrong value (see Figure 3B).

If the temporal kernel $\mathbf{k}(\boldsymbol{\xi}, T)$ decays, this amounts to downweighting spikes observed in the more distant past. In the following, we analyze the behavior of $p(s_T|\boldsymbol{\xi})$ and the optimal temporal kernel $\mathbf{k}(\boldsymbol{\xi}, T)$ for various stimulus autocorrelation functions. Figure 3C shows that a decaying kernel leads to a posterior that widens inbetween spikes. This is incorrect if the stimulus is static, but Figure 3D shows how such a decaying temporal kernel would, in contrast to Figure 3B, allow $p(s_T|\boldsymbol{\xi})$ to track the moving stimulus correctly.

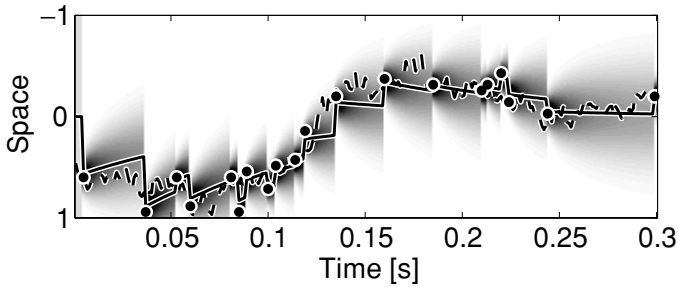

Figure 4: Posterior distribution $p(s_T|\boldsymbol{\xi})$ for OU prior. Same representation as in Figure 1. The dashed line shows the actual stimulus trajectory used to generate the spikes, the dots are the spikes, the posterior distribution is in gray scale, and the solid line shows the posterior mean. Between spikes, the posterior mean decays exponentially back toward the mean $\mathbf{m}$ (here 0), and the variance approaches the static prior variance $\mathcal{C}_{TT}$.

3.2 Nonsmooth (Ornstein-Uhlenbeck) Prior $\zeta = 1$. Setting $\zeta = 1$ in the definition of the prior (see equation 2.7) corresponds to assuming that the stimulus evolves as a random walk with drift to zero (an OU process):

$$ds = -(1 - e^{-\alpha})s(t)dt + \sqrt{c(1 - e^{-2\alpha})}\sqrt{dt}\, dN(t), \tag{3.2}$$

with gaussian noise $N(t) \sim \mathcal{N}(0, 1)$ and parameters as in equation 2.7. The O-U process is the underlying generative process assumed by standard Kalman filters. The simplicity of Kalman-filter-like formulations explains some of its wide applicability and success (Brown et al., 1998; Barbieri, et al., 2004). However, as indicated visually by the example trajectories in Figure 2, the rough trajectories this prior favors are not a good model of smooth biological movements (see also section 5).

Figure 4 shows a sampled stimulus trajectory, sample spikes generated from it, and the posterior distribution $p(s_T|\boldsymbol{\xi})$. The mean of the posterior does a good job of tracking the true underlying stimulus trajectory and is never more than two standard deviations away from it. Between spikes, the mean simply moves back to zero (albeit rather slowly given the parameters associated with the Figure shown).

Figure 5A displays example temporal kernels $\mathbf{k}(\boldsymbol{\xi}, T)$ for inference in this process. They are very close to exponentials (note the logarithmic ordinate). This makes intuitive sense as an OU process is a first-order Markov process (it can be rewritten as a first-order difference equation). In fact, assuming the spikes arrive regularly (replacing each of the interspike intervals (ISI) by their average value $\Delta = \frac{1}{J}\sum_j (t_j - t_{j-1}) \propto \frac{1}{\phi_{\max}}$) allows us to write the

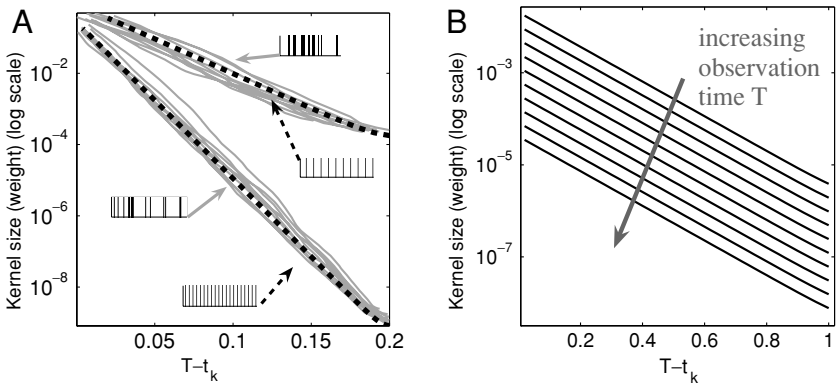

Figure 5: OU temporal kernels for $\zeta = 1$. (A) Example of temporal kernels. The top traces are for lower and the bottom for higher average firing rates. The gray traces show temporal kernels for Poisson spike trains. The components of the vector $\mathbf{k}(\boldsymbol{\xi}, T)$ are plotted against the corresponding spike time. The dashed black traces show temporal kernels for regular spike arrivals (metronomic temporal kernels). The true (gray) temporal kernels are relatively tightly bunched around the metronomic temporal kernel. The firing rate affects the slope of the kernel, but not its overall scale of the kernel. (B) The effect of the time since the last spike on the temporal kernel is an overall multiplicative scaling. There is no effect on the slope.

jth component of $\mathbf{k}(\boldsymbol{\xi}, T)$ as

$$k_j \approx d_1 \lambda_1^{j-1},$$

where d_1 and λ_1 are constants defined in appendix B. For such metronomic spiking, $\mathbf{k}(\boldsymbol{\xi}, T)$ is thus really simply a decaying exponential. Somewhat similar expressions can be obtained for the original case of Poisson-distributed ISIs (see appendix B). Figure 5A shows that the metronomic approximation provides a generally good fit, capturing especially the slope of the true temporal kernels, which depends mostly on the correlation length α and the maximal (or average) firing rate $\phi_{\max}$. The remaining quality of the fit is influenced most strongly by the match between Δ and the time since the last spike $T - t_J$ (which takes its effects through $\mathcal{C}_{T\xi}$ in equation 2.8 and 2.9–2.11). This determines the overall scale of the temporal kernel.

The factors influencing the slope of the temporal kernel and its height do not interact greatly; that is, $T - t_J$ does not affect the slope (shape) of the temporal kernel, only its magnitude, as shown in Figure 5B (metronomic temporal kernels are used for clarity, but the argument applies equally to the exact kernel). Conversely, Δ affects mostly the slope. Replacing the true

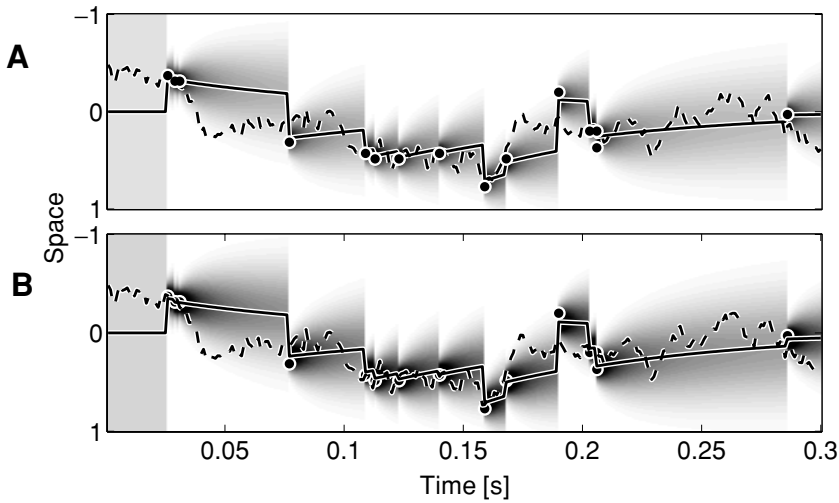

Figure 6: Comparison between exact and metronomic kernels. Same representation as in Figure 4. (A) Exact posterior $p(s_T|\boldsymbol{\xi})$. (B) Approximate posterior derived by replacing all ISIs by Δ but keeping $T - t_J$. This corresponds to approximating the true kernels with the metronomic kernels in Figure 5. The approximation is very close.

temporal kernels by metronomic temporal kernels, that is, replacing all ISIs by Δ but keeping the time since the last spike $T - t_J$, does not greatly degrade $p(s_T|\boldsymbol{\xi})$ (cf. Figures 6A and 6B).

The dependence in Figure 5B can be understood by writing out the integrand of equation 2.2 in detail for the OU prior. This factorizes over potentials involving duplets of spikes because, as we show in appendix B, $\mathcal{C}^{-1}$ is tridiagonal, implying that the elements of $\mathcal{C}^{-1}$ involve only two successive spikes:

$$\begin{aligned} p(\mathbf{s}_\xi, s_T) &\propto \exp\left(-\frac{1}{2}\left[\mathbf{s}_\xi s_T\right]^T \mathcal{C}^{-1}\begin{bmatrix}\mathbf{s}_\xi \\ s_T\end{bmatrix}\right) \\ &= \exp\left(-\frac{1}{2}\left(\sum_{j=1}^{J+1} s_{t_j}^2 \mathcal{C}^{-1}_{t_j t_j} + \sum_{j=1}^{J} s_{t_j}\mathcal{C}^{-1}_{t_j,t_{j+1}} s_{t_{j+1}}\right)\right) \\ p(\mathbf{s}_\xi, s_T) &= \psi(s_T)\prod_{j=1}^{J}\psi(s_{t_j}, s_{t_{j+1}}), \end{aligned} \tag{3.3}$$

where t_J stands for the time of the last spike, t_{J-1} the time of the penultimate one, and so on, and the observation time $T = t_{J+1}$. Note that the last equality

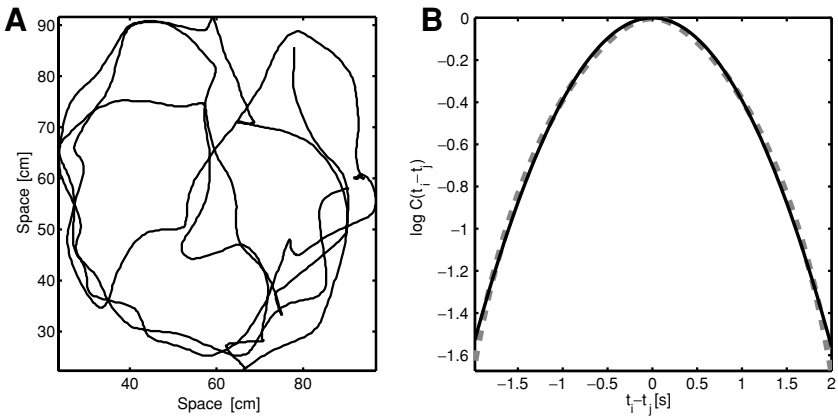

Figure 7: Natural trajectories are smooth. (A) Position of a rat freely exploring a square environment. (B) Covariance function of the position along the ordinate (gray, dashed line) and a quadratic approximation (black, solid line). Note the logarithmic ordinate. The smoothing applied to eliminate artifacts was of a timescale short enough not to interfere with the overall shape of the covariance function.

implies that the determinant also factors over spike pairs. This means that the integrations over each spike in the main equation 2.2 can be written in a recursive form akin to that used in message-passing algorithms (MacKay, 2003) and the exact Kalman filter.

3.3 Smooth Prior $\zeta = 2$. Setting $\zeta = 2$ in the definition of the prior (see equation 2.7) corresponds to assuming that the stimulus evolves as a non-Markov random walk. Trajectories with this autocovariance function are smooth (Figure 2A shows some sample trajectories generated from the prior) and infinitely differentiable. The smoothness makes it a more ecologically relevant prior for Bayesian decoding from movement-related trajectories than nonsmooth priors since natural objects (and limbs) move along smooth trajectories rather than jumping. As an example, Figure 7A shows trajectories of a rat exploring a square environment (data kindly provided by Lever, Wills, Cacucci, Burgess, & O'Keefe, 2002). Not only are these natural trajectories smooth, but Figure 7B also shows that a squared exponential covariance function closely approximates the real covariance function.[1]

[1] Only the center of the covariance function is shown here. Due to the small size of the environment, the rat runs back and forth the entire available length, and there are oscillating flanks to the covariance function for delays larger than those shown.

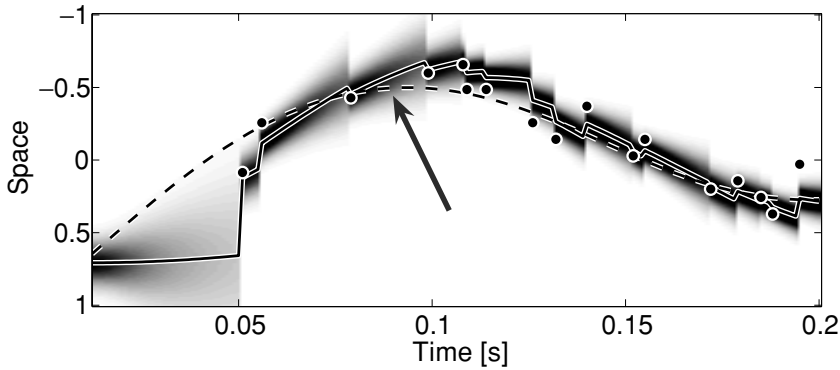

Figure 8: Posterior distribution $p(s_T|\boldsymbol{\xi})$ for the smooth prior. Same representation as in Figure 4. The arrow highlights where the smooth prior uses spike combinations to constrain higher-order statistics of the process, such as velocity, acceleration, and jerk. While the smooth prior correctly predicts that the stimulus will continue away from the mean before returning back, the OU process can predict a decay only back to the mean (see Figure 9). The first spike on the left is the very first spike observed. As the spike history becomes more extensive, the posterior distribution is seen to sharpen up and follow the stimulus accurately.

Figure 8 is the equivalent of Figure 4 for the smooth case and shows the posterior $p(s_T|\boldsymbol{\xi})$. The main dynamical difference between inference in this smooth case and inference in the OU case is indicated by the arrow in the Figure. While the OU process simply decays back to the mean (here, zero for simplicity), the dynamics of the smooth posterior mean are much richer. In the absence of spikes, the mean continues in its current direction for a while before reversing back. As can be seen, this gives a better fit to the underlying stimulus trajectory (the black dashed line) than would otherwise have been achieved. It arises directly from the fact that the correlations extend essentially beyond the last spike (and into the entire past). For comparison, Figure 9 shows the posterior when the wrong prior is used. The stimulus was generated from the smooth prior, but the OU prior was used to infer the posterior. The arrow indicates where the infelicity of the inaccurate posterior is most apparent, falling back to zero instead of predicting that the stimulus will continue to move farther away from zero. In terms of difference equations, the larger extent of correlations intuitively means that the higher-order derivatives of the process are also "constrained" by the covariance $\mathcal{C}$.

The simple exponential temporal kernels observed in the OU process cannot give rise to the reversals observed in the smooth process. Figure 10A shows the temporal kernels for the smooth process, which have a distinctively different flavor from the OU temporal kernels (shown in Figure 5),

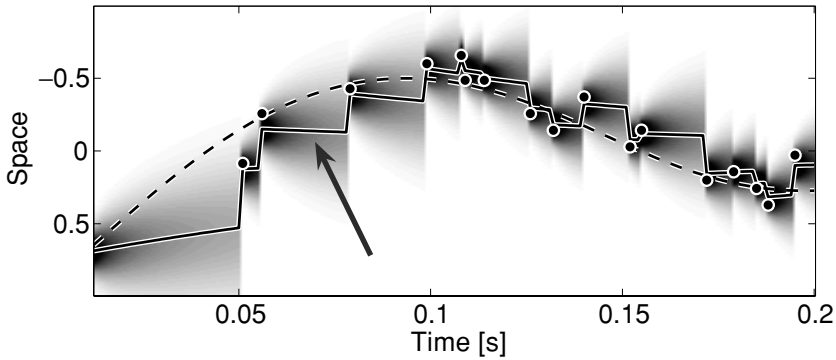

Figure 9: Posterior distribution $p(s_T|\boldsymbol{\xi})$ for smooth stimulus but wrongly assuming an OU prior. The posterior is consistently wider than it should be (cf. Figure 8). The arrow points out where the prediction is qualitatively wrong: the OU prior allows for decay back only to zero, unlike the smooth prior. Note also that the beneficial effect of a larger spike history observed in Figure 8 is absent here.

[thb]

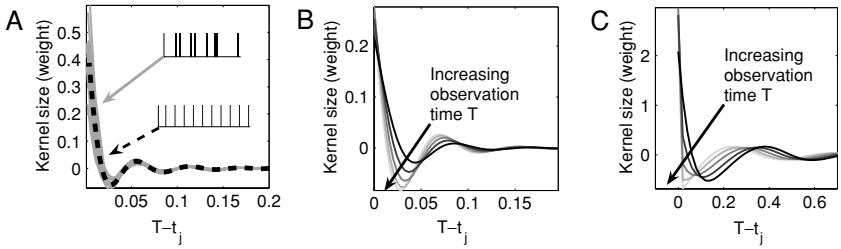

Figure 10: Temporal kernels for the smooth prior. (A) Exact (gray solid) and metronomic (black dashed) temporal kernels for the smooth prior with $\zeta = 2$. The metronomic kernels again provide a close fit. (B) The metronomic temporal kernels change in a complex manner as the observation time T is moved away from the time of the last spike. Unlike in the OU case, this is not just a recursively implementable multiplication. (C) The same qualitative behavior arises for kernels derived from the empirical covariance function of the rat trajectories.

including oscillating terms multiplying the exponential decay. Most important, the oscillating terms allow the weight assigned to a spike to dip below zero; that is, a spike initially signifies proximity of the stimulus to the neuron's preferred stimulus but later swaps over, signaling that the stimulus is not there anymore. This feature of the temporal kernels gives rise to the reversals seen in the posterior mean.

As in the OU case, the metronomic temporal kernel based on equal ISIs gives a good description of the temporal kernel mostly for spikes in the more distant past. Replacing the true temporal kernels by metronomic temporal kernels (but keeping the exact time since the last spike $T - t_J$) again does not affect the posterior strongly. Nevertheless, the Kullback-Leibler divergence between the true posterior and the metronomic posterior is larger in the smooth than in the OU case (data not shown), indicating that the exact timing of spikes is more important in the smooth inference.

Unlike in the OU case, there is no simple analytical expression for the metronomic temporal kernel (let alone the true temporal kernel). In particular, Figure 10B shows that changing the time since the last observed spike $T - t_J$ does not simply scale the temporal kernel, but also changes the shape of the temporal kernel (it produces a complicated phase shift of the oscillating component). Again, for clarity, the metronomic kernels are used as an illustration, but the argument also applies to the exact kernels. Local structure has complex global consequences in the smooth case, with a single new spike requiring individual reweighting of all past spikes depending on their precise times. By comparison, for the OU process, the reweighting involves multiplication by a single factor. Figure 10C shows that this temporal kernel complexity is also a feature of the temporal kernel derived from the covariance function of the empirical rat trajectories in Figure 7.

The fundamental difference between the OU and the smooth temporal kernels arises from the difference in the factorization properties of the prior. Because the inverse of the covariance matrix for $\zeta \notin \{0, 1\}$, and specifically for $\zeta = 2$, is dense, it does not factorize over spike combinations and therefore does not allow a recursive form. To see that a recurrence relation is possible only for the OU prior that factorizes across duplets of spikes, write

$$p(s_T|\boldsymbol{\xi}) = \int ds_J \, d\mathbf{s}_{\overline{J}} \, p(s_T, s_J, \mathbf{s}_{\overline{J}}|\boldsymbol{\xi})$$

by expanding and integrating over the stimulus s_J at the time t_J of the last spike, and $\mathbf{s}_{\overline{J}}$ at the time of all the spikes apart from the last

$$\propto \int ds_J \, p(s_T, s_J) p(\xi_J|s_J) \int d\mathbf{s}_{\overline{J}} \, p(\mathbf{s}_{\overline{J}}, \boldsymbol{\xi}_{\overline{J}}|s_T, s_J)$$

using Bayes rule, and the instantaneity of spiking

$$= \int ds_J \, p(s_T, s_J) p(\xi_J|s_J) \int d\mathbf{s}_{\overline{J}} \, p(\boldsymbol{\xi}_{\overline{J}}|\mathbf{s}_{\overline{J}}) p(\mathbf{s}_{\overline{J}}|s_T, s_J),$$

again because the spikes are instantaneous,

$$= \int ds_J \, p(s_T, s_J) p(\xi_J|s_J) m_T(s_T, s_J, \boldsymbol{\xi}_{\overline{J}}). \tag{3.4}$$

Were $m_T(s_T, s_J, \boldsymbol{\xi}_{\overline{J}})$ independent of s_T, this would be exactly like a recursive update equation, with $p(s_T, s_J)$ being the transition probability from the last observed spike to the inference time T, $p(\xi_J|s_J)$ being the innovation due to the last observation (the likelihood of the last observed spike), and the message $m_T(s_J, \boldsymbol{\xi}_{\overline{J}})$ propagating the uncertainty from all the spikes other than the last to the last one. However, for general priors, $p(\mathbf{s}_{\overline{J}}|s_T, s_J)$, and therefore also $m_T(s_T, s_J, \boldsymbol{\xi}_{\overline{J}})$, do depend on s_T, so all spikes have to be used to infer the posterior at each time T. To make the m_T independent of s_T, the prior has to be Markov in individual spike timings, with

$$p(\mathbf{s}_{\overline{J}}|s_T, s_J) = p(\mathbf{s}_{\overline{J}}|s_J), \tag{3.5}$$

which makes

$$m_T(s_T, s_J, \boldsymbol{\xi}_{\overline{J}}) = \int d\mathbf{s}_{\overline{J}}\, p(\boldsymbol{\xi}_{\overline{J}}|\mathbf{s}_{\overline{J}})\, p(\mathbf{s}_{\overline{J}}|s_J) \tag{3.6}$$

$$\equiv m_T(s_J, \boldsymbol{\xi}_{\overline{J}}), \tag{3.7}$$

which is indeed independent of s_T. So for the OU process, the last message $m_T(s_J, \boldsymbol{\xi}_{\overline{J}})$ merely needs to be multiplied by the transition probability (see Figure 5B). However, the smooth temporal kernel changes shape in a complex way (corresponding to the dependence of the message $m_T(s_T, s_J, \boldsymbol{\xi}_{\overline{J}})$ in equation 3.4 on s_T). Again, this means that all spikes have to be kept in memory for full inference. Note, finally, that this conclusion, and the fact that there is a recursive form for the OU process, do not depend on the particular spiking model assumed, verifying the assertion that the choice of squared exponential tuning functions, although mathematically helpful, does not pose limitations on our conclusions.

3.4 Intermediate (Autoregressive) Processes. There are cases intermediate to the smooth and the OU process that allow a partially recursive formulation. For instance, the metronomic OU process can be generalized to an autoregressive model of nth order by writing

$$s_t = \sum_{i=1}^{n} \beta_i s_{t-i\Delta} + c\sqrt{\Delta}\eta_t. \tag{3.8}$$

In this case, the inverse covariance matrix $\mathcal{C}^{-1}$ is $(2n+1)$-diagonal (see appendix C), with entries determined directly by the β_i. This implies that the posterior factorizes over cliques ψ involving $n+1$ spikes (see equation 3.3), and that inference will be Markov in groups of n spikes. Zhang et al. (1998) find that a two-step Bayesian decoder, which is an AR(2) process in our terms, significantly improves decoding hippocampal place cell data.

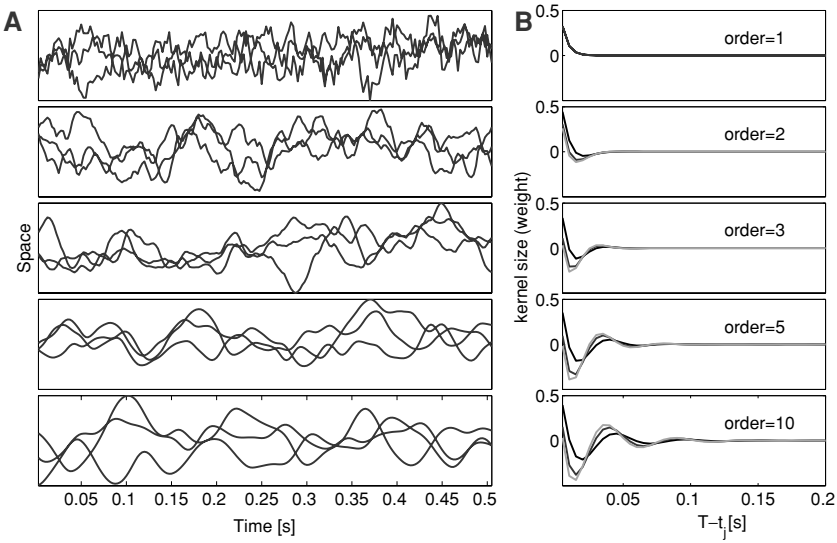

Figure 11: Autoregressive processes of increasing order. (A) Samples from processes of order $n = \{1, 2, 3, 5, 10\}$ from top to bottom. The top process corresponds to an OU process. (B) Metronomic temporal kernels $\mathbf{k}(\boldsymbol{\xi}, T)$ corresponding to the processes in A. The different lines (in varying shades of gray) correspond to increasing the observation time T as in Figures 5 and 10.

Figure 11A shows sample trajectories from such processes of increasing order. The coefficient vectors' $\boldsymbol{\beta}$ was set here such that the nth difference of the processes evolved as an OU process (see appendix C). The higher the order, the smoother the processes that can be generated, and the more oscillations are apparent in the temporal kernels. The OU and the smooth processes (see section 3.3) are at opposite ends of this spectrum, with tridiagonal and dense inverse matrices, respectively.

The higher the order, the greater the complexity of the code. Indeed, the complexity grows exponentially (since groups of n spikes have to be considered and the number of such groups increases exponentially). While natural stimulus trajectories may not be indefinitely differentiable, the exponential increase in complexity implies that any smoothness has great potential to render the code complex.

4 Expert Spikes for Efficient Computation

Complex codes, following, for instance, from the assumption of natural smooth priors, render the information inherent in the spikes hard to extract. Efficient computation in time requires access to all encoded information and

thus requires that the complex temporal structure of the code be taken into account. Here, we show that information present in the complex codes can be re-represented using codes that are straightforward to decode and use in key probabilistic computations.

Specifically, we propose to decode each spike independently and multiply together the contributions from all spikes. This corresponds to treating each spike as an independent expert in a product of experts (PoE) setting (Hinton, 1999):

$$\hat{p}(s_T|\boldsymbol{\xi}) = \frac{1}{Z(T)} \prod_i \exp\left(\sum_t g_i(s,t)\xi^i_{T-t}\right). \tag{4.1}$$

That is, each time a spike ξ^i occurs, it contributes its same projection kernel $\exp(g_i(s,t))$ to the posterior distribution $\hat{p}(s_T|\boldsymbol{\xi})$. To put it another way, for each spike, we add the same stereotyped contribution to the log posterior and then renormalize.

From the discussion in the preceding sections, it is immediately apparent that the PoE approximation is a better approximation for the OU case than for the smooth case. In the following, we first derive an approximate analytical expression for separable projection kernels $g_i(s,t) = f_i(s)h(t)$ based on metronomic spikes and the OU prior. We then remove any restrictions and derive nonparametric, nonseparable $g_i(s,t)$ for both the OU and the smooth temporal kernel and show that these still perform better for the OU process than for the smooth process. Finally, we infer a new set of spikes $\boldsymbol{\rho}_\xi$ such that decoding according to the PoE model produces a posterior distribution $\hat{p}(s_T|\boldsymbol{\rho}_\xi)$ that matches the true posterior distribution $p(s_T|\boldsymbol{\xi})$ well for both OU and smooth priors.

4.1 Approximate Projection Kernels

4.1.1 Metronomic Projection Kernels Section 3.2 showed that for the OU process, the weight accorded a spike is approximately a decreasing exponential function of the time elapsed since its occurrence and that replacing the true temporal kernels by the metronomic temporal kernels (without fixing the time since the last spike at Δ) gives a qualitatively good approximation (see Figure 4). This suggests writing an approximate distribution with spatiotemporally separable projective kernels,

$$\hat{p}(s_T|\boldsymbol{\xi}) \propto \prod_i \phi_i(s)^{\sum_t \xi^i_{T-t} e^{-\beta t}} \tag{4.2}$$

$$= \prod_i \exp\left(\sum_t \log(\phi_i(s)) e^{-\beta t}\, \xi^i_{T-t}\right) \tag{4.3}$$

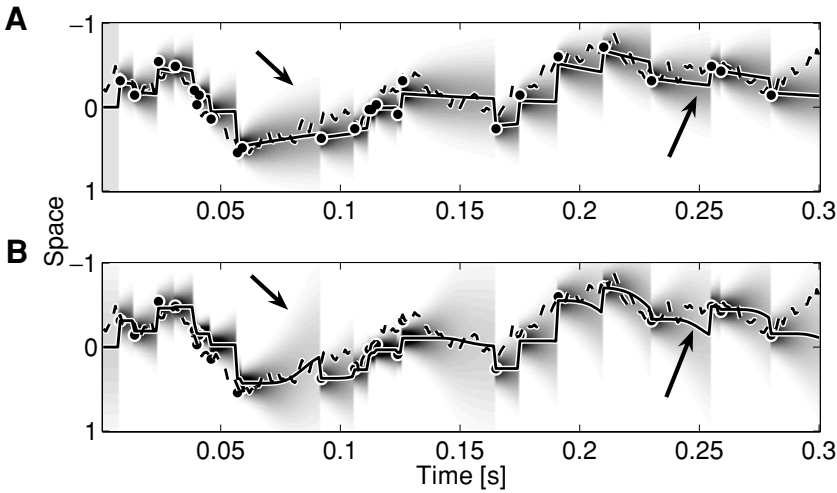

Figure 12: Separable projection kernel for the OU process: Comparison of true $p(s_T|\boldsymbol{\xi})$ (A) and $\hat{p}(s_T|\boldsymbol{\xi})$ from equation 4.3 (B). The left arrows indicate where the variance of the approximate distribution diverges toward ∞ as $T - t_J \to \infty$ rather than approaching $\mathcal{C}_{TT}$. The right arrows show the effect of this on the mean or the approximate posterior, which returns to the prior mean $\mathbf{m} = \mathbf{0}$ more rapidly than the true posterior.

to use exactly the form of equation 4.1. We can thus also write

$$\hat{p}(s_T|\mathbf{A}) \propto \prod_i \phi_i(s)^{A_i(T)}, \tag{4.4}$$

where $A_i(T)$ can be seen as an equivalent "activity" of each neuron. The performance of this approximation is shown in Figure 12 for the OU process (see also Zemel et al., 2005). There are a few differences between Figures 4 and 12. Keeping the $\phi_i(s)$ as before, the variance of this approximation is $\hat{\nu}^2(T) = \sigma^2/\sum_i A_i(T)$. As the last observed spike recedes into the past, this approaches infinity (left arrows in Figure 12), and the mean returns to zero (right arrows in Figure 12). This is different from the case of exact inference, which approaches the static prior with variance $\mathcal{C}_{TT}$. The mean $\hat{\mu}(T) = \sum_i s_i \frac{A_i(T)}{\sum_j A_j(T)}$ is always normalized and returns to zero more slowly than the variance increases. This introduces an inaccuracy, since the true OU temporal kernels (shown in Figure 5) are not normalized $\sum_t k_t(\boldsymbol{\xi}, T) < 1$, which arises because of the weight given to the spatial prior.

For the smooth case, no simple approximation of the form of equation 4.3 is viable. This can be seen, for instance, from the fact that the smooth temporal kernels (see Figure 10) dip below zero (making it tricky to use them in products).

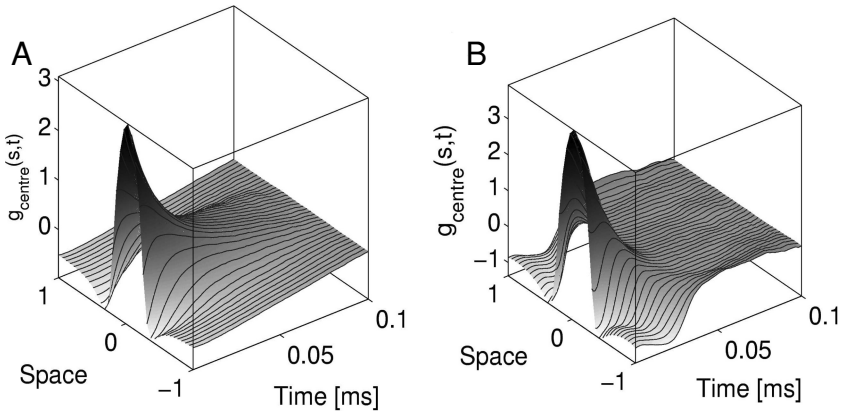

Figure 13: Projection kernels inferred by equation 4.5 for OU (A) and smooth (B) priors. Stimulus trajectories and corresponding population spike trains $\boldsymbol{\xi}$ were generated until the update equations converged (approximately $2 \cdot 10^4$ spike trains). Both kernels have the shape of difference of gaussians for $t = 0$ and fall off exponentially with time. There is little nonseparable structure in both cases.

4.1.2 Inferring Full Spatiotemporal Projection Kernels $g_i(s,t)$. To apply expression 4.1 to the smooth case, we inferred $g_i(s,t)$ in a nonparametric way by discretizing time and space over which the distributions are defined and minimizing the Kullback-Leibler divergence between the discretized versions $p(s_T|\boldsymbol{\xi})$ and $\hat{p}(s_T|\boldsymbol{\xi})$ with respect to the projection kernels,

$$g_i(s,t) \leftarrow g_i(s,t) - \varepsilon \nabla_{g_i(s,t)} D_{KL}(p(s_T|\boldsymbol{\xi})||\hat{p}(s_T|\boldsymbol{\xi})), \tag{4.5}$$

where $D_{KL}(p(s)||q(s)) = \int ds\ p(s) \log \frac{p(s)}{q(s)}$. Given that our approximation 4.1 is related to restricted Boltzmann machines (RBM), it is not surprising that the gradient has a form akin to the wake-sleep algorithm (Hinton, Dayan, Frey, & Neal, 1995):

$$\nabla_{g_i(s,t)} D_{KL}(p(s_T|\boldsymbol{\xi})||\hat{p}(s_T|\boldsymbol{\xi})) = \sum_T [\hat{p}(s_T|\boldsymbol{\xi}) - p(s_T|\boldsymbol{\xi})]\, \xi_i(T-t). \tag{4.6}$$

Figure 13 shows the projection kernels inferred for the OU prior (see Figure 13A) and the smooth prior (see Figure 13B). Both start, for $t = 0$ with a spatial profile similar to a difference of gaussians (DOG), and then fall off as exponentials of time. The kernels $g_i(s,t)$ shown here are for neurons i with s_i close to 0, the center of the gaussian prior over the trajectories. The projection kernels shown are for the same parameter settings as Figures 4 and 8, and the faster decay of the smooth projection kernels is due to the

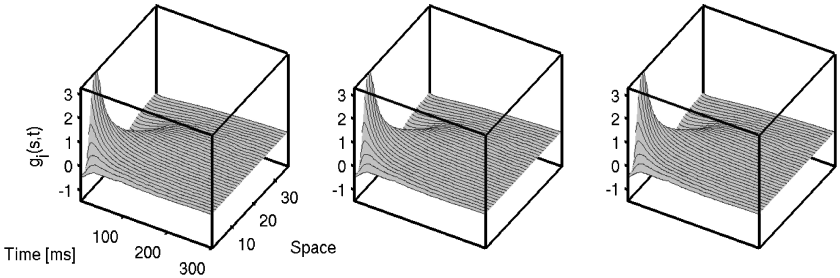

Figure 14: Projection kernels are independent of contrast. The left-most panel shows an OU kernel for the same contrast ($\phi_{\max}$) as in Figure 13; the contrast is doubled in the middle and quadrupled in the right panel. All these are off-center kernels with the same parameters as used in the other Figures. Despite a slight slant toward the mean, the kernels are approximately separable.

shorter correlation timescale. For the OU process, the kernels for neurons i with $s_i > 0$ become slightly slanted toward -1 over time (and the converse holds for those with $s_i < 0$) to capture the decay to the mean (zero), which is only a function of the distance from the mean. This effect is noticeable for the OU but very small for the smooth kernels. Figure 14 shows off-center OU kernels inferred for different contrast (by varying $\phi_{\max}$). As can be seen, the kernels are invariant to the contrast, and the slant effect is small. For the parameter range explored here, both projection kernels are approximately separable, indicating that the analytically derived motivation above may be close to optimal and that, in the PoE framework of equation 4.1, separable projection kernels may be the optimal choice even for the smooth prior. However, simply using these projection kernels to interpret the original spikes $\boldsymbol{\xi}$ results in an approximation that is far from perfect, especially in the smooth case. Figure 15 compares the true posterior distribution and that given by the approximation with the above projection kernels. The cost of independent decoding is quantified in Figure 15A using

$$\left\langle \frac{1}{T} \sum_t \frac{D(p(s_T|\boldsymbol{\xi})||\hat{p}(s_T|\boldsymbol{\xi}))}{\mathcal{H}(p(s_T|\boldsymbol{\xi}))} \right\rangle_{p(\mathbf{s},\boldsymbol{\xi})}, \tag{4.7}$$

where $\mathcal{H}(p)$ is the entropy of p and the average is over many stimulus trajectories $\mathbf{s} \sim \mathcal{N}(0, \mathcal{C})$ and spikes $\boldsymbol{\xi} \sim p(\boldsymbol{\xi}|\mathbf{s})$. This quantity can also be interpreted as a percentage information loss. It is larger for the smooth than for the OU process, showing that the OU process suffers much less from the approximation than the smooth prior. Visually, there are no gross differences between $p(s_T|\boldsymbol{\xi})$ and $\hat{p}(s_T|\boldsymbol{\xi})$ for the OU prior (see Figures 15B and 15D). However, for the smooth prior, the arrows in

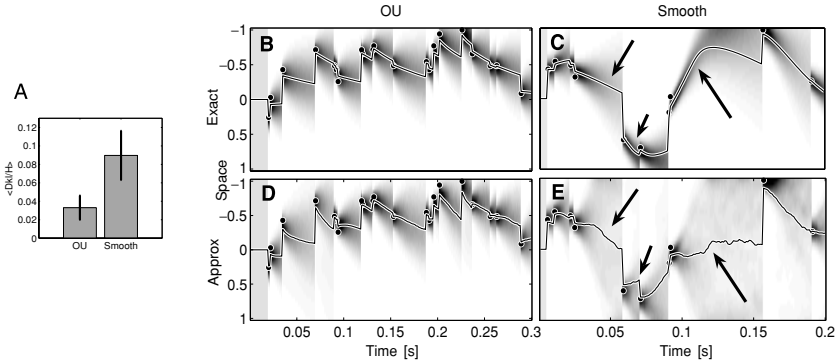

Figure 15: Comparison of true distribution $p(s_T|\boldsymbol{\xi})$ and approximate distribution $\hat{p}(s_T|\boldsymbol{\xi})$ given by equation 4.1 with projection kernels inferred by equation 4.5 and shown in Figure 13. Organization is the same as in previous figures. (A) $\langle \frac{1}{T} \sum_t D(p(s_T|\boldsymbol{\xi})||\hat{p}(s_T|\boldsymbol{\xi}))/H(p(s_T|\boldsymbol{\xi}))\rangle_{p(\boldsymbol{\xi},\mathbf{s})} \pm 1$ standard deviation for both priors. (B, C) $p(s_T|\boldsymbol{\xi})$. (D, E) The corresponding $\hat{p}(s_T|\boldsymbol{\xi})$ for the same spikes. (B, D) A stimulus generated from the OU prior. (C, E) The smooth prior. $\hat{p}(s_T|\boldsymbol{\xi})$ is a good approximation for the OU prior but fails for the smooth prior. The arrows indicate where the approximation fails fundamentally in a similar way to that shown in Figure 9.

Figures 15C and 15E indicate areas where a large mismatch is introduced by the independent treatment of the spikes, which discards all information contained in spike combinations. This mismatch is entirely to be expected.

4.2 Recoding: Finding Expert Spikes. The previous section has shown that an independent interpretation of spikes is more costly with the smooth than with the OU prior. In this section, we show that it is possible to find a new set of "expert" spikes $\boldsymbol{\rho}$, such that each spike can be interpreted independently and the posterior distribution is matched closely for both the OU and the smooth prior. This recoding thus takes spikes $\boldsymbol{\xi}$ that are redundant in a decoding sense and produces a new set of spikes $\boldsymbol{\rho}$ that can be easily used for efficient neural computation because the decoding redundancy has been eliminated. We first infer real-valued activities $\mathbf{a}_\xi$ and then proceed to infer actual spikes $\boldsymbol{\rho}$. We use neurally implausible methods to infer the new set of spikes $\boldsymbol{\rho}$. In a companion paper we will explore the capability of neurally plausible spiking networks to do this recoding and to use the resulting simple code for probabilistic computations in time (see also Zemel, Huys, Natarajan, and Dayan, 2004).

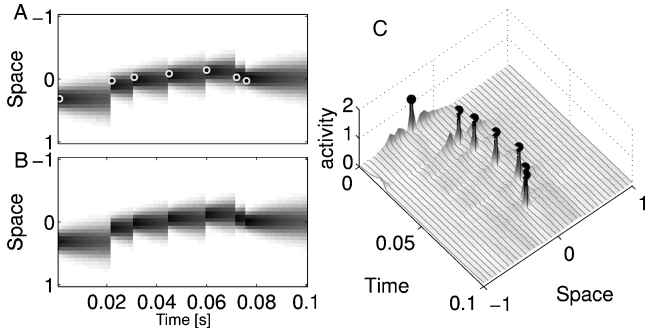

Figure 16: Inferring activities $\mathbf{A}$ for the OU prior. (A) True posterior $p(s_T|\boldsymbol{\xi})$. (B) Approximate posterior $\hat{p}(s_T|\mathbf{A})$, which matches arbitrarily well (for this example, $\langle D_{KL}\rangle_T \sim 10^{-5}$ and the entropy $\langle \mathcal{H}\rangle_T \sim 2$, making the information loss $\Delta I \sim 10^{-5}$). (C) Activities $\mathbf{A}$ for all neurons. The vertical black lines with dots indicate the original spike times $\boldsymbol{\xi}$. Each thin line along the gray surface is the "activity" of one neuron as a function of time. There is a small amount of activity away from the spikes, but zeroing this affects the match between $p(s_T|\boldsymbol{\xi})$ and $\hat{p}(s_T|\mathbf{A})$ only marginally.

4.2.1 Activities. Given a set of projection kernels $g_i(s,t)$ from the previous section, we can go back and infer the optimal activities $\mathbf{A} \geq 0$ of neurons by writing

$$\hat{p}(s_T|\mathbf{A}) \propto \exp\left(\sum_{i,t} A_i(T-t) g_i(s,t)\right). \tag{4.8}$$

If we let $A_i(T-t) = \exp(B_i(T-t))$ and minimize with respect to $\mathbf{B}$ the Küllback-Leibler divergence from the true posterior, we simultaneously enforce $\mathbf{A} \geq 0$:

$$B_i(t) \leftarrow B_i(t) - \varepsilon \nabla_{B_i(t)} D_{KL}\left(p(s_T|\boldsymbol{\xi}) || \hat{p}(s_T|\mathbf{A})\right). \tag{4.9}$$

The results of this procedure are shown for both the OU process (see Figure 16) and for the smooth process (see Figure 17). Figures 16A and 17A show the true spikes $\boldsymbol{\xi}$ and the corresponding distribution $p(s_T|\boldsymbol{\xi})$. Figures 16B and 17B show the approximate distributions $\hat{p}(s_T|\mathbf{A})$ defined in equation 4.8 for the optimal activities $\mathbf{A}$ inferred with equation 4.9. The continuous nature of the activation functions means that they can contain as much information as the distribution itself, and indeed we find empirically that arbitrarily close matches are possible (exemplified by the two Figures; in both cases $\langle D_{KL}\rangle_T \sim 10^{-5}$). Figures 16C and 17C finally show

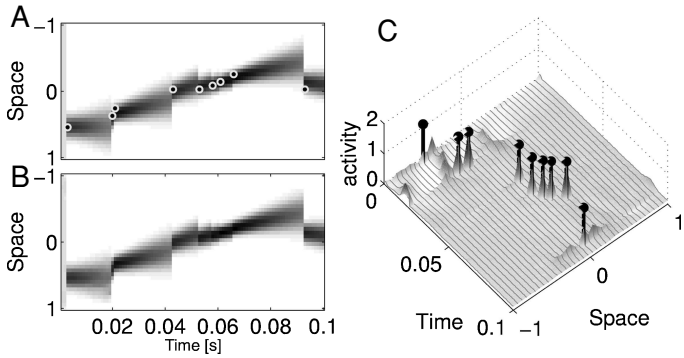

Figure 17: Inferring activities **A** for the smooth prior. (A) True posterior $p(s_T|\boldsymbol{\xi})$. (B) Approximate posterior $\hat{p}(s_T|\mathbf{A})$, which matches arbitrarily well (for this example, $\langle D_{KL}\rangle_T \sim 10^{-5}$ and the entropy $\langle \mathcal{H}\rangle_T \sim 2$, making the information loss $\Delta I \sim 10^{-5}$). (C) Activities **A** for all neurons. The vertical black lines with dots indicate the original spike times $\boldsymbol{\xi}$. Each thin line along the gray surface is the "activity" of one neuron as a function of time. There is a small amount of activity away from the spikes, which allows the approximation $\hat{p}(s_T|\mathbf{A})$ to "bend" between spikes. Unlike in the OU case, zeroing this small activity affects the match between $p(s_T|\boldsymbol{\xi})$ and $\hat{p}(s_T|\mathbf{A})$ strongly.

the inferred activities **A**. Most important, we see that the inferred activities (one gray line for each neuron) are very sparse (in time and across neurons), suggesting that there might indeed be a set of (zero-one) spikes that leads to a good approximation via equation 4.1. On the other hand, the activities line up closely with the original spikes $\boldsymbol{\xi}$ (vertical black lines with dots), and it may be that the approximations with the original spikes in the previous paragraph already gave us the best possible approximation. For the OU prior (see Figure 16), the activities in spikeless times are extremely small, and zeroing them does not significantly worsen the approximation with $\hat{p}(s_T|\mathbf{A})$. However, for the smooth prior (see Figure 17), there is residual activity between the peaks, the zeroing of which significantly worsens the quality of approximation by $\hat{p}(s_T|\mathbf{A})$ (data not shown). The very close approximation found here is not surprising. The distribution to be matched and the activities are discretized on the same spatial grid and are both positive, real quantities. As we have not imposed any constraints on **A** beyond positivity, the entropy of **A** can match any entropy in the distributions $p(s_T|\boldsymbol{\xi})$. This, however, will change drastically when the activities are forced to be binary.

4.2.2 Spikes via Simulated Annealing. To check whether there exists in fact a set of spikes $\boldsymbol{\rho}_\xi$ such that decoding according to equation 4.1 results in a posterior distribution $\hat{p}(s_T|\boldsymbol{\rho})$ that matches $p(s_T|\boldsymbol{\xi})$ closely for the smooth

prior, we assume, as before, the projection kernels $g_i(s,t)$ inferred above and find a set of spikes $\boldsymbol{\rho}$ that minimizes $D_{KL}(p(s_T|\boldsymbol{\xi})||\hat{p}(s_T|\boldsymbol{\rho}))$, where $\hat{p}(s_T|\boldsymbol{\rho})$ given by

$$\hat{p}(s_T|\boldsymbol{\rho}) = \frac{1}{Z(T)} \prod_i \exp\left(\sum_t g_i(s,t)\rho^i_{T-t}\right). \tag{4.10}$$

This is the same interpretation we gave the original spikes in equation 4.1. Our aim is thus to find a new set of spikes that satisfies

$$\boldsymbol{\rho} = \arg\min_{\boldsymbol{\rho}} D_{KL}(p(s_T|\boldsymbol{\xi})||\hat{p}(s_T|\boldsymbol{\rho})). \tag{4.11}$$

This is a highly nonconvex discrete problem, so we applied standard simulated annealing techniques.[2] Figure 18 shows the results. Figure 18A shows the distribution $p(s_T|\boldsymbol{\xi})$ based on the original spikes. Figure 18B shows $\hat{p}(s_T|\boldsymbol{\rho})$ using the projection kernels shown in Figure 13. The arrow in Figure 18B indicates where the new set of spikes performs better than the original, independently interpreted, spikes and matches the shifting distribution by adding a new spike. This is a qualitative improvement on what is possible by interpreting the original spikes according to equation 4.1 (see Figure 15) and is most pronounced for the smooth case. From the close match between $p(s_T|\boldsymbol{\xi})$ and $\hat{p}(s_T|\mathbf{A})$, we expect the overall minimum KL divergence to depend strongly on the projective kernels. Figure 18C shows the effect of scaling the inferred projection kernel $g_i(s,t)$. The increase in accuracy due to an increase in the number of spikes offsets the decrease in accuracy due to the absence of prior information. The more spikes are allowed, the closer this scheme is to the one where rates are allowed. Thus, the prior has here literally been replaced by spikes; the input spikes have been "augmented" with spikes that represent information contributed by past spikes in accordance with a prior over stimulus trajectories. Figure 18D shows relative average KL divergences over 100 sets of new spike trains, for different scalings of the $g_i(s,t)$. In general, the projection kernels found here form an overcomplete basis set. By scaling them down and allowing more spikes, we come closer to the setting in the previous section where we allowed continuous activities rather than 0-1 spikes.

Note the different coding strategy indicated by the arrows, especially in Figure 18C. Here, spikes are positioned such that they take into account

[2] From the very strong sensitivity of our simulated annealing results to the procedure used to reduce the temperature, we infer that the optima are not very well separated, with a number of similar sets of spike trains doing approximately equally well. We rendered the procedure more global by evaluating, at each step, the decrease in cost that would accompany switching every spike and accepting one of the best switches probabilistically.

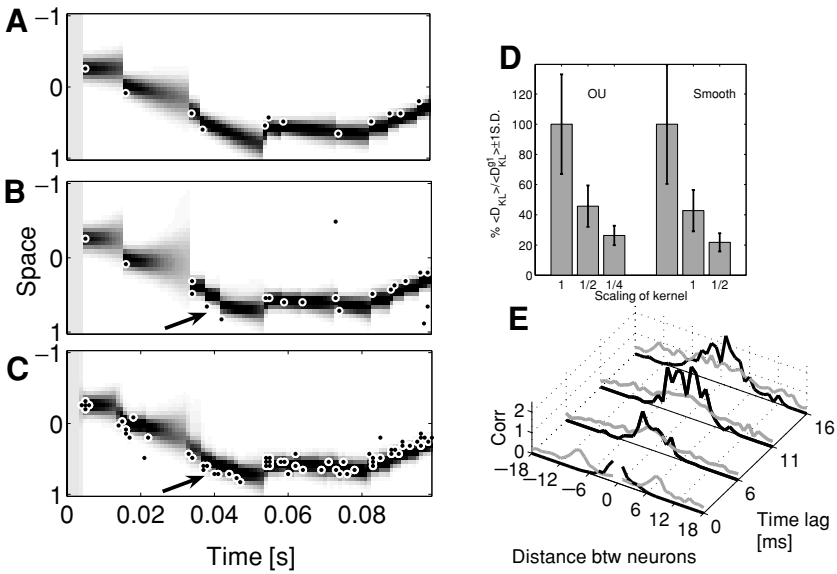

Figure 18: Inferring new spikes for smooth prior. (A) Original spikes with $p(s_T|\boldsymbol{\xi})$. (B) $\hat{p}(s_T|\boldsymbol{\rho})$ with projection kernels $g_i(s,t)$ given by equation 4.5. (C) $\hat{p}(s_T|\boldsymbol{\rho})$ with scaled projection kernels $g_i(s,t)/4$. Note the increased firing rate. Arrows explained in the text. (D) Percentage change in KL divergences between $p(s_T|\boldsymbol{\xi})$ and $\hat{p}(s_T|\boldsymbol{\rho})$ for projection kernels scaled by factors of 1/2 and 1/4 relative to the KL divergence of the unscaled kernel. (E) Cross-correlations between neurons for a few different time lags. Black lines: original spikes $\boldsymbol{\xi}$. Gray lines: recoded spikes $\boldsymbol{\rho}$ for unscaled projection kernels. The autocorrelation has been scaled to unity, except at zero lag, where the autocorrelation was excluded and scaling was performed with respect to the maximal cross-correlation.

what has already been expressed by previous spikes: spikes are positioned with respect to the distribution that has already been encoded. To put it another way, there are explicit relations among the new spikes that are not directly explained by the stimulus itself. Figure 18E shows this more clearly. The black traces show the stimulus-based ("signal") correlations of the original spikes $\boldsymbol{\xi}$. The gray lines show the correlations of the recoded spikes $\boldsymbol{\rho}$. At lag 0 (bottom of the Figure), flanks appear in the cross-correlation functions, but at greater lags, the cross-correlations are flatter for the recoded than for the original spikes. Requiring independently decodable spikes has introduced instantaneous correlations and flattened the spatial profile of cross-correlations over time—a sign of adaptation to temporal statistics. Thus, here we find that the maintenance of a simple code results in the emergence of what appear to be adaptive properties.

5 Discussion

Here, we have analyzed the structure of a Bayesian, optimal decoder in a simple, analytically tractable model. The results are a direct generalization of decoding in the static gaussian-Poisson encoding model (Snippe & Koenderinck, 1992). We show that the structure of the decoder depends on the prior over stimulus trajectories in time, that realistic priors render decoding hard (nonlocal in time and space), and that an independent code in which information is readily available for computational purposes exists. Finally, we showed that apparently adaptive properties of a coding scheme may result from the requirement of a constant, simple code. We are currently working on a biologically plausible network that approximates this recoding and uses the resulting code for flexible probabilistic computations (see also Zemel et al., 2004).

The main innovation in our work is the way we used the gaussian process prior over stimulus trajectories. Figure 7 indicates that the exponential prior with $\zeta = 2$ is a good model of natural movements, as they tend to be smooth. Classically, natural stimuli have been characterized as having autocorrelation structures that fall off exponentially (Dong & Atick, 1995), corresponding to a power spectrum that falls off as a power law function of frequency ($\propto 1/\omega^b$). However, smooth trajectories in time have a much faster (exponential) spectral falloff $\propto \exp(-\omega^2)$.

Most previous work has assumed priors within the OU class (Brown et al., 1998; Smith & Brown, 2003; Barbieri et al., 2004; Kemere et al., 2004; Gao, Black, Bienenstock, Shoham, & Donoghue, 2002), perhaps because of the recursive formulation of decoding. However, Zhang et al. (1998) use a two-step Bayesian decoder corresponding to a second-order autoregressive process (AR(2)) with coefficients that fall off as squared exponentials (their equation 43). This two-step decoder is much more competent than a one-step decoder (corresponding to an AR(1) process) on hippocampal place cell data.

Of course, more complex priors are also possible. For instance, Kemere et al. (2004) point one way forward. They showed the benefits for decoding from motor cortex spikes of using a rich, modular prior based on separate models for each of the (seven) possible arm movements to be extracted.

In terms of our framework, Figure 15A illustrates the cost of neglecting the prior temporal structure and treating all spikes independently. The differences between inference in the smooth and OU case (e.g., the overshoot in Figure 8, which is not seen in Figure 4) also indicate qualitatively what information is lost by applying Kalman-filter-like formulations to decoding. The absolute magnitude of this effect depends on the specifics of the true model, and so remains an empirical question for psychophysical or physiological test. If spikes are dense relative to the movement in the stimulus (the likelihood term in equation 2.2 dominates, either via a very small noise (low σ) or high firing rates (large $\phi_{\max}$)), the contribution of the

prior will be small, and approximation by a recursive prior (as in Brown et al., 1998; Zhang et al., 1998) may suffice. However, if spikes are sparse, the prior will be more important and approximations more costly. Finally, in many cases, the correct prior can be acquired only from experience, which itself may be costly. While it is sensible to expect nervous systems to acquire detailed and correct informative priors (Körding & Wolpert, 2004; Körding, Ku, & Wolpert, 2004; Adams, Graf, & Ernst, 2004), it remains to be seen whether incorporation of informative priors is generally feasible for decoding applications in engineering domains (e.g., brain-machine-interfaces).

The historical approach to population coding is based on ideas of Fisher information. The Fisher information arises from notions of asymptotic normality where there are many "data"—long-spike counting windows and many neurons. In the asymptotic limit, the posterior distribution is well approximated by a gaussian with width $(JI_F)^{-1}$ where J is the number of data points or spikes in our case. This is a linear expansion where each data point (spike) contributes the same amount $1/I_F$ to the variance of the posterior. We, like others before us (Brunel & Nadal, 1998; Bethge, Rotermund, & Pawelzik, 2002), are interested in the case where the population as a whole has emitted few spikes, as indicated in Figure 1—in regimes far from the asymptotic limit. For us, spikes can contribute varied amounts: some (typically the most recent) very much more than others (the most distant). As an analog of Fisher information, it is possible to study the dependence of the posterior variance $\nu^2(T)$ on the width of the encoding tuning functions σ^2 and the dimensionality. In our simple model, we find similar results to previously reported ones (Snippe & Koenderinck, 1992; Zhang & Sejnowski, 1999) (data not shown). However, as we are always in the sparse spike limit, the information per spike is of most relevance, and the posterior variance is strictly increasing in σ, the width of the encoding tuning functions, independent of the dimensionality. If there were dense spiking, the population firing rate (Zhang & Sejnowski, 1999; Silberberg, Bethge, Markram, Pawelzik, & Tsodyks, 2004; DeCharms & Zador, 2000; Knight 1972) might carry enough information to overwhelm any prior.

Three assumptions about the encoding model merit discussion. First, the bell-shaped form of the tuning functions, asymptoting at 0, is only very roughly realistic. However, the arguments in sections 3.2 and 3.3 about the recursive and nonrecursive structure of the decoder depend on the nature of the prior ($\zeta = 1$ and $\zeta = 2$), and so will generalize. The most fundamental change would be that the variance of the posterior would depend not only on spike timing but also on the relative tuning preferences of the neurons that emit the spikes.

Second, we assume an instantaneous relationship between the rate of the inhomogeneous Poisson process and the stimulus. This should be formally straightforward to relax if the dependence on the stimulus history can be approximated by a linear filter (a discrete sum) as is standard for

linear-nonlinear-Poisson like model neurons (Paninski, 2003). In that case, the likelihood term in equation 2.5 will become a function of the stimulus at a number of times, each of which enters equation 2.2. Each spike will thus contribute as many entries to the covariance matrix as its linear filter extends in time. The extra complexity of decoding in this case is not completely clear.

The third questionable assumption is that spiking is independent across time (the Poisson process assumption) and neurons. The actual degree of independence is, of course, hotly debated (Averbeck et al., 2006; DeCharms & Zador, 2000). In fact, the signal correlations that we assume induce some of the same issues for decoding that the noise correlations discussed in those references. Thus, our work can be seen as casting this debate in a slightly different light by showing that it is possible to recode correlated spiking in a particular, independently interpretable, form. Our companion paper (Natarajan et al., 2006) considers a network- (rather than a simulated-annealing-) based implementation of recoding. Note that Nirenberg, Carcieri, Jacobs, and Latham (2001) have studied independent interpretability (in the spikes of retinal ganglion cells) in a less model-dependent manner.

The advantage of independent interpretability (based on equation 4.1) is not confined to decoding. For instance, combining information from different modalities—as in multisensory integration (Ernst and Banks, 2002; Hillis, Ernst, Banks, & Landy, 2002) or sensorimotor integration (Zemel et al., 2005; Körding and Wolpert, 2004)—becomes straightforward and requires only an addition in the log domain or single-neuron (Chance, Abbott, & Reyes, 2002; Salinas & Abbott, 1996; Poirazi, Brannon, & Mel, 2003a, 2003b) or population (Deneve et al., 2001) multiplication.

Recoding produces spike trains that lack temporal redundancy. It is the exact dynamic analog of efficient coding approaches based on producing population activities lacking, for example, spatial correlations (Srinivasan, Laughlin, & Dubs, 1982; Atick, 1992; Nirenberg et al., 2001). As such, it displays phenomena that are strongly reminiscent of adaptation in the static domain. Note, however, that the rationale for adaptation here is different—it is a by-product of the requirement for a computationally efficient code. This rationale may find application outside our particular domain.

Finally, the probabilistic coding here arises only from the ill-posed nature of recovering the spike train from a sparse set of noisy spikes. A more fundamental form of so-called computational uncertainty (Zemel, Dayan, & Pouget, 1998) arises in cases such as the aperture problem (Weiss & Adelson, 1998), when the information in the sensory input is ambiguous in a way that does not depend on noise. Various approaches to computational uncertainty have been suggested in the static case (Anderson, 1994; Barber, Clark, & Anderson, 2003; Zemel et al., 1998; Sahani & Dayan, 2003), but their extension to our dynamic framework remains an open problem.

Appendix A: Matrix Inversion Lemmas

For a matrix partitioned as in equations 2.8, or generally

$$\mathbf{E} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right),$$

it can be shown that the following equalities hold by inverting **E**:

$$(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} = \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \tag{A.1}$$

$$(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} = \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}. \tag{A.2}$$

These identities allow us to perform the final multiplication $p(\boldsymbol{\xi}|s_T)p(s_T)$, here written in the log domain, and then to renormalize:

$$\begin{aligned} \log(p(\boldsymbol{\xi}|s_T)p(s_T)) = -\frac{1}{2}\Big\{ & s_T \left[\mathcal{C}_{TT}^{-1} + \mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi}\left(\mathcal{C}_{\xi\xi} - \mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi} + \mathbf{I}\sigma^2\right)^{-1}\mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\right] s_T \\ & - 2s_T\mathcal{C}_{TT}^{-1}\mathcal{C}_{\xi T}\left[\mathcal{C}_{\xi\xi} - \mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi} + \mathbf{I}\sigma^2\right]^{-1} \cdot \boldsymbol{\theta} + \text{const.}\Big\} \\ \text{thus} \qquad \nu^2(T) = & \left(\mathcal{C}_{TT}^{-1} + \mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi}\left(\mathcal{C}_{\xi\xi} + \mathbf{I}\sigma^2 - \mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi}\right)^{-1}\mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\right)^{-1}. \end{aligned}$$

Making the following substitutions

$$\begin{aligned} \mathbf{A} &= \mathcal{C}_{\xi\xi} + \mathbf{I}\sigma^2 & \mathbf{B} &= \mathcal{C}_{\xi T} \\ \mathbf{C} &= \mathcal{C}_{T\xi} & \mathbf{D} &= \mathcal{C}_{TT} \end{aligned}$$

allows us to apply equation A.1 and write out the variance in equation 2.10. The mean finally is obtained by writing out

$$\begin{aligned} \mu(t) &= \nu^2(T)\mathcal{C}_{TT}^{-1}\mathcal{C}_{\xi T}\left[\mathcal{C}_{\xi\xi} - \mathcal{C}_{\xi T}\mathcal{C}_{TT}^{-1}\mathcal{C}_{T\xi} + \mathbf{I}\sigma^2\right]^{-1} \cdot \boldsymbol{\theta} \\ &= (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} \cdot \boldsymbol{\theta} \end{aligned}$$

and applying equation A.2 to directly yield equation 2.9.

Appendix B: OU Process

Replacing each of the ISIs by the average value Δ, we get a Kac-Murdock-Szego Toeplitz matrix for which the analytical inverse is (Dow, 2003):

$$\mathcal{C} = c \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \qquad \mathcal{C}^{-1} = \frac{1}{c(1-\rho^2)} \begin{bmatrix} 1 & -\rho & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & -\rho & 1 \end{bmatrix},$$

where $\rho = \exp(-\alpha\Delta)$. Rewriting equation 2.11 as $\mathbf{k}(\boldsymbol{\xi}, T) = \mathcal{C}_{T\xi}\mathcal{C}_{\xi\xi}^{-1}/\sigma^2(\mathcal{C}_{\xi\xi}^{-1} + \mathbf{I}/\sigma^2)^{-1}$, we note that $\mathcal{C}_{T\xi}\mathcal{C}_{\xi\xi}^{-1} \approx \delta_{i-1}$; only the first component of this vector is one, and all others are zero. The second factor is

$$\mathbf{A}^{-1} = (\mathcal{C}^{-1} + \mathbf{I}/\sigma^2) = \frac{1}{(a-1)\sigma^2} \begin{bmatrix} a & -\rho & 0 & 0 & 0 \\ -\rho & a+\rho^2 & -\rho & 0 & 0 \\ 0 & -\rho & a+\rho^2 & -\rho & 0 \\ 0 & 0 & -\rho & a+\rho^2 & -\rho \\ 0 & 0 & 0 & -\rho & a \end{bmatrix},$$

where $a = \frac{c}{\sigma^2}(1-p^2) + 1$. We know $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Neglecting the prefactor for a moment, the first row of $\mathbf{A}$ (which is the one of interest) therefore has to satisfy the following recurrence relation:

$$A_{2,1} = (a\,A_{1,1} - 1)/\rho \tag{B.1}$$

$$A_{k+2,1} = (a/\rho + \rho)A_{k+1,1} - A_{k,1} \quad \text{for} \quad n > 3 \tag{B.2}$$

$$A_{N,1}/A_{N-1,1} = \rho/a\,. \tag{B.3}$$

Equation B.2 is a simple two-term linear recurrence equation and can be solved with boundary conditions given by equations B.1 and B.3. The characteristic equation of equation B.2 is

$$r^2 - (a/\rho + \rho)r + 1 = 0 \quad \text{with real roots}$$

$$\lambda_{1,2} = \frac{1}{2}\left(a/\rho + \rho \pm \sqrt{(a/\rho+\rho)^2 - 4}\right).$$

Including the boundary conditions leads to a solution

$$A_{n,1} = d_1\lambda_1^{n-1} + d_2\lambda_2^{n-1}$$

$$d_1 = \left(a - \lambda_1\rho - (a - \lambda_2\rho)\frac{(a\lambda_1 - \rho)}{(a\lambda_2 - \rho)}\left(\frac{\lambda_1}{\lambda_2}\right)^{N-2}\right)^{-1}$$

$$d_2 = \frac{1 - d_1(a - \lambda_1\rho)}{a - \lambda_2\rho}.$$

One of the eigenvalues will always be greater than 1, the other less than 1, but both are positive. As $\mathcal{C}_{\xi\xi}$ is symmetric, so are $\mathbf{A}^{-1}$ and $\mathbf{A}$, and the first column of $\mathbf{A}$ is equal to its first row, which we pick out by premultiplying with $\mathcal{C}_{T\xi}\mathcal{C}_{\xi\xi}^{-1}$. This vector $\mathbf{A}_{1,1:N}$ is exactly the sum of two exponentials we saw when using regular spikes to infer the temporal kernel $\mathbf{k}(\boldsymbol{\xi}, T)$, and the nth component of $\mathbf{k}(\boldsymbol{\xi}, T)$, k_n, is given by

$$k_n = [\mathcal{C}_{T\xi}(\mathcal{C}_{\xi\xi} + \mathbf{I}\sigma^2)^{-1}]_n = (a-1)\sigma^2 A_{n,1} = (a-1)\sigma^2\big(d_1\lambda_1^{n-1} + d_2\lambda_2^{n-1}\big). \tag{B.4}$$

If λ_1 is the larger eigenvalue, we see that the corresponding coefficient d_1 will be $\approx (\lambda_2/\lambda_1)^N$, which is very small. The contribution of the larger λ will grow only very slowly and be seen only for the very distant spikes. On the other hand, d_2 will be $\approx 1/(a - \lambda_2\rho)$. For all intents and purposes, the temporal kernel will be decaying exponentially with a negative spike time constant $\log \lambda_2$. Furthermore, if the second boundary condition (for time 0) is moved to $-\infty$, the result is a pure exponential. Both the analytical and numerical kernels are plotted in Figure 5.

Relaxing the assumption of metronomic spiking gives a matrix $\mathbf{A}^{-1}$, which is still tridiagonal but the elements of which are not equal. Writing matrix $\mathcal{C}$ as

$$\mathcal{C} = \begin{bmatrix} 1 & a & ab & abd \\ a & 1 & b & bd \\ ab & b & 1 & d \\ abd & bd & d & 1 \end{bmatrix} \quad \mathcal{C}^{-1} = \begin{bmatrix} \frac{1}{1-a^2} & -\frac{a}{1-a^2} & 0 & 0 \\ -\frac{a}{1-a^2} & \frac{1-a^2b^2}{(1-a^2)(1-b^2)} & -\frac{b}{1-b^2} & 0 \\ 0 & -\frac{b}{1-b^2} & \frac{1-b^2d^2}{(1-b^2)(1-d^2)} & -\frac{d}{1-d^2} \\ 0 & 0 & -\frac{d}{1-d^2} & \frac{1}{1-d^2} \end{bmatrix}, \tag{B.5}$$

where $a = ce^{-\alpha|t_1-t_2|}$, $b = ce^{-\alpha|t_2-t_3|}$, and soon, leads to a set of equations similar to equations B.2 and B.3 but including more terms.

Appendix C: Autoregressive Processes of Second and Higher Order

An nth-order gaussian autoregressive sequence of length T as produced by equation 3.8 can be written as a sample from a multivariate normal distribution in the following way: Let $\mathbf{b} = [1, -\beta_1, -\beta_2, -\beta_3, \ldots, \beta_N]$ and

let $\mathcal{B}_t = [\mathbf{0}_t, \boldsymbol{\beta}, \mathbf{0}_{T-n-t}]$, where $\mathbf{0}_t$ stands for a vector of zeros of length t. The inverse covariance matrix of the process is given by

$$\mathcal{C}^{-1} = \sum_{t=0}^{T-n-1} \mathcal{B}_t \mathcal{B}_t^T. \tag{C.1}$$

For the coefficients of $\mathbf{b}$ to define a stationary and finite process, $\mathcal{C}$ must be Toeplitz. One way of generating a finite process from the $\mathbf{b}$ is by letting the nth derivative of the process evolve as an OU process,

$$s_t^{(n)} = \beta_0 s_{t-1}^{(n)} + c\sqrt{\Delta}\eta_t, \tag{C.2}$$

in which case the coefficients of the vector $\mathbf{b}$ are given by

$$\boldsymbol{\beta}_i = {}^nC_i(-\beta_0)^{i-1}, \tag{C.3}$$

where ${}^nC_i$ is the binomial coefficient. To enforce stationarity, we have to finally perform a subtraction,

$$\mathcal{C}^{-1} = \left(\sum_{t=0}^{T-1} \mathcal{B}_t \mathcal{B}_t^T\right) - \sum_{t'=T-n}^{T} \mathcal{B}_t^{-1} \mathcal{B}_t^{-T}, \tag{C.4}$$

where we abuse notation and $\mathcal{B}^{-1}$ stands for $\mathcal{B}_t^{-1} = [\mathbf{0}_t, \beta_N, \beta_{N-1}, \ldots, \beta_1, \mathbf{0}_{T-n-t}]$.

Acknowledgments

We thank Jeff Beck, Sophie Denève, Peter Latham, Máté Lengyel, Liam Paninski, and Peggy Seriès for helpful discussions and for reading versions of the manuscript. This work was supported by the BIBA consortium (Q.H., P.D.), the UCL Medical School M.B./Ph.D. program (Q.H.), the Gatsby Charitable Foundation (P.D.), and NSERC and CIHRC NET program (R.N., R.Z.).

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the light-from-above prior. *Nat. Neurosci.*, *7*(10), 1057–1058.

Anderson, C. H. (1994). Basic elements of biological computational systems. *Int. J. Modern Physics C*, *5*(2), 135–137.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, *3*, 213–251.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci. 7*(5), 358–366.

Barber, M. J., Clark, J. W., & Anderson, C. H. (2003). Neural representation of probabilistic information. *Neural Comp., 15*, 1843–1864.

Barbieri, R., Frank, L. M., Nguyen, D. P., Quirk, M. C., Solo, V., Wilson, M. A., & Brown, E. N. (2004). Dynamic analyzes of information encoding in neural ensembles. *Neural Comp., 16*, 277–307.

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *J. Physiol., 137*, 69–88.

Bethge, M., Rotermund, D., & Pawelzik, K. (2002). Optimal short-term population coding: When Fisher information fails. *Neural Comp., 14*(2), 303–319.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science, 252*(5014), 1854–1857.

Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., & Wilson, M. A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *J. Neurosci., 18*(18), 7411–7425.

Brunel, N., & Nadal, J.-P. (1998). Mutual information, Fisher information, and population coding. *Neural Comp., 10*, 1731–1757.

Chance, F. S., Abbott, L. F., & Reyes, A. D. (2002). Gain modulation from background synaptic input. *Neuron, 35*(4), 773–782.

DeCharms, C., & Zador, A. (2000). Neural representation and the cortical code. *Annu. Rev. Neurosci., 23*, 613–647.

Deneve, S., Latham, P. E., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nat. Neurosci., 4*(8), 826–831.

Dong, D. W., & Atick, J. J. (1995). Statistic of natural time-varying images. *Network: Computation in Neural Systems, 6*, 345–358.

Dow, M. (2003). Explicit inverses of Toeplitz and associated matrices. *Austr. New Zealand Industr. Appl. Math. J. E, 44*, E185–E215.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433.

Gao, Y., Black, M. J., Bienenstock, E., Shoham, S., & Donoghue, J. P. (2002). Probabilistic inference of hand motion from neural activity in motor cortex. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14*. Cambridge, MA: MIT Press.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1983). Neuronal population coding of movement direction. *Science, 233*(4771), 1416–1419.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298*(5598), 1627–1630.

Hinton, G. E. (1999). Products of experts. In *Ninth International Conference on Artificial Neural Networks, (ICANN 9)* (Vol. 1, pp. 1–6). Piscataway, NJ: IEEE.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science, 268*(5214), 1158–1161.

Johansson, R. S., & Birznieks, I. (2004). First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nat. Neurosci., 7*, 170–177.

Kemere, C., Santhaman, G., Yu, B. M., Ryu, S., Meng, T., & Shenoy, K. V. (2004). Model-based decoding of reaching movements for prosthetic systems. In *Proceedings of the 26th Annual International Conference of the IEEE EMBS* (pp. 4524–4528). Piscataway, NJ: IEEE Press.

Knight, B. W. (1972). Dynamics of encoding in a population of neurons. *J. Gen. Physiol., 59*, 734–766.

Körding, K. P., Ku, S. P., & Wolpert, D. M. (2004). Bayesian integration in force estimation. *J. Neurophysiol., 92*(5), 3161–3165.

Körding, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), 244–247.

Lever, C., Wills, T., Cacucci, F., Burgess, N., & O'Keefe, J. (2002). Long-term plasticity in the hippocampal place cell representation of environmental geometry. *Nature, 426*, 90–94.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., & Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature, 411*, 698–701.

Paninski, L. (2003). Convergence properties of three spike-triggered analysis techniques. *Network, 14*(3), 437–464.

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biol. Cybern., 58*(1), 35–49.

Poirazi, P., Brannon, T., & Mel, B. W. (2003a). Arithmetic of subthreshold synaptic summation in a model CA1 pyramidal cell. *Neuron, 37*, 977–987.

Poirazi, P., Brannon, T., & Mel, B. W. (2003b). Pyramidal neuron as 2-layer neural network. *Neuron, 37*, 989–999.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annu. Rev. Neurosci., 26*, 318–410.

Pouget, A., Zhang, K., Deneve, S., & Latham, P. E. (1998). Statistically efficient estimation using population coding. *Neural Comp., 10*(2), 373–401.

Rao, R. P. N., Olshausen, B. A., & Lewicki, M. S. (Eds.) (2002). *Probabilistic models of the brain: Perception and neural function.* Cambridge, MA: MIT Press.

Reinagel, P., & Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *J. Neurosci., 20*(14), 5392–5400.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Sahani, M., & Dayan, P. (2003). Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Comp., 15*, 2255–2279.

Salinas, E., & Abbott, L. F. (1996). A model of multiplicative neural responses in parietal cortex. *Proc. Natl. Acad. Sci. USA, 93*, 11956–11961.

Schwartz, A. B. (1994). Direct cortical representation of drawing. *Science, 265*, 540–543.

Seung, S. H., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proc. Natl. Acad. Sci. USA, 90*(22), 10749–10753.

Silberberg, G., Bethge, M., Markram, H., Pawelzik, K., & Tsodyks, M. (2004). Dynamics of population rate codes in ensembles of neocortical neurons. *J. Neurophysiol., 91*(2), 704–709.

Smith, A. C., & Brown, E. N. (2003). Estimating a state-space model from point process observations. *Neural Comp., 15*, 965–991.

Snippe, H. P., & Koenderinck, J. J. (1992). Discrimination thresholds for channel-coded systems. *Bio. Cybern., 66*, 543–551.

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B Biol. Sci., 216*(1253), 427–459.

Twum-Danso, N., & Brockett, R. (2001). Trajectory estimation from place cell data. *Neural Networks, 14*, 835–844.

Van Rullen, R., & Thorpe, S. J. (2001). Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. *Neural Comp., 13*(6), 1255–1283.

Wang, X.-J., Liu, Y., Sanchez-Vives, M. V., & McCormick, D. A. (2003). Adaptation and temporal decorrelation by single neurons in the primary visual cortex. *J. Neurophysiol., 89*(6), 3279–3293.

Weiss, Y., & Adelson, E. H. (1998). *Slow and smooth: A Bayesian theory for the combination of local motion signals in human vision* (A. I. Memo 1624). Cambridge, MA: MIT, Artificial Intelligence Laboratory.

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science, 261*(5124), 1055–1058.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Comp., 10*(2), 403–430.

Zemel, R. S., Huys, Q. J. M., Natarajan, R., & Dayan, P. (2005). Probabilistic computation in spiking populations. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 1609–1616). Cambridge, MA: MIT Press.

Zhang, K., Ginzburg, I., McNaughton, B. L., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *J. Neurophysiol., 79*, 1017–1044.

Zhang, K., & Sejnowski, T. J. (1999). Neuronal tuning: To sharpen or to broaden. *Neural Comp., 11*(1), 75–84.

Received October 17, 2005; accepted June 17, 2006.