Published in final edited form as:

J Comput Neurosci. 2011 February ; 30(1): 181–200. doi:10.1007/s10827-010-0248-1.

Automating the design of informative sequences of sensory stimuli

Jeremy Lewi,
Georgia Institute of Technology, Atlanta, GA 30332, USA

David M. Schneider,
Columbia University, New York, NY 10027-6902, USA

Sarah M. N. Woolley, and
Columbia University, New York, NY 10027-6902, USA

Liam Paninski
Columbia University, New York, NY 10027-6902, USA

Abstract

Adaptive stimulus design methods can potentially improve the efficiency of sensory neurophysiology experiments significantly; however, designing optimal stimulus sequences in real time remains a serious technical challenge. Here we describe two approximate methods for generating informative stimulus sequences: the first approach provides a fast method for scoring the informativeness of a batch of specific potential stimulus sequences, while the second method attempts to compute an optimal stimulus distribution from which the experimenter may easily sample. We apply these methods to single-neuron spike train data recorded from the auditory midbrain of zebra finches, and demonstrate that the resulting stimulus sequences do in fact provide more information about neuronal tuning in a shorter amount of time than do more standard experimental designs.

Keywords

Information theory; Generalized linear model; Birdsong; Active learning; Optimal experimental design

1 Introduction

Adaptive stimulus optimization methods hold a great deal of promise for improving the efficiency of sensory neurophysiology experiments (Tzanakou et al. 1979; Mackay 1992; deCharms et al. 1998; Anderson and Micheli-Tzanakou 1998; Foldiak 2001; Machens 2002; Edin et al. 2004; Machens et al. 2005; O'Connor et al. 2005; Chen et al. 2005; Paninski 2005; Benda et al. 2007; Yamane et al. 2008). We previously developed methods for efficiently computing the stimulus that will provide the most information about a neuron's tuning properties (Lewi et al. 2007, 2009). These methods were tested in a variety of simulated examples, and demonstrated the potential to speed convergence in our parameter estimates (and therefore reduce the total experimental time needed to characterize neural tuning) by an order of magnitude.

liam@stat.columbia.edu URL: www.stat.columbia.edu/~liam.

However, we face two major problems when applying these methods directly in real experimental settings. First, often stimuli need to be presented in “batches,” instead of one stimulus frame at a time. In auditory experiments, for example, we typically do not present a single 10 ms “frame” of an auditory stimulus; instead, we usually present longer stimuli (e.g., bird song, or other naturalistic sequences) with strong temporal correlations, since auditory neurons integrate information over longer timescales to detect acoustic features. Thus, we need a method to optimize temporally-coherent *sequences* of stimuli, instead of choosing the best stimulus on a frame-by-frame basis, since the latter approach will typically lead to choppy, temporally-incoherent, highly non-naturalistic stimuli.

Second, the methods described in Lewi et al. (2007, 2009) require the development of highly-optimized, real-time software; in particular, for the real-time adaptive stimulus design approach to be feasible, the stimulus generation and spike detection software necessarily need to be tightly integrated. This may be quite challenging, depending on the recording setup available in a given sensory neurophysiology lab. Thus, it would be very helpful to relax our goals somewhat: instead of computing the optimal stimulus one presentation frame at a time, we would like to develop related methods to compute a highly-informative “batch” of stimuli that can be presented over the next (possibly large) set of presentation frames. This greatly reduces the requirement for real-time integration of stimulus generation and spike detection, since the necessary processing can be done over large batches of presentation frames instead of single frames (which will typically require processing on a time scale of milliseconds).

In this paper we present methods for solving both of these problems and demonstrate their applicability to single-neuron spike train data recorded from the auditory midbrain of songbirds.

2 Overview and review

Our setting is as follows. We are recording from a neuron which responds to a vector input s_t presented at time t according to some conditional distribution $p(r_t|\theta, s_t)$, where r_t is the response at time t and θ denotes the vector of parameters that control the neuron’s response properties. As in Paninski et al. (2007), Lewi et al. (2009), we will focus our attention on a generalized linear model (GLM) for the neural responses:

$$r_t|\theta, s_t \sim \text{Poiss}\,[\,f(\theta^T s_t)dt], \tag{1}$$

with the rectifying function $f(.)$ chosen to be convex and log-concave; as discussed in more detail in Paninski (2004), these constraints ensure that the loglikelihood in this model will be a concave function of the parameter θ. We will describe an example implementation of the GLM in much more detail in Section 3.1.1 below.

Given the t previously-observed input-response pairs $(s_{1:t}, r_{1:t})$, our knowledge about θ is summarized by the posterior distribution $p(\theta|s_{1:t}, r_{1:t})$. Our main goal when choosing the next sequence of b inputs $s_{t+1:t+b}$ is to reduce the entropy (uncertainty) of this conditional distribution as much as possible; more precisely, we want to choose $s_{t+1:t+b}$ to optimize the conditional mutual information (Cover and Thomas 1991) between the parameter θ and the response sequence $r_{t+1:t+b}$ given the input sequence $s_{t+1:t+b}$,

$$I\,(\theta;r_{t+1:t+b}\,|\{s_{1:t+b}, r_{1:t}\}) = H[\,p(\theta|s_{1:t}, r_{1:t})] - H[\,p(\theta\Big|s_{1:t+b}, r_{1:t+b})], \tag{2}$$

where $H[p]$ denotes the entropy of the distribution p. The conditional mutual information is the difference between our parameter uncertainty at time t, $H[p(\theta|s_{1:t}, r_{1:t})]$, and our expected uncertainty at time $t + b$, $H[p(\theta|s_{t+1:t+b}, r_{t+1:t+b})]$, and therefore this quantity measures the amount of information we expect an experiment to provide about θ.

As discussed at length in Lewi et al. (2009), it is quite difficult to compute and optimize this mutual information exactly. However, good approximations are available. In particular, we approximate the posterior distribution $p(\theta|s_{1:t}, r_{1:t})$ as a Gaussian density with respect to the parameter θ,

$$p(\theta|s_{1:t}, r_{1:t}) \approx \mathcal{N}_{\mu_t,C_t}(\theta), \tag{3}$$

where μ_t and C_t denote the posterior mean and covariance of θ given $(s_{1:t}, r_{1:t})$; intuitively, μ_t represents our “best guess” about θ given the data observed in the first t responses, and C_t summarizes our posterior uncertainty about θ. This Gaussian approximation turns out to be asymptotically precise in the large-sample limit $t \to \infty$ (Paninski 2005), and is also fairly accurate even for finite t for the log-concave neural response models we consider here (Paninski et al. 2007; Lewi et al. 2009). Using this Gaussian approximation, we can simplify the posterior entropy:

$$I(\theta; r_{t+1:t+b} \,|s_{1:t+b}, r_{1:t}) \approx H[\,p(\theta\,|s_{1:t}, r_{1:t})] + \frac{1}{2} E_\theta E_{r_{t+1:t+b}\,|s_{1:t+b}, r_{1:t}, \theta} \log \left| C_t^{-1} + \sum_{i=1}^{b} J_{\mathrm{obs}}(r_{t+i}, s_{t+i}) \right| + \mathrm{const.}, \tag{4}$$

where $E_x f(x)$ denotes the expectation of $f(x)$ under the distribution of the random variable x. Since the first term on the right hand side is constant with respect to the input sequence $s_{t+1:t+b}$, we will focus our attention on the second term. We have already discussed the posterior covariance C_t; $J_{\mathrm{obs}}(r_{t+i}, s_{t+i})$ denotes the observed Fisher information matrix about θ in the response r_{t+i}, given the input s_{t+i}:

$$J_{\mathrm{obs}}(r_{t+i}, s_{t+i}) = -\nabla\nabla_\theta \log p(r_{t+i}|\theta, s_{t+i}), \tag{5}$$

where $\nabla\nabla_\theta$ denotes the Hessian matrix with respect to θ.

Now, with these preliminaries out of the way, we consider three different cases:

1. The single-stimulus case, $b = 1$;
2. The finite-sequence case, $1 < b < \infty$;
3. The long-sequence limit, $b \to \infty$.

Note that case 1, $b = 1$, which was considered in depth in Lewi et al. (2009), is essentially a special case of case 2. We will discuss case 2 in Section 3 and case 3 in Section 4 below. In neither case are we able to solve the information-maximization problem exactly with the full desired generality. Instead, for case 2, we develop a useful lower bound for the information in a sequence of stimuli which is easier to optimize than the original expression. This lower bound turns out to be closely related to the $b = 1$ case. We will see that optimizing this lower bound leads to significantly more efficient experiments. Similarly, we will show that we can simplify the problem by letting $b \to \infty$. Unfortunately, even in this limiting case it appears that optimizing the information is in general still quite difficult. Therefore we focus our attention on a useful special case; again, we find that the resulting optimized experimental designs are more efficient. Before moving on to the details of these latter two cases, we briefly review our approach to case 1.

2.1 Case 1: the single-frame case, *b* = 1

The special structure of the GLM (Eq. (1)) allows us to compute the Gaussian information (Eq. (4)) fairly explictly. In particular, since the firing rate in the GLM at time $t + 1$ depends only on the one-dimensional projection $\rho_{t+1} = \theta^T s_{t+1}$ of the input onto the parameter vector θ, the Fisher information matrix $J_{\text{obs}}(r_{t+1}, s_{t+1})$ is guaranteed to have rank one. Thus, by applying the Woodbury matrix lemma to $C_t^{-1}+J_{\text{obs}}(r_{t+1}, s_{t+1})$ we can derive an expression for the mutual information which depends on just two scalar variables, $\mu_\rho=s_{t+1}^T\mu_t$ and $\sigma_\rho^2=s_{t+1}^T C_t s_{t+1}$:

$$I(\theta;r_{t+1}|s_{1:t+1}, r_{1:t}) \approx \frac{1}{2}E_{\theta|\mu_t,C_t}\ E_{r_{t+1}|s_{t+1},\theta} \log\left(1+D(r_{t+1},\rho_{t+1})\sigma_\rho^2\right)+\text{const.} \tag{6}$$

$$=\frac{1}{2}E_{\rho_{t+1}|\mu_\rho,\sigma_\rho^2}\ E_{r_{t+1}|\rho_{t+1}} \log\left(1+D(r_{t+1},\rho_{t+1})\sigma_\rho^2\right)+\text{const.}, \tag{7}$$

where ρ_{t+1} is a linear projection of the Gaussian variable θ and is therefore itself Gaussian with mean μ_ρ and variance σ_ρ^2, and we have abbreviated the Fisher information along the projection ρ_t,

$$D(r_t,\rho_t)=-\left.\frac{\partial^2 \log p(r_t|\rho)}{\partial\rho^2}\right|_{\rho=\rho_t}. \tag{8}$$

In the special case that $f(.) = \exp(.)$ (in the GLM literature, this is referred to as the “canonical” link function for Poisson responses, so we will refer to this as the “canonical Poisson” case), the Fisher information does not depend on the observed response r_{t+1}, and we can compute the above quantities directly to obtain

$$I(\theta;r_{t+1}|s_{1:t+1}, r_{1:t})=\frac{1}{2}E_{\rho_{t+1}|\mu_\rho,\sigma_\rho^2} \log\left(1+\exp(\rho_{t+1})\sigma_\rho^2\right)+\text{const.}, \tag{9}$$

which can be further reduced and computed analytically, via standard Gaussian formulae, when the approximation $\log(1 + x) \approx x$ (for x small) is valid for $x=\exp(\rho_{t+1})\sigma_\rho^2$. Given this relatively simple formula for the stimulus informativeness, we can choose a good stimulus, present this stimulus to the neuron and record the response r_t, update our approximate mean μ_t and covariance C_t using the online maximum a posteriori methods described in Lewi et al. (2009), and then choose a new stimulus according to our now-updated objective function, in a closed loop. Again, see Lewi et al. (2009) for further details.

3 Case 2: constructing a tractable lower bound for the informativeness of stimulus sequences of finite length *b*

Now we turn to the question of computing the second term on the right-hand side of Eq. (4) for *b* > 1. We have not been able to develop efficient algorithms for computing this term directly. However, we can easily derive a tractable lower bound on this term: we use Jensen's inequality applied to the concave log-determinant function (Cover and Thomas 1991) to move the summation over stimuli outside the determinant:

$$\log\left|C_t^{-1}+\sum_{i=1}^{b}J_{\text{obs}}(r_{t+i},\, s_{t+i})\right|=\log\left|C_t^{-1}+\sum_{i=1}^{b}D(r_{t+i},\rho_{t+i})s_{t+i}s_{t+i}^T\right| \tag{10}$$

$$=\log\left|\sum_{i=1}^{b}\frac{1}{b}C_t^{-1}+\frac{b}{b}D(r_{t+i},\rho_{t+i})s_{t+i}s_{t+i}^T\right| \tag{11}$$

$$\geq\sum_{i=1}^{b}\frac{1}{b}\log|C_t^{-1}+b\,D\,(r_{t+i},\rho_{t+i})s_{t+i}s_{t+i}^T| \tag{12}$$

$$=\sum_{i=1}^{b}\frac{1}{b}\left(\log|C_t^{-1}|+\log(1+b\,D\,(r_{t+i},\,\rho_{t+i})s_{t+i}^TC_ts_{t+i})\right) \tag{13}$$

$$=\log|C_t^{-1}|+\frac{1}{b}\sum_{i=1}^{b}\log\left(1+b\,D(r_{t+i},\,\rho_{t+i})s_{t+i}^TC_ts_{t+i})\right) \tag{14}$$

$$=\log|C_t^{-1}|+\frac{1}{b}\sum_{i=1}^{b}\log\,(1+b\,D(r_{t+i},\,\rho_{t+i})\sigma_{\rho,t+i}^2)), \tag{15}$$

where we have applied the Woodbury lemma and in the last line defined

$$\sigma_{\rho,t+i}^2=s_{t+i}^TC_ts_{t+i}, \tag{16}$$

with the projected Fisher information $D(r_{t+i}, \rho_{t+i})$ defined as in the last section. When *b* = 1 this inequality is an equality and is exactly the same expression we discussed above (c.f. the term inside the expectation in Eq. (7); the $\log|C_t^{-1}|$ term is cancelled when we subtract this conditional entropy from the prior entropy to form the mutual information). When *b* > 1 this expression may be plugged into the expectation in Eq. (4) to obtain a sum of terms which are essentially similar to the objective function in the *b* = 1 case. As a result, we may use the methods in Lewi et al. (2009) to efficiently compute the terms in the summation. As before, the key insight is that for the GLM under consideration here, each term in the resulting summation is simply a function of the two scalar variables $(\mu_{\rho,t+i}, \sigma^2_{\rho,t+i})$, so to efficiently evaluate the bound we can simply precompute this function on some suitable range of $(\mu_{\rho,t+i}, \sigma^2_{\rho,t+i})$; alternatively, in the canonical Poisson case, as discussed above, these formulas may be simplified even further and in some cases computed analytically.[1] Again, see Lewi et al. (2009) for full details.

We can use this result to optimize our experiments in a straightforward manner. Suppose we are given *n* stimulus sequences, each of length *b*, and we want to choose the best sequence from among these *n* examples. Using the above expression, we compute the lower bound evaluated for each of these sequences, and then we select the sequence which maximizes this lower bound,[2] present the sequence, and update our posterior parameters μ_t and C_t based on the observed responses, as in Lewi et al. (2009). See Fig. 1 for an illustration, and the following section for further details. This lower bound can also be applied in more elaborate receding-horizon settings (Kwon and Han 2005), where we update our chosen stimulus sequence every time step (instead of every *b* time steps), but we will stick to the simpler case (update every *b* steps) below for clarity.

3.1 Application to zebra finch auditory data

One key question left unanswered by our previous work (Lewi et al. 2009) was how well our techniques will work with real data. We can use the methods presented above to begin to address this question. The idea is to take the set of stimulus-response pairs obtained during an actual experiment and check whether the infomax approach leads to a more informative ordering of the data. A more informative ordering of the data is one for which our parameter estimates converge faster (as a function of the number of stimulus sequences presented) to the final estimate of the model trained using all the data. Here we will not choose stimuli arbitrarily, from the space of all possible auditory stimuli, but rather we will restrict our attention to stimuli which were actually presented, i.e., to simple reorderings of the trials, because we want to use the actual neural response observed with each presented stimulus.

We begin by describing the physiological experiments and data collected. Experimental details have been previously described in Woolley and Casseday (2004, 2005). We played auditory stimuli to adult male zebra finches while recording the responses of neurons in the mesencephalicus lateralis pars dorsalis (MLd) using extracellular electrodes; for further details, see Woolley et al. (2006). MLd is a midbrain auditory nucleus which is the avian homolog of the mammalian inferior colliculus. The set of stimuli consisted of samples of the songs of 20 different adult zebra finches and 10 modulation-limited noise stimuli (ml-noise), which is described below. Each stimulus had a duration of approximately 2 s and was repeated 10 different times to the bird in a random order. Before and after each stimulus was played there was a randomized period of silence of 1,200–1,600 ms. This period of silence allowed the neuron to return to its resting state before the next stimulus was played, thereby minimizing the effects of adaptation.

Examples of each stimulus type and an accompanying raster plot for one neuron are shown in Fig. 2. For comparison, ml-noise stimuli were also presented to the birds; this is a form of

[1] The careful reader will have noticed that Eq. (15) depends on the future responses r_{t+i}, through the projected Fisher information $D(r_{t+i}, \rho_{t+i})$. However, recall that we are taking the expectation of Eq. (15), by plugging into Eq. (4), and the expectation of each term in Eq. (15) may be computed directly using the methods described above, without violating any causality constraints.

[2] In Lewi et al. (2009), we discuss analytical methods for computing the optimal stimuli over sets of infinite cardinality, e.g., ellipsoidal sets of bounded squared norm in stimulus space. These methods relied strongly on the one-dimensional nature of the projected stimulus ρ_t and are unfortunately not applicable in the $b > 1$ case, where the projected stimulus is *b*-dimensional instead of just one-dimensional.

broadband noise which is designed to have the same power and maximum spectral and temporal modulations that occur in the songs of adult zebra finches (Hsu et al. 2004; Woolley et al. 2006) Thus, ml-noise can be used to contrast the responses to conspecific vocalizations compared to noise stimuli with a similar spectral range.

3.1.1 Fitting the GLM to birdsong data—

In this section we describe our efforts to estimate the receptive field of MLd neurons using GLM methods. We used the canonical Poisson model (*f*(.) = exp(.)) for simplicity; in this case, as discussed above, the expected Fisher information is exponential and independent of the response r_t. This property makes the computations required to optimize the design much more tractable, and the results in Paninski (2004), Lewi et al. (2009) indicate that the results should be fairly robust with respect to the choice of the rectifying nonlinearity *f*(.) here. Finally, we have found in Calabrese et al. (2010) that the canonical Poisson model provides good predictions of spike train responses in this system.

The canonical Poisson model assigns a probability to the number of spikes we expect to observe in some window of time, as a function of the stimulus, past responses, and background firing rate. The log-likelihood of the response at time *t* is

$$\begin{aligned} \log p(r_t|s_t,\theta) &= -\log r_t! + r_t s_t^T\theta - \exp(s_t^T\theta)dt \\ s_t^T &= \left\{ \vec{x}_{t-t_k}^{\,T}, \ldots, \vec{x}_t^{\,T}, r_{t-t_a}, \ldots, r_{t-1}, 1 \right\}. \end{aligned} \tag{17}$$

The response, r_t, is the number of spikes observed in a single trial in some small time window of length *dt*. The input, s_t, consists of the most recent $t_k + 1$ stimuli, $\{\vec{x}_{t-t_k}, \ldots, \vec{x}_t\}$, the most recent t_a responses of the neuron, and a constant term that allows us to include a bias which sets the background firing rate of the neuron. The stimulus $\vec{x}_t$ is defined more explicitly below.

For auditory neurons, the receptive field of the neuron is typically represented in the spectrotemporal domain because the early auditory system is known to perform a frequency decomposition. Furthermore, transforming the input into the power spectral domain is a nonlinear transformation which generally improves the accuracy of the linear model for auditory data (Gill et al. 2006). The spectro-temporal receptive field (STRF) of the neuron, $\theta_x(\tau, \omega)$, is a two-dimensional filter which relates the firing rate at time *t* to the amount of energy at frequency ω and time $t - \tau$ in the stimulus. The subscript on θ is used to distinguish the elements of θ which measure the dependence of the response on the stimulus, spike-history, and bias terms respectively.

The stimuli and responses were computed from the experimental data by dividing the recordings into time bins of 2.5 ms. The time bin was small enough that more than one spike was almost never observed in any bins. To construct the corresponding stimulus, $\vec{x}_t$, we computed the power spectrum over a small interval of time centered on *t* (Gill et al. 2006). The power was computed for frequencies in the range 300–8,000 Hz, in intervals of approximately 100 Hz; previous work has suggested that this frequency spacing is suitable for computing the STRF in this context (Singh and Theunissen 2003; Gill et al. 2006).

We initially estimated a GLM with a STRF that had a duration of 50 ms ($t_k + 1 = 20$ time bins) and had $t_a = 8$ spike history terms as well as a bias term, for a total of 1,589 unknown parameters.[3] The durations of the STRF and spike history dependence were chosen based on

[3] 79 × 20 coefficients of the STRF +8 spike history coefficients +1 bias term = 1,589 unknown parameters.

prior knowledge that these durations were long enough to capture most of the salient features of the STRF and spike history dependence (Woolley et al. 2006). Examples of the estimated STRF, spikehistory, and bias terms are shown in Fig. 3. The STRFs have similar temporal and frequency tuning to the STRFs trained on ml-noise using reverse-correlation methods presented in previous work (Woolley et al. 2006); see Calabrese et al. (2010) for further details. Also plotted is the estimated spike history filter. The largest coefficients are negative and occur for delays close to zero; thus the effect of the spike-history terms is to inhibit spiking immediately after the neuron fires (i.e., a relative refractory effect). The bias terms were also significantly negative, corresponding to low background firing rates; for the neurons shown in Fig. 3 the background firing rates are ≈ 3–5 Hz.

To produce the STRFs shown in Fig. 3 we incorporated a regularizer in the fitting procedure to remove high-frequency noise (Theunissen et al. 2000, 2001; Machens et al. 2003; Smyth et al. 2003; Theunissen et al. 2004). Our preliminary results (data not shown) indicated that removing this regularization led to overfitting, which is unavoidable given the dimensionality of the STRF and the size of the dataset: the 30 stimuli, each ≈ 2 s in duration, translate into ≈ 20,000 distinct inputs when using a 50 ms STRF. Furthermore, most of these inputs are highly correlated due to the temporal structure of birdsong and the fact that we generate the inputs $\vec{s}_t$ by sliding a window over the spectrogram. To deal with this issue efficiently, we represent the STRF in the frequency domain and incorporate a prior on the amplitudes of the frequency coefficients; fitting the STRF by maximum penalized likelihood (instead of maximum likelihood) biases the STRF towards smoother features when the available data are limited. This regularization procedure is discussed in detail in Appendix A.

3.1.2 Quantifying the improvement due to infomax stimulus design: Our goal now is to measure how many fewer trials an infomax design would require to accurately estimate these neurons' receptive fields. We simulated a closed-loop infomax experiment by iterating the following steps. First, we initialized our Gaussian prior given no observed stimulus-response data; this Gaussian corresponds to the exponent of the quadratic low-pass regularizing penalty term described in Appendix A. Next, we considered all sequences of b consecutive stimuli and computed the lower bound on the informativeness of each sequence, given all of the previously-presented stimuli and corresponding responses, using Eq. (15). In these experiments we chose $b = 20$, so that each subsequence had a duration of roughly 50 ms, the approximate length of most relevant auditory features in the zebra finch song. We then selected the stimulus sequence which maximized this lower bound on the informativeness. To update our posterior mean μ_t and covariance matrix $\boldsymbol{C}_t$ we used the actual responses recorded given this sequence of inputs; using this data $(s_{t+1:t+b}, r_{t+1:t+b})$, we updated the posterior (Eq. (3)) via the methods described in Lewi et al. (2009). We then recomputed the informativeness of the remaining stimulus segments (excluding all previously chosen inputs), using the new, updated posterior mean μ_{t+b} and covariance C_{t+b}. We repeated these steps until all of the stimulus-response segments were processed. Below we will compare the performance of this procedure against a "shuffled" procedure in which stimulus-response pairs were drawn without replacement from the true data set at each time step; i.e., the same data are presented, just in a shuffled order.

We used 28 different training stimulus files to generate input/response pairs (s_t, r_t). Each stimulus was presented 10 times to the bird. The total number of spikes in the training sets ranged from roughly 300–4,000 per neuron. For each neuron, we trained a single model on all the training data (including both ml-noise and bird song stimuli), as opposed to fitting separate GLMs on the bird song and ml-noise training sets, so that the training set would span the input space as much as possible. For each neuron, the responses to one bird song and one ml-noise stimuli were held out as test stimuli to evaluate the fitted models. The expected log-likelihood (averaged over our posterior uncertainty about the model parameters, $p(\theta|\mu_t, C_t)$) provides a measure of how well the model predicts the responses to novel stimuli:

$$Q(t)=\frac{1}{T}\sum_{i=1}^{T} E_{\theta|\vec{\mu}_t,C_t} \log p(r_i|s_i,\theta), \tag{18}$$

where the summation over i is over each stimulus-response pair in the test set and T is the number of stimuli in the test set. Note that if the neuron's response is perfectly predictable, and the GLM achieves a perfect level of prediction, then $Q(t) = 0$ (since in this case the conditional entropy of the response r_t will be zero). In practice, $Q(t)$ will converge to some value less than 0, since neural responses are noisy and the GLM is an imperfect model. The results of this analysis are shown in Fig. 4 for two different neurons. In each case, the infomax design reduces the number of trials needed to accurately estimate the neuron's receptive field.

To quantify the improvement in efficiency over a larger neural population, we exponentiated the expected log-likeliood $Q(t)$ to obtain a quantity bounded between 0 and 1. We then examined the ratio of the number of trials needed to achieve a given level of accuracy (as measured by $\exp(Q(t))$, normalized by its asymptotic maximum value computed using the full dataset) for each neuron, comparing the infomax approach to the original (random) order in which the stimuli were presented. This ratio (measured in units of % improvement) is summarized in Fig. 5(b), which plots the average and standard deviation of the speedup over all of the neurons in our dataset. The results show that on average the shuffled design required three times as many trials to produce a model that fit the data as well as a model trained using the information maximizing design. Table 1 lists the median as well as maximum and minimum speedup for all 11 neurons examined here.

It is also worth noting that we expect the results in Fig. 5(b) to underestimate the potential improvement in actual experiments, because in our simulations the infomax design could only pick inputs which were actually presented to the birds. We know from Lewi et al. (2009) that restricting the input to a poorly chosen set of stimuli can dramatically reduce the relative advantage of the infomax design. Therefore, we would expect an infomax design which is allowed to choose stimuli from a much larger stimulus space than the relatively small set of songs and ml-noise stimuli considered in these experiments to perform significantly better.

4 Case 3: constructing informative random sequences in the long-sequence limit, $b \to \infty$

In the previous section we chose informative stimuli with complex temporal features by optimizing sequences of b inputs. However, even for values of b for which the stimulus segment is reoptimized just once every 50 ms or so, a good deal of real-time computation is required. Thus it is natural to ask whether the computations may be simplified if we let b become even larger. In particular, can we form some convenient approximation of our objective function as $b \to \infty$? This would correspond to a situation in which we re-optimize our stimuli only occasionally during an experiment, and would greatly reduce the required computational overhead. In addition, once we pick an arbitrarily long sequence of inputs, we can continue the experiment indefinitely. Thus, while we are computing an updated, optimized sequence using the most recent data, we can continue our experiment using the previously chosen sequence, further reducing any real-time computational requirements.

How can we evaluate this $b \to \infty$ limit? If we look at the form of Eq. (4) once again, we see that the log-determinant term involves the inverse prior covariance C_t^{-1} plus a sum of Fisher information matrices J_{obs} over b responses. Thus we might hope that we can neglect the prior term C_t^{-1} and apply the law of large numbers to more easily evaluate the summed information matrices J_{obs} in the large-b limit. This idea is well-known in the experimental design literature (Fedorov 1972); see also Paninski (2005) for some concrete results related to the setting discussed here.

Unfortunately, the resulting limiting objective function (Eq. (19) below) is still not very tractable. Thus we relax our problem further, in three steps:

1. We restrict our consideration to stimulus sequences with a Gaussian distribution. Thus, instead of searching over infinitely long stimulus sequences, we just need to optimize our objective function over the much more tractable space of mean vectors and covariance functions.
2. We restrict our attention to the canonical Poisson GLM, which allows us to compute the objective function explicitly.
3. We construct a lower bound on this objective function to arrive at a function we can optimize tractably.

4.1 Deriving a tractable objective function in the large-*b* limit—We begin by showing that as $b \to \infty$ maximizing the mutual information is equivalent to maximizing the average information per trial. If we think of the mutual information as measuring the total information acquired from b trials then the average information per trial (or the information rate) is just the mutual information normalized by the number of trials. More concretely, let's rewrite Eq. (4) slightly:

$$E_\theta\, E_{r_{t+1:t+b}\,|s_{t+1:t+b},\theta} \log \left| C_t^{-1} + \sum_{i=t+1}^{t+b} J_{\mathrm{obs}}(r_i, s_i) \right| = E_\theta\, E_{r_{t+1:t+b}\,|s_{t+1:t+b},\theta} \log \left| \frac{C_t^{-1}}{b} + \frac{1}{b} \sum_{i=t+1}^{t+b} J_{\mathrm{obs}}(r_i, s_i) \right| + \dim(\theta) \log b.$$

The $\dim(\theta) \log b$ term does not depend on the input sequence, and may therefore be ignored.

If C_t is invertible and the average Fisher information $\frac{1}{b} \sum_{i=t+1}^{t+b} j_{\mathrm{obs}}(r_i, s_i)$ is of full rank,[4] then $(1/b)C_t^{-1}$ will be negligible as $b \to \infty$, and

$$\lim_{b\to\infty} E_\theta\, E_{r_{t+1:t+b}\,|s_{t+1:t+b},\theta} \log \left| \frac{C_t^{-1}}{b} + \frac{1}{b} \sum_{i=t+1}^{t+b} J_{\mathrm{obs}}(r_i, s_i) \right| = \lim_{b\to\infty} E_\theta\, E_{r_{t+1:t+b}\,|s_{t+1:t+b},\theta} \log \left| \frac{1}{b} + \sum_{i=t+1}^{t+b} J_{\mathrm{obs}}(r_i, s_i) \right|. \tag{19}$$

[4] If C_t is low-rank then C_t^{-1} is infinite in some directions, and the derivation will not hold because the contribution of C_t^{-1} will not become negligible as $b \to \infty$. In this case we can simply use a truncated design: i.e., we maximize the information in directions for which our prior uncertainty is not zero. To accomplish this we simply project θ into the lower-dimensional space corresponding to the space spanned by non-zero eigenvectors of C_t. Alternately, in the case that C_t has some very small but positive eigenvalues, it may be possible to approach the full objective function directly, though we have not pursued this direction systematically.

The result is that maximizing the average Fisher information per trial is asymptotically equivalent to maximizing the total information. See e.g. Chaloner and Verdinelli (1995) for background on the resulting "Bayesian D-optimal" design criterion. As discussed in Paninski (2005), this log-determinant of the average Fisher information is a concave function of the input distribution, $p(s_t)$, and therefore in principle we can optimize this function by ascent methods over the convex space of all possible input distributions $p(s_t)$. Unfortunately, since we are considering very high-dimensional inputs s_t here (e.g., bird songs), this optimization over $p(s_t)$ will not be tractable in general, and we have to search for a simpler relaxation of our problem.

4.1.1 Restricting our attention to Gaussian stimulus sequences: The first simplification is to restrict our attention to Gaussian stimulus distributions $p(s_t)$. The advantage here is obvious: Gaussian distributions are easy to sample from, and are specified completely by their mean vector μ_s and covariance matrix C_s. Thus, if s_t is d-dimensional, we can reduce our original infinite-dimensional problem (search over all possible distributions $p(s_t)$) to a much simpler $O(d^2)$-dimensional problem (search over the space of allowable (μ_s, C_s)). So our optimization problem is reduced from maximizing Eq. (19) over all input distributions $p(s_t)$ to

$$\max_{\mu_s,C_s} E_\theta \log \left| E_{s|\mu_s,C_s} E_{r|s,\theta} \frac{1}{b}\sum_{i=1}^{b} J_{\text{obs}}(r_i, s_i) \right|. \tag{20}$$

4.1.2 Specializing to the canonical Poisson case: As emphasized above, computations involving Fisher information simplify dramatically in the canonical Poisson model, since the observed Fisher information does not depend on the response r_t. In fact, the Fisher information has a particularly simple form here:

$$J_{\text{exp}}(s^T\theta)=E_{r|s,\theta}J_{\text{obs}}(r, s)=\exp(s^T\theta)ss^T. \tag{21}$$

Note that Eq. (20) involves a Gaussian integral over the Fisher information; the simple exponential form of J_{exp} allows us to evaluate this expectation analytically here:

$$E_\theta \log \left|E_{s|\mu_s,C_s} E_{r|s,\theta}J_{\text{obs}}(r, s)\right| =E_\theta \log \left|E_{s|\mu_s,C_s} \exp(s^T\theta)ss\right| \tag{22}$$

$$=E_\theta \log \left|\exp\left(\theta^T\mu_s+\frac{1}{2}\theta^T C_s\theta\right) \times \left((\mu_s+C_s\theta)(\mu_s+C_s\theta)^T+C_s\right)\right| \tag{23}$$

$$=E_\theta \left(d\theta^T\mu_s+\frac{d}{2}\theta^T C_s\theta+\log|C_s|+\log(1+(\mu_s+C_s\theta)^T C_s^{-1}(\mu_s+C_s\theta))\right) \tag{24}$$

$$=d\vec{\mu}_t^T\mu_s+\frac{d}{2}Tr(C_s\vec{\mu}_t\vec{\mu}_t^T+C_sC_t)+\log|C_s|+E_\theta \log(1+(\mu_s+C_s\theta)^T C_s^{-1}(\mu_s+C_s\theta)) \tag{25}$$

(Note that we have incorporated the sum over s_i in Eq. (20) into the expectation over s in the first line here.) More generally, for other GLMs, the expected Fisher information $E_{p(s)}$ $J_{\exp}(s^T\theta)$ may be computed in terms of a rank-2 perturbation of the stimulus covariance matrix C_s (see Appendix B). This is still tractable in principle, but we have not explored this direction systematically.

4.1.3 Computing a tractable lower bound: Computing the last term in Eq. (25) is difficult. However, it is easy to see that this term must be nonnegative, since

$$(\mu_s+C_s\theta)^T C_s^{-1}(\mu_s+C_s\theta) \geq 0 \Rightarrow \log(1+(\mu_s+C_s\theta)^T C_s^{-1}(\mu_s+C_s\theta)) \geq 0. \tag{26}$$

Thus, by dropping the log term5 we obtain a rather simple lower bound:

$$E_\theta \log |E_s \exp(s^T\theta)ss| \geq d\vec{\mu}_t^T \mu_s + \frac{d}{2} Tr(C_s\vec{\mu}_t\vec{\mu}_t^T + C_sC_t) + \log |C_s|. \tag{27}$$

Qualitatively, this lower bound leads to a reasonable objective function for optimizing the design. Our goal is to pick inputs which maximize the amount of new information provided by the experiment. The utility of an input is thus a function of 1) the informativeness of the experiment as measured by the Fisher information, which is independent of what we already know, and 2) our posterior covariance, which quantifies what we already know. We can interpret each of the terms in Eq. (27) as reflecting these goals. For example, in the canonical Poisson model the Fisher information increases with $\exp(s_t^T\theta)$. Thus, to increase the Fisher information of the inputs we want to maximize the projection of the inputs on θ. Clearly, maximizing $Tr(C_s\vec{\mu}_t\vec{\mu}_t^T)=\vec{\mu}_t^T C_s\vec{\mu}_t$ and $\vec{\mu}_t^T \mu_s$ entails placing as much stimulus power as we can in the direction of μ_t, which is our best estimate of θ at time t. As a result, the first two terms quantify the extent to which the design picks inputs with large Fisher information.

In contrast, the effect of the log $|C_s|$ term is to whiten the design, since this term is maximized (given a bound on $Tr(C_s)$, i.e., a bound on the total mean square power of C_s) when all the eigenvalues of C_s are equal. Similarly, the $Tr(C_sC_t)$ term is directly related to our prior uncertainty and the Fisher information. The $Tr(C_sC_t)$ term forces us to explore areas of uncertainty: maximizing $Tr(C_sC_t)$ subject to a constraint on $Tr(C_s)$ entails putting all stimulus power along the largest eigenvector of C_t. Thus, maximizing $Tr(C_sC_t)$ favors designs which explore regions of θ space where our uncertainty is high.

All of these terms may be increased trivially by increasing the magnitude of the stimuli. Therefore, we must constrain the stimuli in order to get a well defined optimization problem. A reasonable constraint is the average power of the stimuli,

$$E(s^T s)=Tr(C_s)+\mu_s^T\mu_s. \tag{28}$$

Thus, finally, we arrive at our desired optimization problem:

[5] It is also worth noting that the log term will typically be much smaller than the other terms when the stimulus dimension d_S is large, since the first three terms in Eq. (25) scale linearly with d_S.

$$\max_{(\mu_s,C_s):Tr(C_s)+\mu_s^T\mu_s<m}\left(d\vec{\mu}_t^{\,T}\mu_s+\frac{d}{2}Tr(C_s\vec{\mu}_t\vec{\mu}_t^{\,T}+C_sC_t)+\log|C_s|\right), \tag{29}$$

over the space of semi-positive definite stimulus covariance matrices C_s. It turns out to be fairly straightforward to solve this optimization problem semianalytically; only a one-dimensional numerical search over a Lagrange multiplier is required. See Appendix C for the details.

4.2 Results

We tested our methods using simulated experiments in which we generated synthetic responses from a GLM whose parameters were estimated directly from the data described in Section 3.1.1 above. In particular, for the simulations shown here, we used the STRF and bias term fit to the neuron shown in Fig. 3(a); similar results were observed with parameters estimated from other neurons (data not shown). We chose stimuli either by sampling from an optimized Gaussian process using the methods discussed above, or by sampling i.i.d. stimuli from a Gaussian distribution with mean zero and covariance proportional to the identity matrix; the proportionality constant was chosen so that both Gaussian processes were subject to the same average power constraint. The prior $p(\theta)$ was chosen as discussed in the previous section.

In Fig. 6 we compare the posterior mean estimate of the parameters using the two designs as a function of t. In Fig. 7 we compute the expected error

$$E_{\theta|\mu_t,C_t}||\theta-\theta_o||_2,$$

where the expectation is computed using our posterior on θ and θ_o is the true parameter vector. While the posterior mean parameter estimates μ_t qualitatively seem to converge at a similar rate in Fig. 6, it is clear that the uncertainty in our estimates converges to zero faster when we use the optimized Gaussian stimulus distribution.

In Fig. 8 we plot the observed firing rate as a function of time for the synthetic neuron. This plot shows that the optimized design ends up picking inputs which drive the neuron to fire at a higher rate; recall that for the canonical Poisson the Fisher information increases with the firing rate.

5 Conclusion

In this work we have developed two methods for choosing informative stimulus sequences for use in on-line, adaptive sensory neurophysiology experiments. Our primary goal was to extend the methods introduced in Lewi et al. (2007, 2009), which focused on the problem of choosing a stimulus for which the corresponding response (measured as the spike count in a single short time bin) will provide as much information as possible about the observed neuron’s response properties. The extension pursued here—to the problem of choosing a *sequence* of stimuli whose corresponding sequence of responses will be as informative as possible—turns out to be computationally quite challenging, and so we have taken a more modest approximate approach instead. In Sections 3 and 4 we developed two lower bounds on our original information-theoretic objective function; these bounds provide rather natural extensions of the single-stimulus objective function to the stimulus-sequence case, and are much more computationally tractable than the full stimulus-sequence informativeness. The first method (described in Section 3) leads to efficient scoring of sequences of stimuli, so we can quickly pick the most informative stimulus sequence out of a large batch of candidate sequences. The second method (Section 4) finds a good distribution over sequences, from which we may then draw sample stimuli quite easily. In each case, despite a number of approximations and simplifications to ensure the tractability of the resulting algorithm, the chosen stimulus sequences decreased the error significantly faster than did standard experimental designs when tested on real and simulated birdsong auditory responses. We emphasize that the methods described here are simple enough to be implemented in on-line experiments without extraordinary effort, as compared to the single-stimulus methods discussed in Lewi et al. (2009), which require implementation of rather sophisticated real-time spike train processing and stimulus generation methods.

We close by noting a few attractive directions for future work. First, as emphasized in Section 4.1 and in the Appendix, it should be possible to develop tighter lower bounds and better approximations for the informativeness, perhaps at the expense of some computational tractability. By maximizing better approximations to the original information-theoretic objective function, we would hope to obtain even better performance. Second, it would be very useful to extend these methods to compute informative stimulus sequences in the context of multiple-neuron recordings, which have proven especially powerful in studying the early visual system (Segev et al. 2004; Ohki et al. 2005; Pillow et al. 2008) and which hold great promise in other sensory modalities (Luczak et al. 2007). Third, higher-level neurons often show nonlinear selectivity for specific feature conjunctions, which makes discovering optimal stimuli difficult, but ripe for efficient stimulus optimization methods. While the simple GLM approach we have pursued here is poorly suited for such neurons, it may be possible to adapt the nonlinear methods described in our previous work (Lewi et al. 2008, 2009) to handle these cases. Similarly, central neurons frequently exhibit strong response adaptation, which is often partially stimulus-specific. The method discussed in Section 4 tends to drive sampling toward stimuli that have a strong projection onto the cell’s receptive field, leading to an increasingly strongly-driving and homogeneous subset of stimuli (cf. Fig. 8). Incorporating more profoundly stimulus-dependent adaptation terms into our approach remains an important open challenge. Finally, we are currently pursuing applications to real online experiments, in order to better understand the role of plasticity and spectral filtering in the songbird auditory system.

Acknowledgments

We thank A. Ramirez, C. Rozell, and the anonymous referees for very helpful comments on the manuscript. JL was supported by the Computational Science Graduate Fellowship Program administered by the DOE under contract DE-FG02-97ER25308. DMS is supported by a Gatsby Initiative in Brain Circuitry fellowship, and by NIH Predoctoral Fellowship F31DC010301. SW is supported by the Searle Scholar’s Fund and by NIH grant 1R01DC009810. LP is supported by a McKnight Scholar award and an NSF CAREER award. LP and SW were jointly supported by a Gatsby Initiative in Brain Circuitry pilot grant. Description of some early preliminary results appeared in conference proceedings (Lewi et al. 2008).

Appendix A: Using a frequency representation to smooth the STRF

To represent the STRF in the Fourier domain, we applied the Fourier transform separately to the spectral and temporal dimensions of the STRF. Applying the separable Fourier transform to the STRF is just a linear transformation. This transformation maps the STRF into a coordinate system in which the basis functions are rank one matrices. Each of these matrices is the product of 1-dimensional sine-waves in the spectral and temporal directions of the STRF. Using these basis functions we can write the STRF such that each row and column of the STRF is a linear combination of 1-d sine-waves,

$$\theta(i,j)=\sum_{\alpha=1}^{m_f}\sum_{\beta=1}^{m_t}\gamma^1_{\alpha,\beta}\sin(2\pi\cdot f_{o,f}\cdot\alpha\cdot i)\times\sin(2\pi\cdot f_{o,t}\cdot\beta\cdot j) \tag{30}$$

$$+\sum_{\alpha=1}^{m_f}\sum_{\beta=0}^{m_t}\gamma^2_{\alpha,\beta}\sin(2\pi\cdot f_{o,f}\cdot\alpha\cdot i)\times\cos(2\pi\cdot f_{o,t}\cdot\beta\cdot j) \tag{31}$$

$$+\sum_{\alpha=0}^{m_f}\sum_{\beta=1}^{m_t}\gamma^3_{\alpha,\beta}\cos(2\pi\cdot f_{o,f}\cdot\alpha\cdot i)\times\sin(2\pi\cdot f_{o,t}\cdot\beta\cdot j) \tag{32}$$

$$+\sum_{\alpha=0}^{m_f}\sum_{\beta=0}^{m_t}\gamma^4_{\alpha,\beta}\cos(2\pi\cdot f_{o,f}\cdot\alpha\cdot i)\times\cos(2\pi\cdot f_{o,t}\cdot\beta\cdot j). \tag{33}$$

The functions $\sin(2\pi \cdot f_{o,f} \cdot \alpha \cdot i)$ and $\cos(2\pi \cdot f_{o,f} \cdot \alpha \cdot i)$ determine how each basis function varies across the spectral dimension of the STRF while the functions $\sin(2\pi \cdot f_{o,t} \cdot \beta \cdot j)$ and $\cos(2\pi \cdot f_{o,t} \cdot \beta \cdot j)$ determine how the basis functions vary across time in the STRF. Each pair of sine-waves measures the amount of energy at particular frequencies in the spectral and temporal dimensions. The amplitude of each frequency is determined by the coefficients $\gamma^i_{\alpha,\beta}$. To form an orthogonal basis for the STRF we need to project the STRF onto sinusoids with frequencies

$$\{0, f_{o,f}, 2f_{o,f}, \dots m_f f_{o,f}\}\{0, f_{o,t}, 2f_{o,t}, \dots, m_t\, f_{o,t}\} \tag{34}$$

$$f_{o,f}=\frac{1}{n_f} f_{o,t}=\frac{1}{n_t} \tag{35}$$

$$m_f=\lceil\frac{1}{2f_{o,f}}-1\rceil m_t=\lceil\frac{1}{2f_{o,t}}-1\rceil; \tag{36}$$

$f_{o,f}$ and $f_{o,t}$ are the fundamental frequencies and are set so that one period corresponds to the dimensions of the STRF (n_t and n_f denote the dimensions of the STRF in the time and frequency dimensions, respectively), and m_f and m_t are the largest integers such that $m_f f_{o,f}$ and $m_t f_{o,t}$ are less than the Nyquist frequency. We subtract 1 and take the ceiling to make sure the frequencies of our basis functions are less than the Nyquist frequency. The unknown parameters in this new coordinate system are the amplitudes, $\vec{\gamma}=\{\gamma^1_{\alpha,\beta}, \gamma^2_{\alpha,\beta}, \gamma^3_{\alpha,\beta}, \gamma^4_{\alpha,\beta}\}$. For simplicity, we will continue to refer to the unknown parameters as θ, realizing that the STRF is represented using this new basis. Since this transformation is linear we can continue to apply our methods for fitting theGLMand optimizing the stimuli.

To low pass filter the STRF we can simply force the coefficients of θ corresponding to high frequencies to zero; i.e we pick cutoffs n_{tc} and n_{fc} for the time and spectral directions respectively and set

$$\gamma^{i}_{\alpha,\beta}=0\mathrm{i}f\alpha>n_{f_c}\ \text{or}\ \beta>n_{tc}. \tag{37}$$

Decreasing the cutoff frequencies not only makes the estimated STRFs smoother, it also reduces the dimensionality of the model. Reducing the dimensionality makes it easier to fit the GLM and optimize the stimuli, but the risk is that the lower-dimensional model may be too simple to adequately model auditory neurons. We can mitigate this risk by using a soft cutoff. Rather than force all high-frequency components to zero, we can adjust our prior to reflect our strong belief that high frequencies should have little energy; we simply set the prior mean of these coefficients to zero and decrease their prior variance. If we now estimate the STRF using the maximum of the posterior then the amplitudes at high frequencies will be biased by our prior towards zero. However, given sufficient evidence the posterior mean will yield non-zero estimates for the amplitudes of high frequencies. See Theunissen et al. (2001) for details and David et al. (2007), Calabrese et al. (2010) for further discussion.

We chose to impose a hard cutoff because we wanted to reduce the dimensionality to make online estimation of the model and online optimization of the stimuli more tractable. To pick the cutoff frequencies, we picked a single neuron and estimated the STRF using maximum-likelihood for a variety of cutoff frequencies. We evaluated the quality of each model by computing the log-likelihood of the bird's responses to inputs in a test set. The test set consisted of one bird song and one ml-noise stimulus which were not used to train the models. We chose the cutoff frequencies to be n_{fc} = 10 and n_{tc} = 4 because these values provided good predictive performance for both the bird song and ml-noise while keeping the number of unknown parameters tractable (in this case the STRF has 189 unknown parameters).

Appendix B: Computing the average information for a Gaussian process

In this section we show how the average information per stimulus,

$$E_\theta \log \left|E_s \exp(s^T\theta)ss^T\right|,$$

can be computed when the input distribution is a Gaussian process. For the GLM the expected Fisher information matrix $E_s\ J_{\exp}(s, \theta)$ has a simple 1-dimensional dependence on θ,

$$J_{\exp}(s,\theta)=J_{\exp}(s^T\theta)ss^T \tag{38}$$

$$J_{\exp}(s^T\theta)=-E_r\frac{\partial^2\log p(r|\rho=s^T\theta)}{\partial\rho^2}ss^T \tag{39}$$

$$=J_{\exp}(\rho=s^T\theta)ss^T. \tag{40}$$

This 1-dimensional structure along with the fact that $p(s)$ is Gaussian makes computing the expectations tractable. We start by defining a new coordinate system in which the first axis is aligned with θ. This coordinate system is defined by the orthonormal matrix, $\mathcal{R}_\theta$. The first column of $\mathcal{R}_\theta$ is $\frac{\theta}{\|\theta\|_2}$ and the remaining columns are a suitable set of orthonormal vectors. We can thus define the transformation of s and θ into this new coordinate system,

$$\theta' = \mathcal{R}_\theta^T \theta \tag{41}$$

$$\vec{w} = \mathcal{R}_\theta^T s. \tag{42}$$

This coordinate system has the convenient properties

$$\theta_i' = 0 \forall i \neq 1 \tag{43}$$

$$\Rightarrow \vec{w}^T \theta' = w_1 \theta_1'. \tag{44}$$

We can now rewrite our objective function

$$\mathcal{F}(p(s)) = E_\theta \log | E_{p(s)} J_{\exp}(\rho) s s^T | \tag{45}$$

$$= E_\theta \log | E_{p(\vec{w})} J_{\exp}(w_1 \theta_1') \vec{w}\vec{w}^T | \tag{46}$$

$$= E_\theta \log | E_{w_1} J_{\exp}(w_1 \theta_1') E_{w_2,\ldots,w_{\dim(s)}|w_1} \vec{w}\vec{w}^T |. \tag{47}$$

Since $p(s)$ is Gaussian and $\vec{w} = \mathcal{R}_\theta^T\ s,\ p(\vec{w})$ is Gaussian with mean $\mathcal{R}_\theta^T\ \mu_s$ and covariance matrix $\mathcal{R}_\theta^T\ C_s \mathcal{R}_\theta^T$. Consequently, $p(\vec{w}|w_1)$ is also Gaussian and can be computed using the standard Gaussian conditioning formulas,

$$p(\vec{w}|w_1) = \mathcal{N}\left(\mathcal{R}_\theta^T \mu_s + \frac{1}{\sigma_{\omega_1}^2} \mathcal{R}_\theta^T \gamma\,(w_1 - \mu_{\omega_1}), \mathcal{R}_\theta^T C_s \mathcal{R}_\theta - \frac{1}{\sigma_{\omega_1}^2} \mathcal{R}_\theta^T\ \gamma\gamma^T \mathcal{R}_\theta\right) \tag{48}$$

$$\mu_{\omega_1} = \frac{\theta^T}{\|\theta\|_2} \mu_s \tag{49}$$

$$\sigma^2_{\omega_1} = \frac{\theta^T}{||\theta||_2} C_s \frac{\theta}{||\theta||_2} \tag{50}$$

$$\gamma = C_s \frac{\theta}{||\theta||_2}. \tag{51}$$

Using this distribution we can easily compute the conditional expectation,

$$E_{\vec{w}|w_1} \vec{w}\vec{w}^T = \mathscr{R}_\theta^T \left(C_s - \frac{1}{\sigma^2_{\omega_1}} \gamma\gamma^T + \left(\mu_s + \frac{1}{\sigma^2_{\omega_1}} \gamma \left(w_1 - \mu_{\omega_1} \right) \right) \times \left(\mu_s + \frac{1}{\sigma^2_{\omega_1}} \gamma \left(w_1 - \mu_{\omega_1} \right) \right)^T \right) \mathscr{R}_\theta \tag{52}$$

$$= \mathscr{R}_\theta^T \left(C_s + \left(\mu_s + \frac{1}{\sigma^2_{\omega_1}} \gamma \left(w_1 - \mu_{\omega_1} \right) - \frac{1}{\sqrt{\sigma^2_{\omega_1}}} \gamma \right) \times \left(\mu_s + \frac{1}{\sigma^2_{\omega_1}} \gamma \left(w_1 - \mu_{\omega_1} \right) + \frac{1}{\sqrt{\sigma^2_{\omega_1}}} \gamma \right)^T \right) \mathscr{R}_\theta \tag{53}$$

$$= \mathscr{R}_\theta^T \left(C_s + \left(\vec{\kappa} w_1 + \vec{\delta} \right) \left(\vec{\kappa} w_1 + \vec{\eta} \right)^T \right) \mathscr{R}_\theta \tag{54}$$

$$\vec{\kappa} = \frac{\gamma}{\sigma^2_{\omega_1}} \tag{55}$$

$$\vec{\delta} = \mu_s - \frac{\gamma}{\sigma^2_{\omega_1}} \mu_{\omega_1} - \frac{\gamma}{\sqrt{\sigma^2_{\omega_1}}} \tag{56}$$

$$\vec{\eta} = \mu_s - \frac{\gamma}{\sigma^2_{\omega_1}} \mu_{\omega_1} + \frac{\gamma}{\sqrt{\sigma^2_{\omega_1}}} \tag{57}$$

The key point is that the expected value is just a rank-1 perturbation of a rotated C_s. We can now evaluate the expectation over w_1,

$$E_{w_1} J_{\exp}(w_1\theta_1^{'})\, E_{\vec{w}|w_1} \vec{w}\,\vec{w}^T = \mathscr{R}_\theta^T \left(C_s\varpi_1 + \varpi_3 \left[\left(\vec{\kappa} + \frac{\varpi_2}{\varpi_3}\vec{\delta}\right)\left(\vec{\kappa} + \frac{\varpi_2}{\varpi_3}\vec{\eta}\right)^T + \left(\frac{\varpi_1}{\varpi_3} - \left(\frac{\varpi_2}{\varpi_3}\right)^2\right)\vec{\delta}\,\vec{\eta}^T\right]\right)\mathscr{R}_\theta$$

(58)

$$\varpi_1 = E_{w_1} J_{\exp}(w_1\theta_1^{'}) \quad (59)$$

$$\varpi_2 = E_{w_1} J_{\exp}(w_1\theta_1^{'})w_1 \quad (60)$$

$$\varpi_2 = E_{w_1} J_{\exp}(w_1\theta_1^{'})w_1^2; \quad (61)$$

$w_1 = \frac{\theta^T}{||\theta||_2} s$, and $p(w_1)$ is Gaussian with mean and variance $(\mu_{\omega_1}, \sigma^2_{\omega_1})$. The above are just 1-dimensional expectations so for any value of θ we could compute them numerically.

Equation (58) is a rank 2 update of C_s. Therefore we can use the matrix determinant lemma to compute $|E_{w_1} E_{\vec{w}|w_1} J_{\exp} \vec{w}\vec{w}^T|$,

$$\log |E_{w_1} E_{\vec{w}|w_1} J_{\exp}(w_1\theta_1^{'})\vec{w}\,\vec{w}^T| = \dim(C_s) \log \varpi_1 + \log |I + V^T(\varpi_1 C_s)^{-1} U| + \log |C_s| \quad (62)$$

$$U = \varpi_3 \left[\left(\vec{\kappa} + \frac{\varpi_2}{\varpi_3}\vec{\delta}\right), \left(\frac{\varpi_1}{\varpi_3} - \left(\frac{\varpi_2}{\varpi_3}\right)^2\right)\vec{\delta}\right] \quad (63)$$

$$V = \left[\left(\vec{\kappa} + \frac{\varpi_2}{\varpi_3}\vec{\eta}\right), \vec{\eta}\right]. \quad (64)$$

Since $I + V^T (\varpi_1 C_s)^{-1} U$ is a 2-d matrix, we can compute its determinant analytically. Taking the expectation with respect to θ yields,

$$E_\theta \log |E_{w_1} E_{\vec{w}|w_1} J_{\exp}(w_1\theta_1^{'})\vec{w}\,\vec{w}^T| = \dim(C_s) E_\theta \log \varpi_1 + E_\theta \log |I + V^T(\varpi_1 C_s)^{-1} U| + \log |C_s|.$$

(65)

Appendix C: Solving the optimization problem described in Section 4.1.3

Our goal is to solve the optimization problem

$$\arg\max_{\mu_s, C_s} d\vec{\mu}_t^T \mu_s + \frac{d}{2} Tr(C_s R) + \log |C_s|, \tag{66}$$

where we have abbreviated

$$R = \vec{\mu}_t \vec{\mu}_t^T + C_t, \tag{67}$$

over all (μ_s, C_s) subject to the constraints

$$s^T C_s s > 0 \forall s \neq 0 \tag{68}$$

$$Tr(C_s) < m - ||\mu_s||^2, \tag{69}$$

where *m* is the maximum allowed average stimulus power. Clearly the optimal μ_s will be parallel to $\vec{\mu}_t$. Therefore, only the magnitude of the optimal μ_s is unknown. We can therefore rewrite the objective function as

$$\arg\max_{||\mu_s||} \left[d||\vec{\mu}_t||||\mu_s|| + \arg\max_{C_s} \left(\frac{d}{2} Tr(C_s R) + \log |C_s| \right) \right] \tag{70}$$

We rewrite the inner problem to make the dependence on $||\mu_s||$ (through the power constraint) more explicit:

$$\arg\max_{C_s} \frac{d}{2} Tr(C_s R) + \log |C_s| \tag{71}$$

$$s.t s^T C_s s > 0 \forall s \neq 0 \tag{72}$$

$$Tr(C_s) < m - ||\mu_s||^2. \tag{73}$$

We solve this optimization problem by introducing a Lagrange multiplier,

$$L = \frac{d}{2} Tr(C_s R) + \log |C_s| - \lambda Tr(C_s) \tag{74}$$

$$=Tr\left[C_s\left(\frac{d}{2}(R-\lambda I)\right)\right]+\log|C_s|. \tag{75}$$

This Lagrangian is exactly isomorphic to twice the loglikelihood in a multivariate Gaussian model with zero mean, if we interpret C_s as the inverse covariance matrix in the Gaussian model and $-\frac{d}{2}(R-\lambda I)$ as the observed sample covariance matrix. Standard arguments involving a change of basis now imply that the optimal C_s is given by

$$C_s=-\left(\frac{d}{2}(R-\lambda I)\right)^{-1}, \tag{76}$$

for any λ > $\max_i\{r_i\}$, where r_i denotes the i-th eigenvalue of R. This condition on λ is required to ensure that the resulting C_s maximizes the Lagrangian L, and guarantees that C_s is positive definite.

We now solve for λ by plugging this C_s into our power constraint:

$$Tr(C_s)=\frac{2}{d}\sum_i\frac{1}{\lambda-r_i}=m-\|\mu_s\|^2 \tag{77}$$

We can easily solve this equation numerically on the allowed range $\lambda > \max_i r_i$ to compute λ as a function of $\|\mu_s\|$. We can then in principle do a search over all $\|\mu_s\|$ to find the optimal value (μ_s, C_s). In fact, a more efficient method is to instead just compute the optimal ($\|\mu_s\|$, C_s) for each value of λ; thus, a single 1-d search over λ is guaranteed to find the optimal (μ_s, C_s). Also note that we can compute the inverse in Eq. (76) efficiently for any value of λ by computing the eigendecomposition of R once and then using the fact that $eig(R-\lambda I) = eig(R)-\lambda$; we used this formula already in Eq. (77).

Enforcing stationarity by incorporating Toeplitz constraints

It is worth noting, in the case of stimulus filters θ that extend over more than one time bin (i.e., n_t as defined in Appendix A is greater than one), that the stimulus sequence drawn from the Gaussian distribution defined above will not be temporally stationary. Instead, the stimulus sequence will consist of a series of appended n_t-long segments of draws from a Gaussian distribution, and therefore the marginal distribution of the inputs s_t will in general be an n_t-mixture of Gaussians, instead of a single Gaussian distribution. We recover a single Gaussian only in the special stationary case that the stimulus covariance C_s is constrained to have a Toeplitz structure and the mean vector μ_s is constrained to be constant with respect to time-shifts.

Since we are observing the neural responses r_t given the input s_t presented at each time point t, we should arguably optimize our information over this marginal mixture distribution, instead of the single Gaussian distribution optimized in this appendix. Alternately, we could enforce stationarity in our optimized Gaussian process by including Toeplitz constraints on C_s in our derivation above. We have had limited success deriving a computationally efficient optimization strategy in either of these cases, but this remains an attractive direction for future research. Meanwhile, the results described in Section 4.2 (with a non-stationary optimized Gaussian stimulus ensemble) remain encouraging.

References

Anderson, M.; Micheli-Tzanakou, E. Computer-brain interaction to optimize auditory stimuli based on neuronal responses; Bioengineering conference, 1998. Proceedings of the IEEE 24th Annual Northeast; 1998. p. 18-20.

Benda J, Gollisch T, Machens CK, Herz AV. From response to stimulus: Adaptive sampling in sensory physiology. Current Opinion in Neurobiology. 2007; 17(4):430–436. [PubMed: 17689952]

Calabrese, A.; Schumacher, J.; Schneider, D.; Woolley, S.; Paninski, L. A generalized linear model for estimating receptive fields from midbrain responses to natural sounds; Conference Abstract: Computational and Systems Neuroscience; 2010.

Chaloner K, Verdinelli I. Bayesian experimental design: A review. Statistical Science. 1995; 10(3): 273–304.

Chen Z, Becker S, Bondy J, Bruce IC, Haykin SC. A novel model-based hearing compensation design using a gradient-free optimization method. Neural Computation. 2005; 17(12):2648–2671. [PubMed: 16212766]

Cover, TM.; Thomas, JA. Elements of information theory. New York: Wiley; 1991.

David S, Mesgarani N, Shamma S. Estimating sparse spectro-temporal receptive fields with natural stimuli. Network. 2007; 18:191–212. [PubMed: 17852750]

deCharms RC, Blake DT, Merzenich MM. Optimizing sound features for cortical neurons. Science. 1998; 280(5368):1439–1443. [PubMed: 9603734]

Edin F, Machens C, Schutze H, Herz A. Searching for optimal sensory signals: Iterative stimulus reconstruction in closed-loop experiments. Journal of Computational Neuroscience. 2004; 17(1):47–56. [PubMed: 15218353]

Fedorov, VV. Theory of optimal experiments. New York: Academic Press; 1972.

Foldiak P. Stimulus optimisation in primary visual cortex. Neurocomputing. 2001; 38–40:1217–1222.

Gill P, Zhang J, Woolley S, Fremouw T, Theunissen F. Sound representation methods for spectro-temporal receptive field estimation. Journal of Computational Neuroscience. 2006; 21:5–20. [PubMed: 16633939]

Hsu A, Woolley SMN, Fremouw TE, Theunissen FE. Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. Journal of Neuroscience. 2004; 24(41):9201–9211. [PubMed: 15483139]

Kwon, WH.; Han, S. Receding horizon control: Model predictive control for state models. New York: Springer; 2005.

Lewi J, Butera R, Paninski L. Efficient active learning with generalized linear models. AISTATS07. 2007

Lewi J, Butera R, Schneider DM, Woolley SMN, Paninski L. Designing neurophysiology experiments to optimally constrain receptive field models along parametric submanifolds. NIPS. 2008:945–952.

Lewi J, Butera R, Paninski L. Sequential optimal design of neurophysiology experiments. Neural Computation. 2009; 21(3):619–687. [PubMed: 18928364]

Luczak A, Bartho P, Marguet S, Buzsaki G, Harris K. Sequential structure of neocortical spontaneous activity *in vivo*. PNAS. 2007; 104:347–352. [PubMed: 17185420]

Machens C. Adaptive sampling by information maximization. Physical Review Letters. 2002; 88:228104–228107. [PubMed: 12059456]

Machens C, Gollisch T, Kolesnikova O, Herz A. Testing the efficiency of sensory coding with optimal stimulus ensembles. Neuron. 2005; 47(3):447–456. [PubMed: 16055067]

Machens CK, Wehr M, Zador AM. Spectro-temporal receptive fields of subthreshold responses in auditory cortex. Advances in Neural Information Processing Systems. 2003; 15:133–140.

Mackay DJC. Information-based objective functions for active data selection. Neural Computation. 1992; 4(4):590–604.

O’Connor KN, Petkov CI, Sutter ML. Adaptive stimulus optimization for auditory cortical neurons. Journal of Neurophysiology. 2005; 94(6):4051–4067. [PubMed: 16135553]

Ohki K, Chung S, Ch'ng Y, Kara P, Reid C. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. Nature. 2005; 433:597–603. [PubMed: 15660108]

Paninski L. Maximum likelihood estimation of cascade point-process neural encoding models. Network: Computation in Neural Systems. 2004; 15:243–262.

Paninski L. Asymptotic theory of information-theoretic experimental design. Neural Computation. 2005; 17(7):1480–1507. [PubMed: 15901405]

Paninski, L.; Pillow, J.; Lewi, J. Statistical models for neural encoding, decoding, and optimal stimulus design. In: Cisek, P.; Drew, T.; Kalaska, J., editors. Computational neuroscience: Progress in brain research. New York: Elsevier; 2007.

Pillow J, Shlens J, Paninski L, Sher A, Litke A, Chichilnisky E, et al. Spatiotemporal correlations and visual signaling in a complete neuronal population. Nature. 2008; 454:995–999. [PubMed: 18650810]

Segev R, Goodhouse J, Puchalla J, Berry MJ. Recording spikes from a large fraction of the ganglion cells in a retinal patch. Nature Neuroscience. 2004; 7(10):1155–1162.

Singh NC, Theunissen FE. Modulation spectra of natural sounds and ethological theories of auditory processing. The Journal of the Acoustical Society of America. 2003; 114(6):3394–3411. [PubMed: 14714819]

Smyth D, Willmore B, Baker G, Thompson I, Tolhurst D. The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. Journal of Neuroscience. 2003; 23:4746–4759. [PubMed: 12805314]

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network-Computation in Neural Systems. 2001; 12(3):289–316.

Theunissen FE, Sen K, Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. Journal of Neuroscience. 2000; 20(6):2315–2331. [PubMed: 10704507]

Theunissen FE, Woolley SMN, Hsu A, Fremouw T. Methods for the analysis of auditory processing in the brain. Annals of the New York Academy of Sciences. 2004; 1016:187–207. [PubMed: 15313776]

Tzanakou E, Michalak R, Harth E. The alopex process: Visual receptive fields by response feedback. Biological Cybernetics. 1979; 35:161–174. [PubMed: 518937]

Woolley SM, Gill PR, Theunissen FE. Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. The Journal of Neuroscience. 2006; 26:2499–2512. [PubMed: 16510728]

Woolley SMN, Casseday JH. Response properties of single neurons in the zebra finch auditory midbrain: Response patterns, frequency coding, intensity coding, and spike latencies. Journal of Neurophysiology. 2004; 91(1):136–151. [PubMed: 14523072]

Woolley SMN, Casseday JH. Processing of modulated sounds in the zebra finch auditory midbrain: Responses to noise, frequency sweeps, and sinusoidal amplitude modulations. Journal of Neurophysiology. 2005; 94(2):1143–1157. [PubMed: 15817647]

Yamane Y, Carlson E, Bowman K, Wang Z, Connor CE. A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nature Neuroscience. 2008; 11:1352–1360.

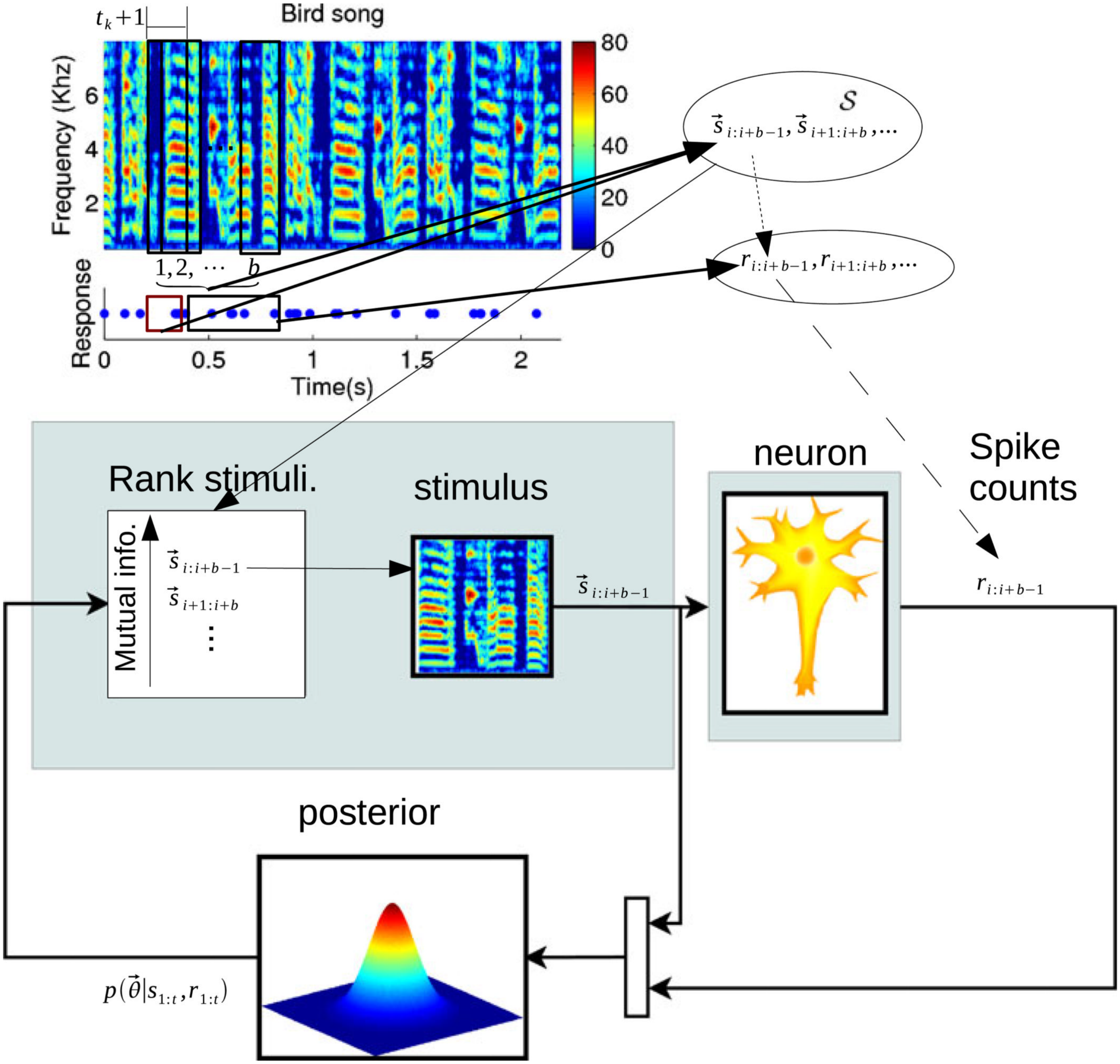
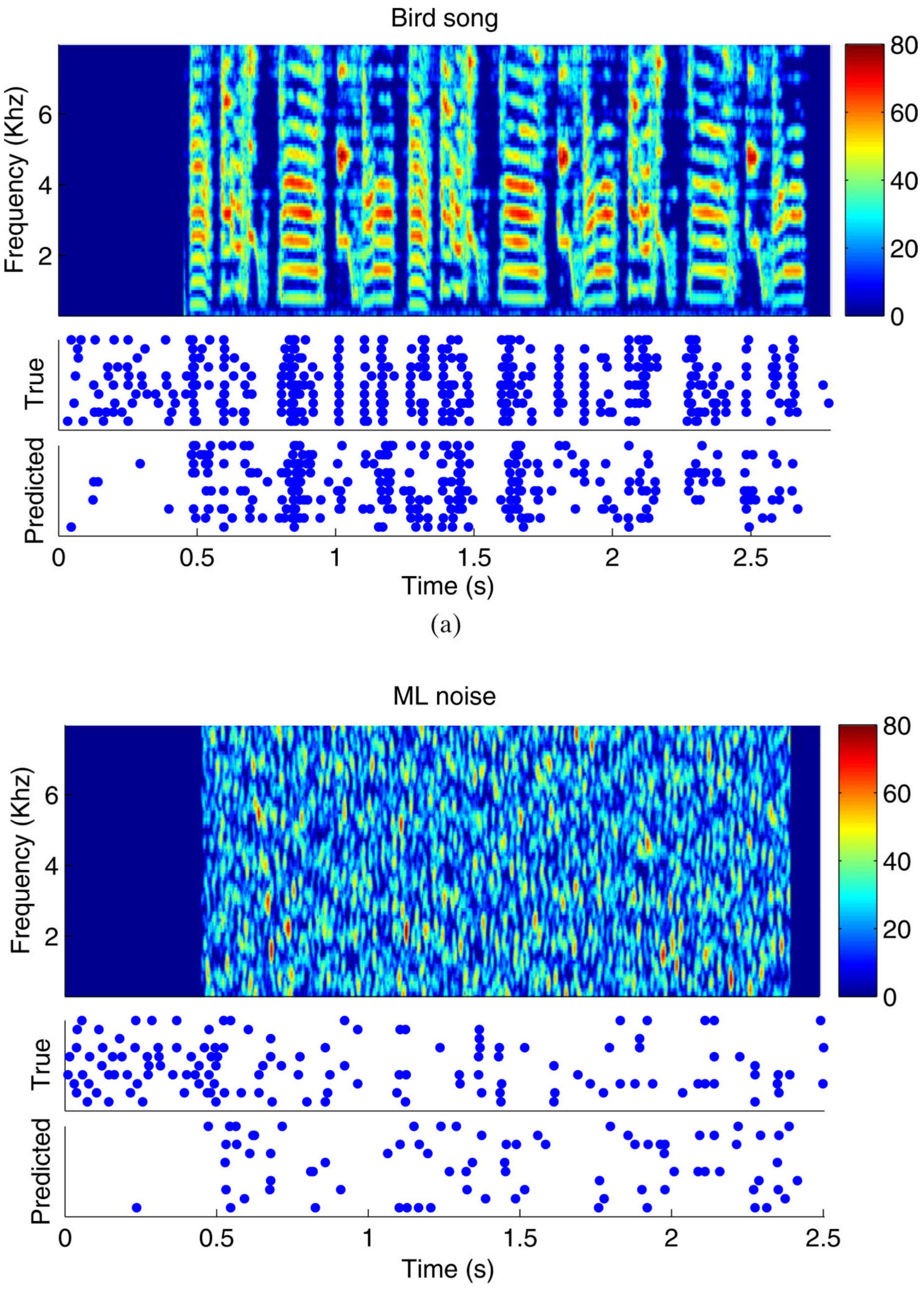


Fig. 1.

A schematic of the online experimental design approach. *Top*: an illustration of how stimulus sequences and the associated responses are obtained from the actual data. Each input, $\vec{s}_t$, consists of $t_k + 1$ columns of the spectrogram (with each column of width $dt = 2.5$ ms here) as indicated by the black boxes on the spectrogram (note the boxes are not drawn to scale, for improved visibility). Each input also consists of the recent spike history as indicated by the red box on the raster plot. On each iteration we pick a sequence of b consecutive inputs. Taken together these b inputs yield one possible stimulus sequence $\vec{s}_{i:i+b-1}$ in the set of possible input sequences, $\mathscr{S}$ The responses to these stimuli, indicated by the black box in the raster plot, are the responses that we will observe when this stimulus sequence is chosen. *Bottom*: the iterative loop that describes our experiment. At each timestep we rank all remaining sequences in $\mathscr{S}$that have not been selected yet. The sequences are ranked according to the lower bound of the mutual information (plugging Eq. (15) into Eq. (4) using our current posterior $p(\theta|s_{t:t}, r_{t:t})$. The sequence which optimizes this objective function is chosen and the neuron’s response to this sequence is used to update our Gaussian approximation of the posterior (Eq. (3)). Using our updated posterior we then re-rank the remaining stimuli in $\mathscr{S}$ having removed the stimulus sequence we just presented, and continue the process



(a)

(b)

Fig. 2.

(**a**) The top plot shows the spectrogram of one of the bird songs used during the experiments (color scale is in units of sound intensity). The middle plot shows the raster plot of the recorded neuron's spiking in response to this stimulus, over ten identical repetitions. The bottom plot shows the predicted raster plot computed using a GLM fitted to the training set. Each row of the raster plots shows the firing of the neuron on independent presentations of the input. (Note that these responses to repeated stimuli are useful for validating any model we fit to the observed data, but repeated responses are not necessary for the stimulus-design procedures developed here.) The training set did not include this song or the ml-noise stimulus shown in (**b**). (**b**) The same as panel (**a**), except the stimulus is ml-noise instead of birdsong. When fitting a GLM, the stimulus, $\vec{x}_t$, corresponds to one column of the spectrogram matrix; the “input” s_t corresponds to $t_k + 1$ adjacent columns, as illustrated in Fig. 1

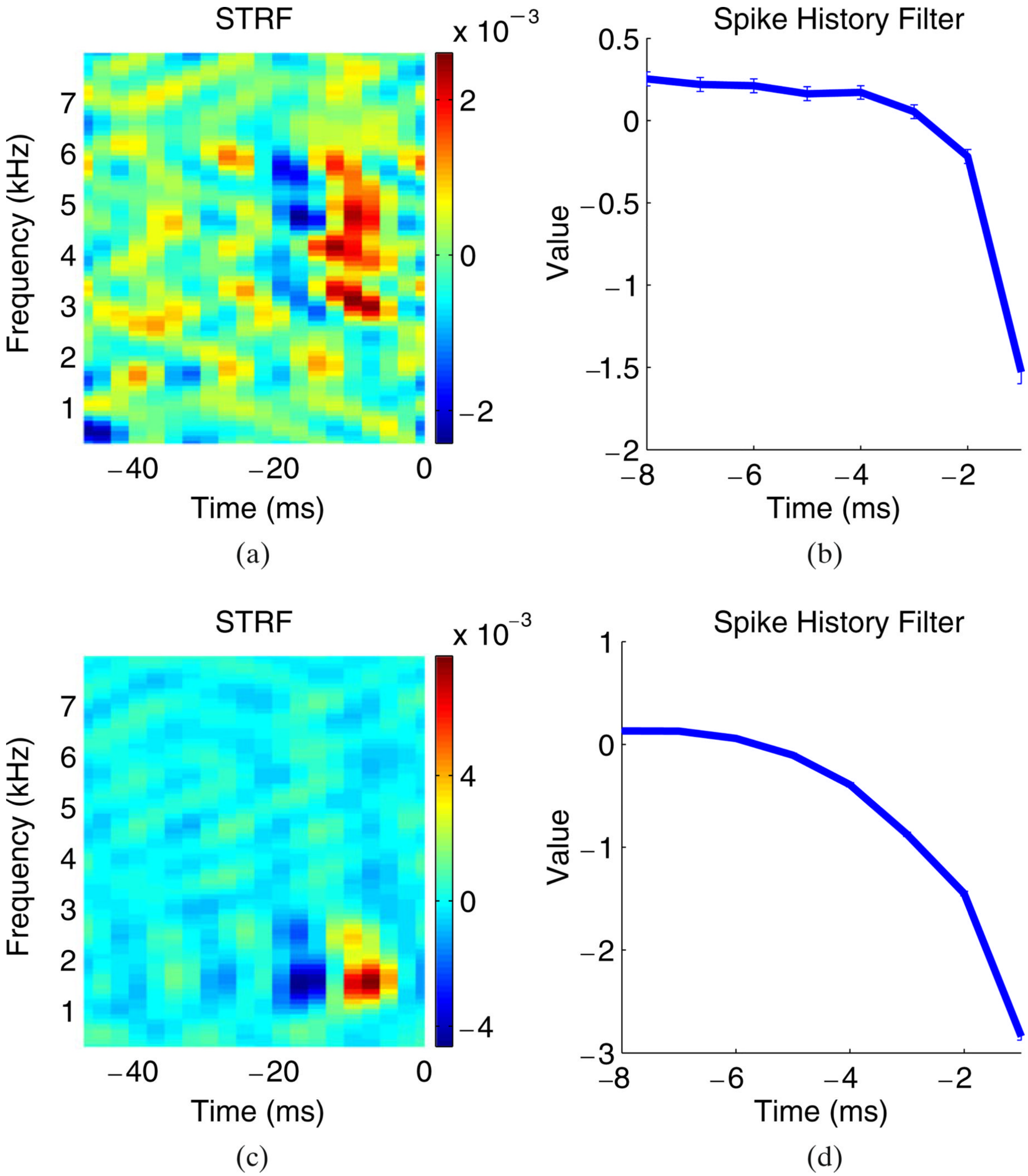


Fig. 3.

The receptive fields for two different neurons estimated with cutoff frequencies $n_{fc} = 10$ and $n_{tc} = 4$; see Appendix A for details. (**a**) The STRF for the first neuron. (**b**) The spike history coefficients (the curve shows the values of the filter coefficients at different delays). The bias in this case was −4.3. The error bars indicate plus and minus one standard deviation for each coefficient. (**c**) The STRF for the second neuron. (**d**) The spike history for the second neuron. The bias in this case was −4.6

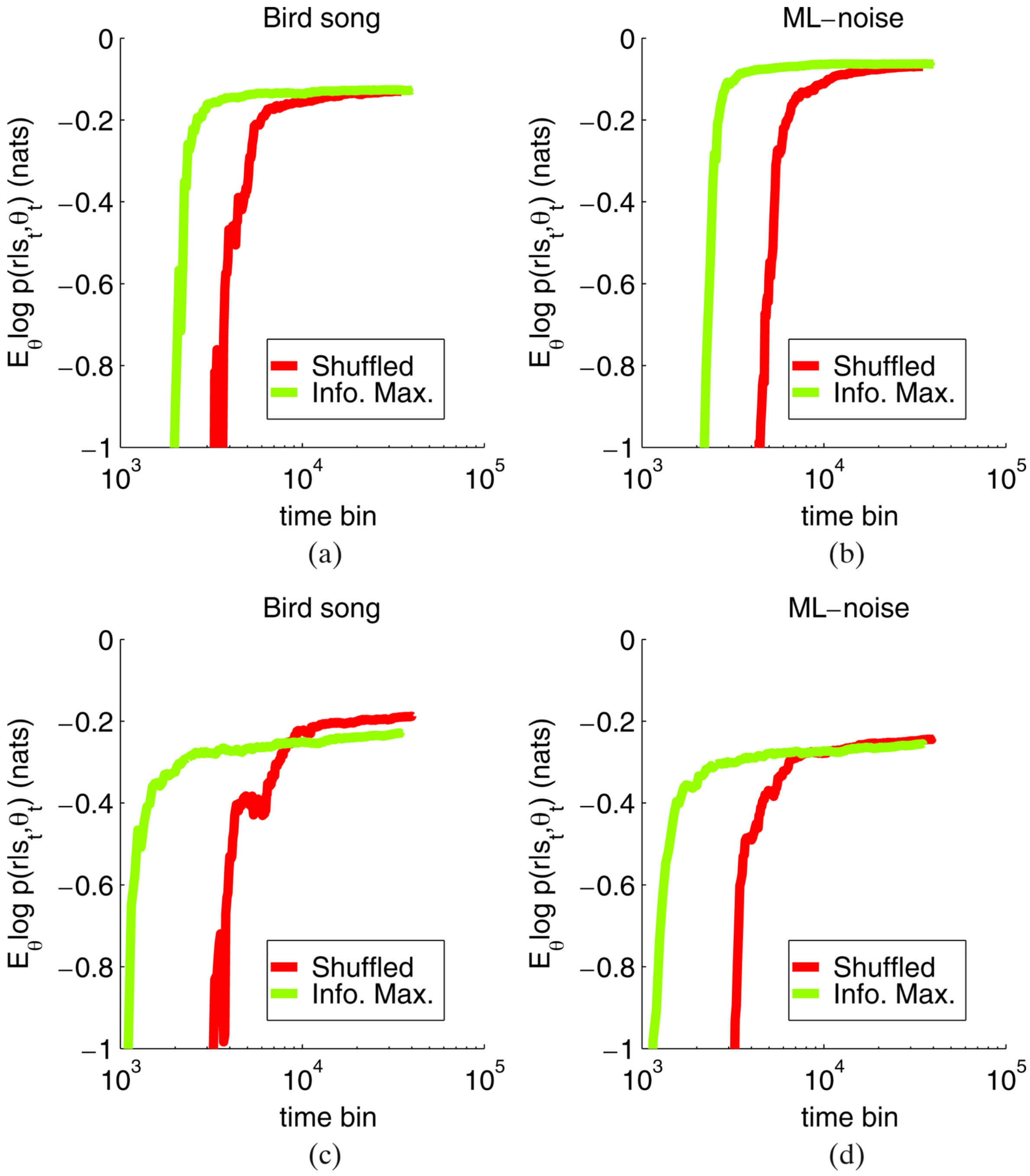


Fig. 4.

Quantifying the relative performance of the info. max. vs. the shuffled experimental design. Each panel shows the expected log-likelihood, up to a normalization constant, computed on the test sets for a different neuron. The test set for each neuron consisted of one bird song and one ml-noise stimulus. The expected log-likelihood is plotted as a function of the number of time bins used to train a model using inputs chosen by either an info. max. or shuffled design as described in the text. The results clearly show that the info. max. design achieves a higher level of prediction accuracy using fewer trials. We quantify the improvement as the "speedup" factor defined in the main text; see Fig. 5(b) and Table 1 for details and quantitative comparisons. As noted in the text, the expected log-likelihood is proportional to the amount of variability of the observations that can be explained by knowing θ and *s*. In this case, a larger number means more of the variability can be explained by knowing θ and *s*. The units are nats

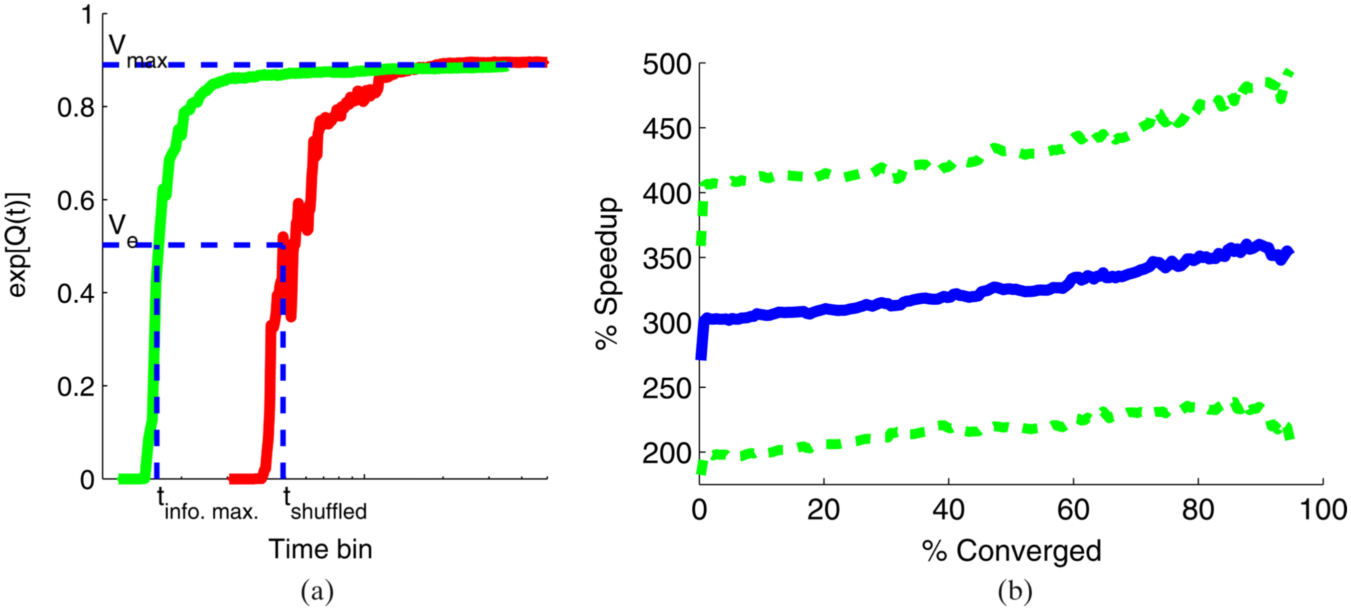


Fig. 5.

(**a**) A plot illustrating the quantities used to compute the speedup due to the infomax approach. For each design, we make a plot of exp[$Q(t)$] (as defined in Eq. (18)) vs. t. The maximum value, V_{max}, is the horizontal asymptote that both traces converge to. V_{max} measures how well we can predict the neuron's responses using a GLM if we train on all the data. For any value, V_e, we can read off the number of stimuli, $t_{\text{info. max.}}$ and t_{shuffled}, required by each design to train a GLM that can account for V_e percent of the response variability. The ratio $\frac{t_{\text{shuffled}}}{t_{\text{info. max.}}}$ is the Speedup as a function of V_e. b) A plot of the speedup achieved by using the info. max. design instead of a shuffled design. The speedup is plotted as a function of % Converged, as described in the text following Eq. (18). The solid blue line shows the average speedup across all 11 neurons and the dashed green lines show plus and minus one standard deviation. The results show that using a shuffled design would require roughly 3 times as many trials to achieve the same level of prediction accuracy as the info. max. design

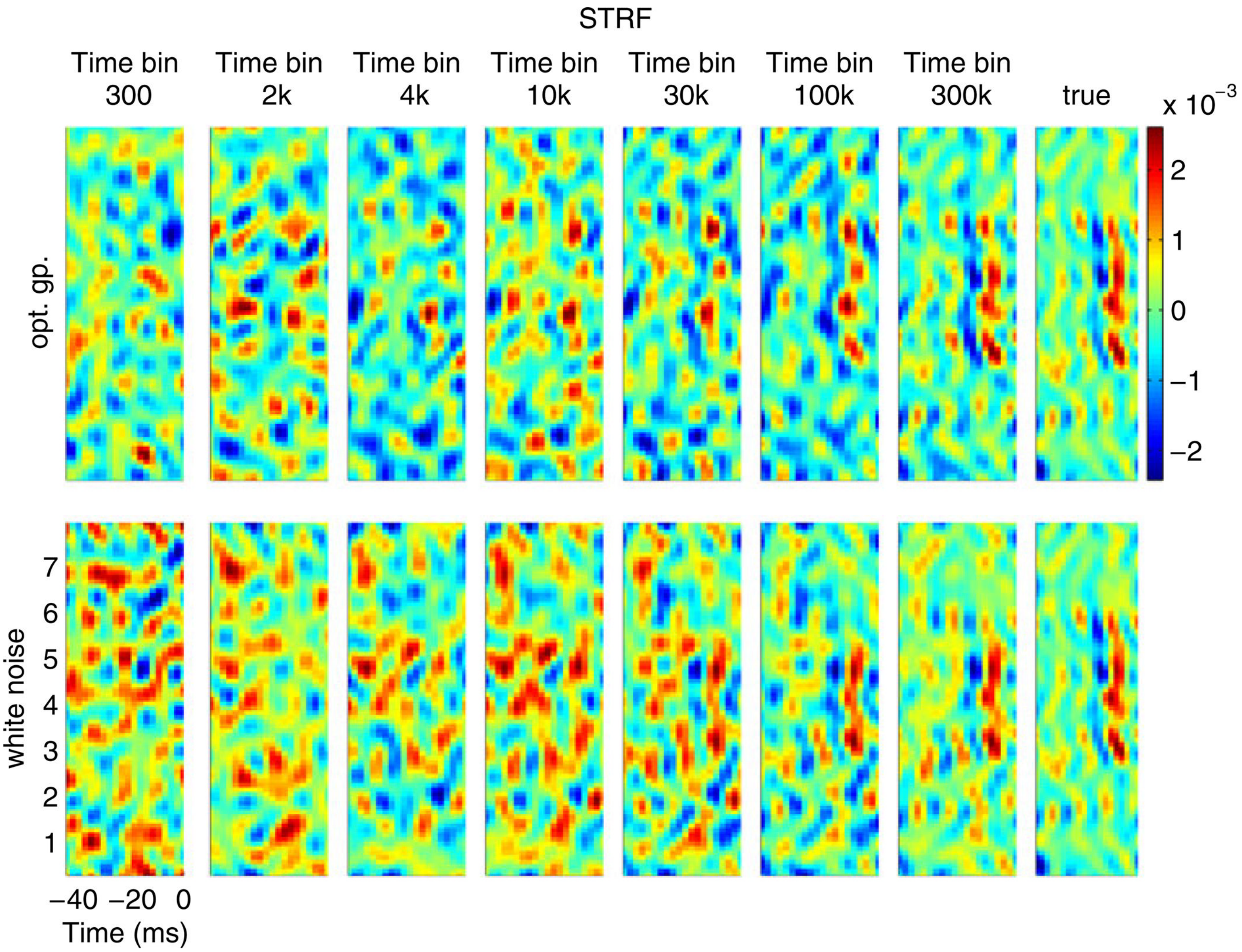


Fig. 6.

Simulation results comparing the posterior mean estimates of the STRF estimated using a white vs. optimized Gaussian process. The STRF used to generate the data was the STRF fitted to the neuron in Fig. 3(a). In this case the posterior mean estimates seem to converge to the true STRF at the same qualitative rate; however, as shown in Fig. 7 below, the uncertainty of our estimate shrinks more rapidly under the optimized Gaussian process design

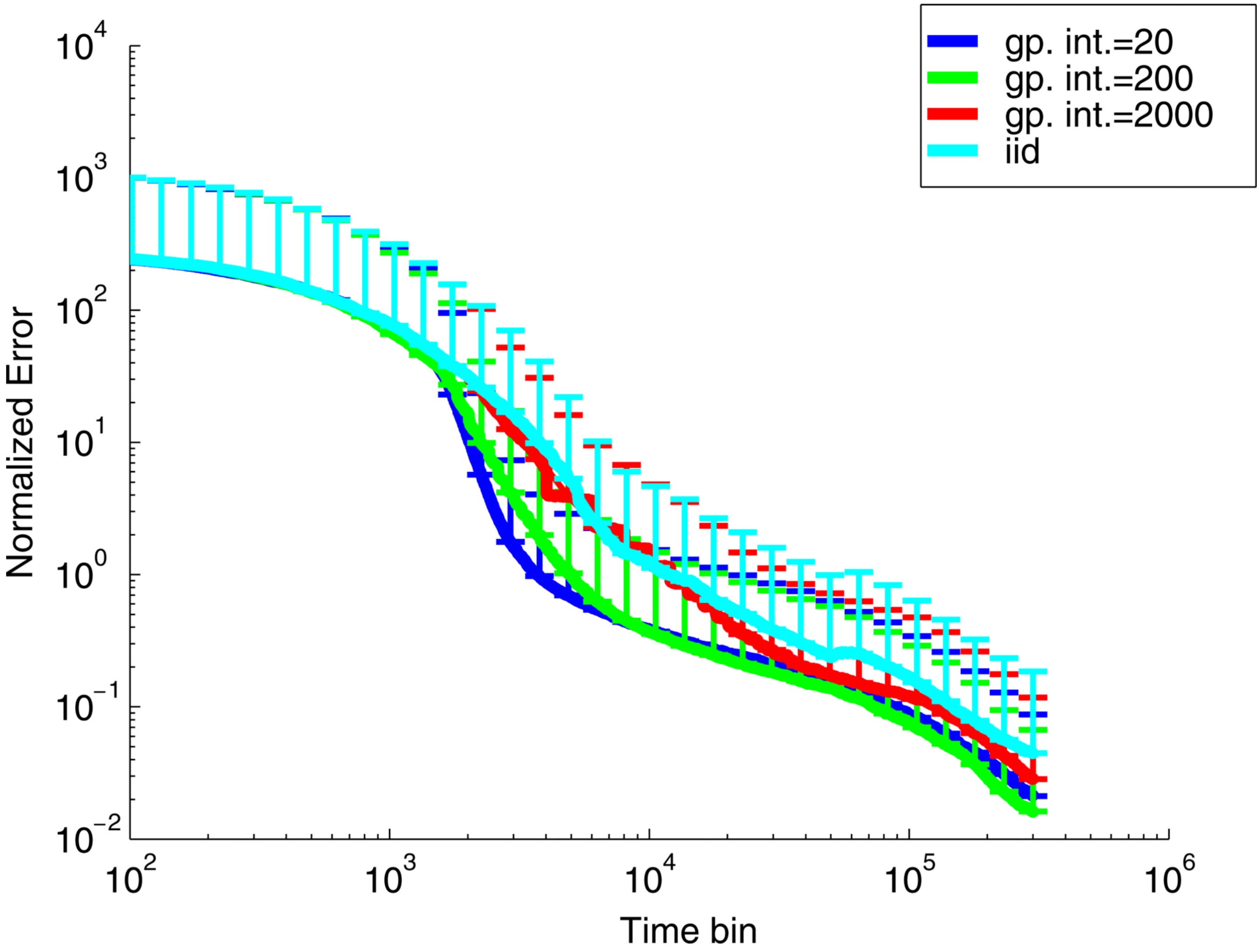


Fig. 7.

Plots of the expected squared error between the posterior mean and the corresponding true value of θ. The error is only computed for the stimulus coefficients (i.e the spike history and bias terms were not included). The expected error is normalized by the square of the magnitude of the STRF. For the optimal design, we tried several different designs in which we varied the interval over which we updated our Gaussian stimulus distribution. The interval corresponding to each trace is shown in the legend. The results show that we can achieve an improvement over an i.i.d. design even if we only update the Gaussian process every 200 timesteps (≈ 500 ms in this case); no major improvement is seen (relative to the i.i.d. design) if we only update every 2,000 timesteps. For each design, we repeated the simulation 10 times, using the same real STRF each time (shown in Fig. 3(a)), and computed the mean and standard deviation of the expected squared error. The results did not depend strongly on the STRF used to simulate the data. The plot shows the mean of the error and the errorbars show plus one standard deviation

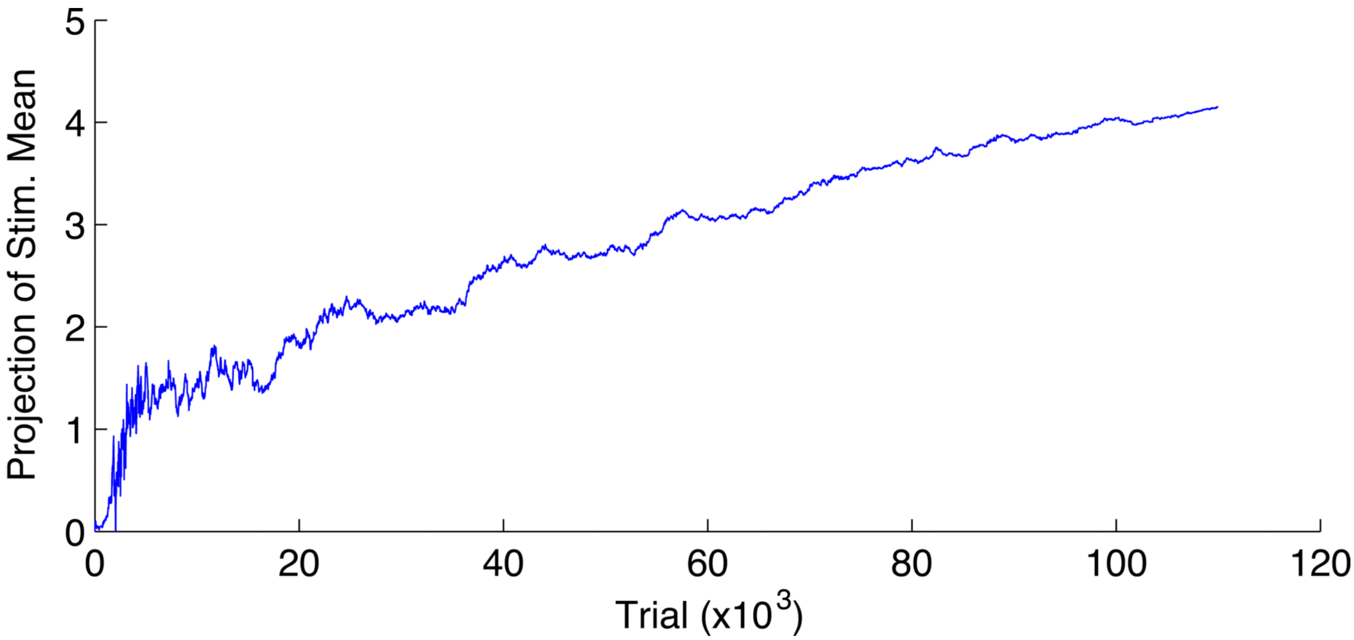


Fig. 8.

A plot of the projection of the stimulus mean onto the posterior mean estimate μ_t, for the design using the optimized Gaussian process. The projection tends to increase with *t*, which corresponds to an increase in the expected firing rate and in the informativeness of each stimulus

Table 1

A table listing the median as well as minimum and maximum values of the speedup evaluated over all 11 neurons

	Bird song	**ml-noise**
Median	330%	310%
Min	160%	200%
Max	630%	480%

The statistics are computed separately on the bird song and ml-noise stimuli in the test set. The statistics were computed at 50% converged (c.f. Fig. 5(b))