A new inference approach for training shallow and deep generalized linear models of noisy interacting neurons

Gabriel Mahuas[1,3], **Giulio Isacchini**[1,4], **Olivier Marre**[2], **Ulisse Ferrari**[2, ‡], **Thierry Mora**[1, ‡]

[1]Laboratoire de physique de l'École normale supérieure, PSL University, CNRS, Sorbonne University, and University of Paris, 75005 Paris, France
[2]Institut de la Vision, Sorbonne University, INSERM, CNRS, 75012 Paris, France
[3]Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria
[4]Max Planck Institute for Dynamics and Self-organization, 37077 Göttingen, Germany
[‡]These authors contributed equally to this work

Abstract

Generalized linear models are one of the most efficient paradigms for predicting the correlated stochastic activity of neuronal networks in response to external stimuli, with applications in many brain areas. However, when dealing with complex stimuli, their parameters often do not generalize across different stimulus statistics, leading to degraded performance and blowup instabilities. Here, we develop a two-step inference strategy that allows us to train robust generalized linear models of interacting neurons, by explicitly separating the effects of stimulus correlations and noise correlations in each training step. Applying this approach to the responses of retinal ganglion cells to complex visual stimuli, we show that, compared to classical methods, the models trained in this way exhibit improved performance, are more stable, yield robust interaction networks, and generalize well across complex visual statistics. The method can be extended to deep convolutional neural networks, leading to models with high predictive accuracy for both the neuron firing rates and their correlations.

1 Introduction

The pioneering work of J.W. Pillow and colleagues [1] showed how the Generalized Linear Model (GLM) can be used for predicting the stochastic response of neurons to external stimuli. Thanks to its versatility [2], high performance, and easy inference, the GLM has become one of the reference models in computational neuroscience. Nowadays, its applications range from retinal ganglion cells [1], to neurons in the LGN [3], visual [4], motor [5], parietal [6] cortices, as well as other brain regions [7, 8]. However, the GLM has also shown some significant limitations that has prevented its application to an even wider spectrum of contexts. In particular, the GLM shows unsatisfying performance when applied to the response to complex stimuli with spatio-temporal correlations much stronger than white noise, as for example naturalistic images [9] or videos [10].

A first limitation is that the inferred parameters depend on the stimulus used for training. This happens not only for the part of the model that deals with the external stimulus, which typically suffers a change in the stimulus statistics, but also for the couplings parameters quantifying interactions between the neurons of the network. However, if these couplings are to reflect an underlying network of biological interactions, they should be stimulus independent. In addition, and as we show in this paper, this lack of generalizability comes with errors in the prediction of correlated noise between neurons. This

Preprint. Under review.

issue can strongly limit the application of GLM for unveiling direct synaptic connections between the recorded neurons [11] and for estimating the impact of noise correlations in information transmission [1].

A second issue is that the GLM can be subject to uncontrollable and unnatural self-excitation transients [12, 13, 14]. During these strong and positive feedback loops, the network's past activity may drive its current state to excitations above naturalistic levels, in turn activating neurons in subsequent time steps and resulting in a transient of very high, unrealistic activity. This problem limits the use of the GLM as a generative model—it is often necessary to remove those self-excitation runs by hand. Ref. [12] proposed an extension of the GLM that also includes quadratic terms limiting self-excitations of the network, but this comes at the price of more fitting parameters and harder inference. Ref. [13] showed that a GLM that predicts the responses several time-steps ahead in time [15] limits self-excitation, but this implies higher computational complexity and the risk of missing fine temporal structures. Alternatively, Ref. [14] proposed an approximation to estimate the stability of the inferred GLM model, and then used a stability criterion to constrain the parameter space over stable models. However the resulting models are sub-optimal, with degraded performance.

Thirdly, because neuronal responses are highly non-linear and hard to model for complex stimuli, the GLM fails to predict those responses correctly, even for early visual areas such as the retina [10]. Recently deep convolutional neural networks (CNNs) have been shown to outperform the GLM at predicting individual neuron mean responses [9, 16, 17, 18]. Compared to the GLM, these deep CNNs benefit from a more flexible and richer network architecture allowing for strong performance improvements [9]. However, the GLM retains an advantage over CNNs: thanks to the couplings between neurons in the same layer, it can account for both shared noise across the population and self-inhibition due to refractoriness. This feature, which is missing from deep CNNs [9], can be used to study how noise correlated in space and time impacts the population response [1]. A joint model combining the benefits of the deep architecture of CNNs and the neuronal couplings of the GLM is still lacking. It would allow for a more detailed description of the neuronal response to stimulus.

In this paper we develop a two-step inference strategy for the GLM that solves these three issues. We apply it to recordings in the rat retina subject to different visual stimulations. The main idea is to use the responses to a repeated stimulus to infer the GLM couplings without including the stimulus processing component. Then, in a second, independent step, we infer the parameters of the model pertaining to stimulus processing. Our approach allows for a wide variety of architectures, including deep CNNs. Finally, we introduce an approximation scheme to put together the two inference results into a single model that can predict the joint network response from the stimulus.

2 Recordings

Retinal ganglion cells of a long-evans rat were recorded through a multi-electrode array experiment [19, 20] and spike-sorted with *SpyKING CIRCUS* [21]. Cell activity was stimulated with one unrepeated and two repeated videos of checkerboard (white-noise) and moving bars. For the checkerboard, we used the unrepeated ($1350s$) and one of the two repeated videos ($996s$ in total for 120 repetitions) for training, and the second repeated video for testing ($756s$ in total for 120 repetitions). Similarly, for the moving bar video we used the unrepeated ($1750s$) and one of the two repeated videos ($165s$ in total for 50 repetitions) for training, and the second repeated video for testing ($330s$ in total for 50 repetitions). In addition, we also recorded responses from a full-field movie with naturalistic statistics [19].

After sorting, we applied a spike-triggered average analysis to locate the receptive fields of each cell. Then, we used the response to full-field stimulation to cluster cells into different cell-types. In this work we focus on a population of $N = 25$ OFF Alpha retinal ganglion cells, which tile the visual field through a regular mosaic. The responses to both checkerboard and moving bar stimulations showed strong correlations, which we decompose into the sum of stimulus and noise correlations. Stimulus correlations are correlations between the cell mean responses (Peristimulus time histogram or PSTH). They are large only for the bar video, mostly because the video itself has strong and long-ranged correlations. Noise correlations, on the other hand, are due to shared noise from upstream neurons and gap junctions between cells in the same layer [22], and mostly reflect the architecture of the underlying biological network. Consistently, noise correlations were similar in the response of the two stimulations. In Suppl. sect. S1 we present additional statistics of the data.

3 Generalized linear model

In our Poisson GLM framework, $n^i(t)$, the number of spikes emitted by cell i in the time-bin t of duration $dt = 1.67ms$, follows a Poisson distribution with mean $\lambda^i(t)$: $n^i(t) \sim \text{Pois}(\lambda^i(t))$. The vector of the cells' firing rate $\{\lambda^i(t)\}_{i=1}^N$, with $N = 25$ is then estimated as

$$\lambda^i(t) = \exp\left\{h^i_{\text{stim}}(t) + h^i_{\text{int}}(t)\right\} \ , \tag{1}$$

where

$$h^i_{\text{int}}(t) = \sum_j \sum_\tau J_{ij}(\tau) n^j(t-\tau) \tag{2}$$

accounts for both past firing history of cell i itself and the contribution coming from all other cells in the network: J_{ii} are the spike-history filters, whereas $J_{i\neq j}$ are coupling filters. Both integrate the past up to 40ms. $h^i_{\text{stim}}(t)$ is a contribution accounting for stimulus drives, which takes the form of a linear spatio-temporal convolution in the classical GLM:

$$h^i_{\text{stim}}(t) = \sum_\tau \sum_{xy} K_{x,y}(\tau) S_{x,y}(t-\tau) \ , \tag{3}$$

where $S_{x,y}(t)$ is the stimulus movie at time t, $\{x, y\}$ being the pixel coordinates and $K_{x,y}(\tau)$ is a linear filter that integrates the past up to 500 ms. Later in the paper, we will go beyond this classical architecture and will allow for deep, non-linear architectures.

In order to regularize couplings and spike-history filters during the inferences, we projected their temporal part over a raised cosine basis [1] of 4 and 7 elements respectively, and added an $L1$-regularization $= 0.1$, which we kept the same for all the inferences. In addition, we imposed an absolute refractory period of τ^i_{refr} time-bins during simulations and consequently $J_{ii}(\tau)$ were set to zero for $\tau \leq \tau^i_{\text{refr}}$. In order to lower its dimension, the temporal behavior of stimulus filter $K_{x,y}(\tau)$ was projected on a raised cosine basis with 10 element. In addition we included an L1 regularization over the basis weights and a L2 regularization over the spatial Laplacian to induce smoothness.

All the inferences were done by log-likelihood (log-ℓ) maximization with Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, using the empirical past spike activity during training [1]. For easy comparison, all the performances discussed below are summarized in Table 1.

4 Failure of GLM for complex stimuli

We inferred the GLM by whole log-ℓ maximization from both the response to the checkerboard and moving bar non-repeated stimulations, and then simulated its response to the repeated videos (Fig. 1). Consistent with [1], in the case of the checkerboard stimulus, the model can predict with high accuracy the PSTH of all cells (Fig. 1A, mean Pearson's $\rho = 0.82 \pm 0.05$). It also reproduces the values of the zero-lag (17 ms window) noise correlations for all cell pairs (Fig. 1B, coefficient of determination CoD$= 0.94$), and the temporal structure of noise cross-correlations (Fig. 1C).

The model performance is very degraded for the moving bars video—a stimulus characterised by long-range correlations. The model reproduces the empirical PSTH with rather good accuracy (Fig. 1D, $\rho = 0.71 \pm 0.10$) and shows fair overall accuracy on the noise correlations (Fig. 1E, CoD$= 0.55$). However it overestimates the value of the noise correlation for certain distant cell pairs (Fig. 1E&F). A closer look reveals that the model overestimates noise correlations for pairs of cells that are strongly stimulus-correlated (Fig. 1G). Here the error in the estimates is normalized over the empirical value of the noise correlations with a cut-off at three standard deviations. Interestingly, the effect is strong only for the moving bar video, as stimulus correlations are small for checkerboard stimulation. These results show that the inferred couplings of the GLM do not depend only on the correlated noise among the neurons, but can also be influenced by stimulus correlations. This prevents the inferred couplings from generalizing across stimuli. In addition, we observed several self-excitation transients when simulating the GLM inferred from the moving-bars stimulus (10% of the time, in 36% of the repetitions, Fig.1H, versus 0% for the model inferred from the checkerboard stimulus). This effect is probably the consequence of the over-estimation of those cell-to-cell couplings in the moving-bars stimulus, which drive the over-excitation of the network.

All these issues can be ascribed to the fact that by maximising the whole log-ℓ over all the parameters simultaneously, the GLM mixes the impact of stimulus correlations with neuronal past activity. In the

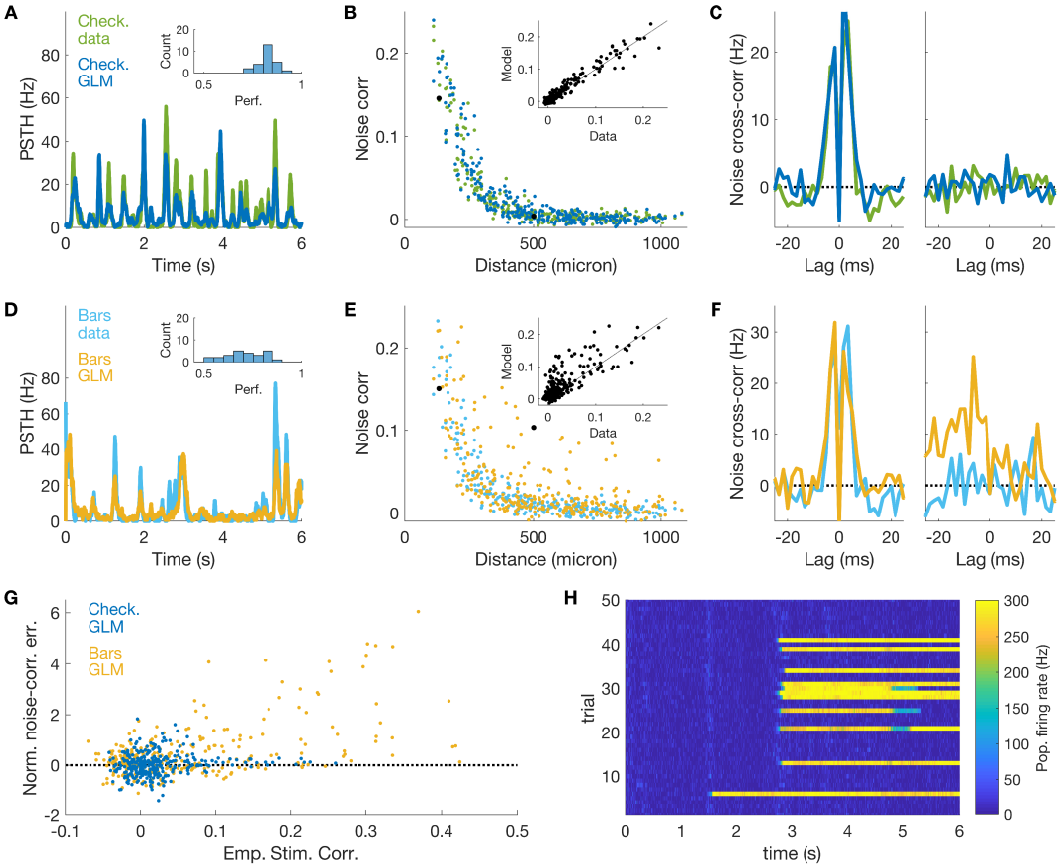

Figure 1: **GLM fails to predict noise correlations in the presence of strong stimulus correlations.** A) PSTH prediction for the response of an example cell to checkerboard stimulation. Inset: histogram of the model performance (Pearson correlation between empirical and model PSTH) for all cells in the population. B) Empirical and model predicted noise correlations versus distance between the cells. Inset: scatterplot. C) Empirical and model predicted noise cross-correlation between a nearby and a distant example cells. D,E,F) same as A,B,C, but for the response to moving bar stimulation. Note that the model overestimates noise correlations between certain pairs of distant cells. G) Error in the prediction of noise correlations normalized over their empirical value versus the empirical value of stimulus correlations. H) Population firing rate in time during model simulations of the responses to the moving bar stimulus. Note the transient of unnatural high activity due to self-excitation within the model.

next section we develop an inference strategy that disentangles stimulus from noise correlations and infer their parameters independently.

5 A two-step inference approach

In order to disentangle the inference of the couplings between neurons from that of the stimulus filters, we split the model training into two independent steps. We name this approach "two-step" inference (Fig. 2).

Filter inference. We run the inference of a GLM model without the couplings between neurons or with themselves (spike-history filter) using the responses to the unrepeated stimulus (single-neuron linear-nonlinear Poisson (LNP) models [23], Fig. 2B). This inference allows us to predict the mean firing rate λ of each neuron.

Coupling inference. We run a log-ℓ maximization inference over the response to a repeated video stimulation. Instead of inferring the parameters of a stimulus filter ($K_{x,y}(\tau)$ in Eq. 3), we treat

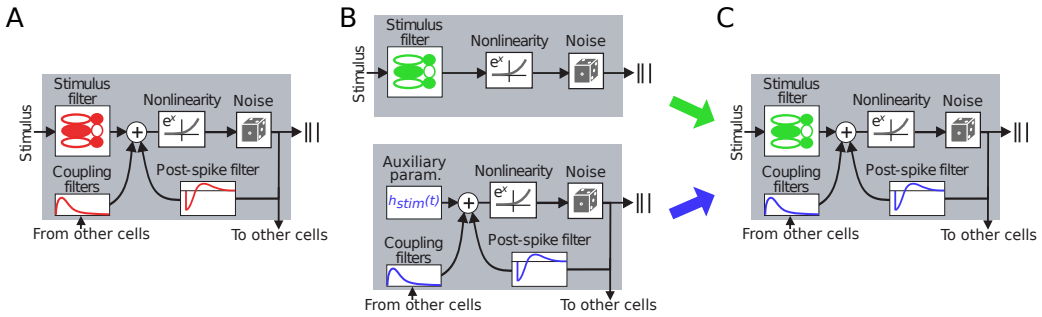

Figure 2: **Two-step inference of couplings and spike-history filters.** A) Whole log-ℓ maximization [1] trains couplings and spike-history filters together with the stimulus filter. B) Two-step inference trains couplings filters and stimulus filters by running two independent log-ℓ maximizations. Top: we remove coupling filters and infer the equivalent of an LNP model for each cell. Bottom: we run an inference over repeated data where we add auxiliary variables (instead of the stimulus filter) to exactly enforce the PSTH prediction. C) We build together the model by using the previously inferred parameters. A correction needs to be added (not shown, see text).

the terms $h^i_{\text{stim}}(t)$ of Eq. 1 as auxiliary parameters that we infer directly from data (Fig. 2B). The log-ℓ derivative over these parameters is proportional to the difference between empirical and model-predicted PSTH. As a consequence, and thanks to repeated data, the addition of these parameters allows for enforcing the value of the PSTH exactly when the corresponding log-ℓ gradient vanishes. In this way, stimulus correlations are perfectly accounted for, and the couplings only reflect correlated noise between neurons. As for the GLM inferred with whole log-ℓ maximization, we imposed an absolute refractory period of τ^i_{refr} time-bins during simulations and thus set $J_{ii}(\tau)$ to zero for $\tau \leq \tau^i_{\text{refr}}$.

Full model. Once couplings and stimulus filters are inferred, we can combine them to build up the full model (Fig. 2C). This cannot be done straightforwardly because the addition of the couplings will change the firing rate prediction of LNP model. To correct for this effect, we subtract the mean contribution of the coupling term: $h^i_{\text{int}}(t) \to h^i_{\text{int}}(t) - \langle h^i_{\text{int}}(t)\rangle_{\text{noise}\sim\text{Pois}}$. This correction is equivalent to modify Eq. 2 into

$$\sum_j \sum_\tau J_{ij}(\tau) n^j(t-\tau) \to \sum_j \sum_\tau J_{ij}(\tau)\Big(n^j(t-\tau) - \lambda^i(t-\tau)\Big) \,. \tag{4}$$

Lastly, in order to account for the addition of absolute refractory periods, we added a term $\sum_{\tau=1}^{\tau^i_{\text{refr}}} \lambda^i(t-\tau)$ for each neurons (Suppl. Sect. S2). To compute all the corrections, we therefore only need the past firing rates $\lambda^i(t)$ of all neurons in the absence of the couplings, which are given by the LNP model predictions. This allows the full model to predict the neuronal response to unseen (testing) data.

We first applied our two-step inference to the response to checkerboard stimulation and obtained very similar results to whole log-ℓ maximization (Table 1). By constrast, performance was improved in the case of the moving bar stimulus (Fig. 3). The two inference approaches yielded similar performances for the PSTH (Fig. 3A, $\rho = 0.72 \pm 0.10$, versus $\rho = 0.71 \pm 0.10$), but for noise correlations we obtained much better results (Fig. 3B, CoD$= 0.91$, versus CoD$= 0.55$). In particular, the model avoids the overestimation the noise correlations for distant pairs (Fig. 3B&C) that we obtained with whole log-ℓ maximization (Fig. 1E&F). With the two-step inference, the strong stimulus correlations of the moving bar video do not affect the model inference as was the case for whole log-ℓ maximization (Fig. 3D). In addition the model is much more stable, and we never observed self-excitation for either stimulus when simulating the model (Fig.3E, versus 10% of the time, Fig. 1H). In Table 1 we report all the performance for the different cases.

6 Two-step inference allows for generalizing across stimuli

So far we have shown how our two-step approach can disentangle the inference of neuronal couplings from stimulus correlations. If these couplings are only due to network effects, one should expect

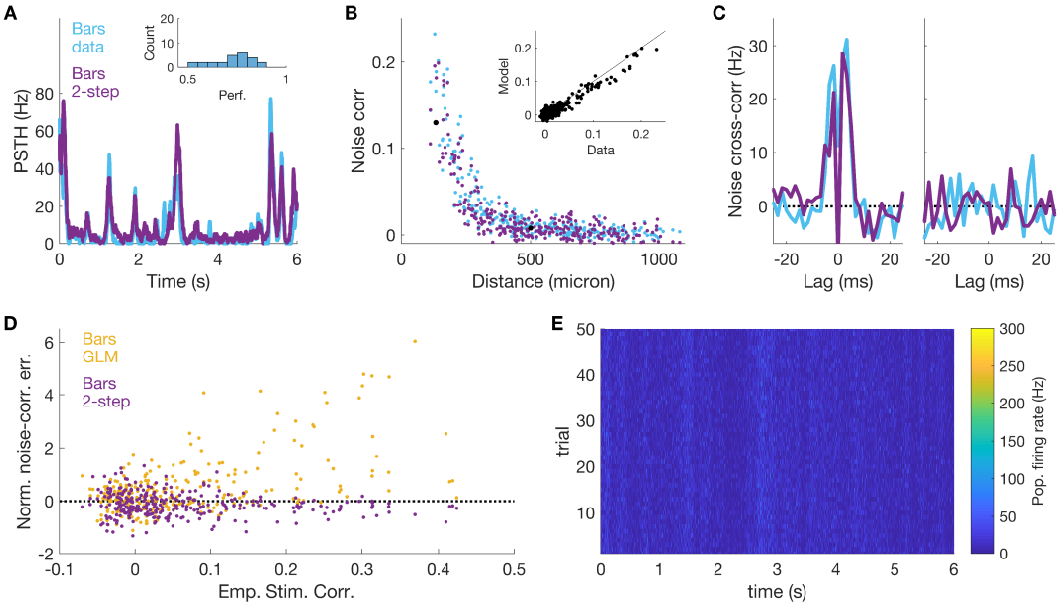

Figure 3: **Two-step inference retrieves noise correlations independently of the strong stimulus correlations in the moving bar video.** A) PSTH prediction for an example cell. Inset: histogram of the model performance for all cells. B) Empirical and model-predicted noise correlations versus distance between the cells. Inset: scatterplot. C) Empirical and model predicted noise cross-correlation between example pairs of nearby and distant cells. D) Normalized error in the prediction of noise correlations plotted versus the empirical value of the stimulus correlations. E) Population activity during model simulation shows no self-excitation transients.

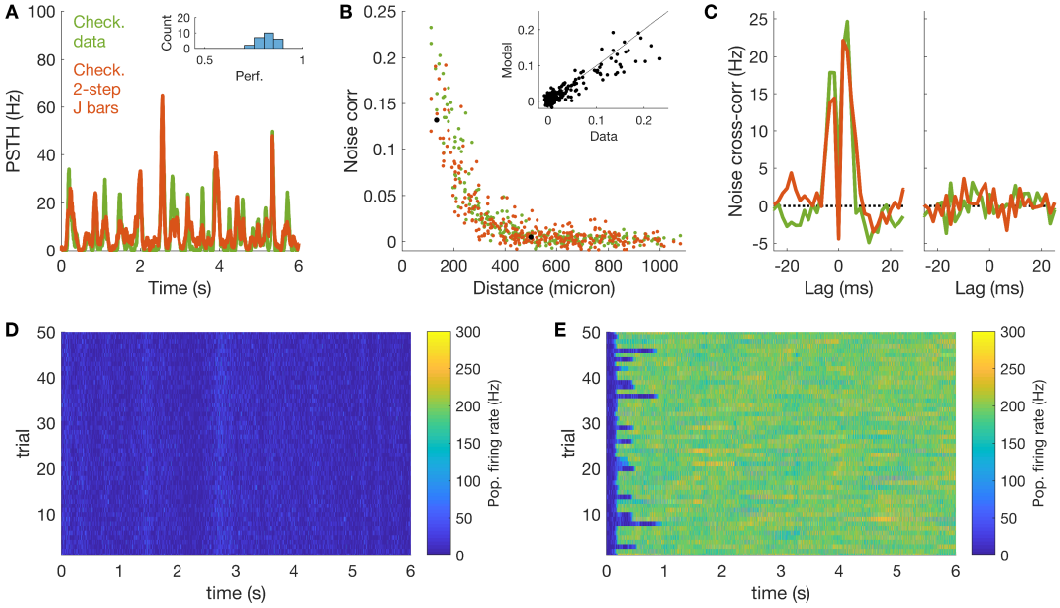

Figure 4: **Two-step inference allows for generalizing across stimulus ensembles** A,B,C,D) Simulation of the checkerboard responses for a model where stimulus filters were inferred from the response to checkerboard, and couplings filter were inferred from the moving bar data with our two-step inference. A) PSTH predictions. B) Noise correlations. C) Noise cross-correlation. D) Population activity showed no self-excitation transients E) Simulation of checkerboard responses when couplings filters are those inferred from moving bar data with whole log-ℓ maximization. The model shows self-excitation during all runs.

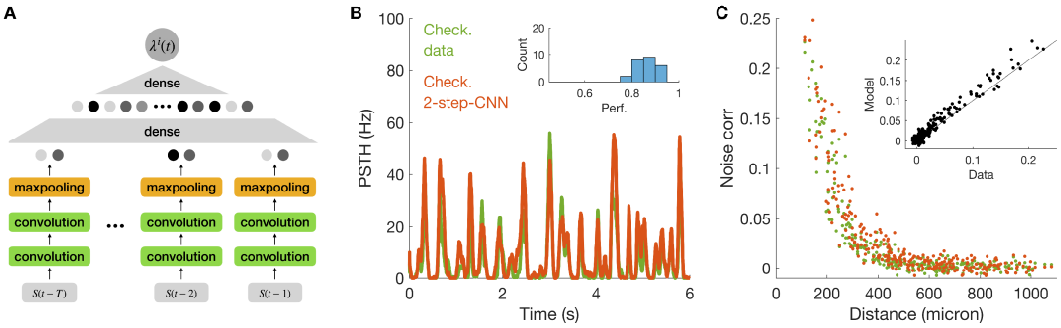

Figure 5: **Deep CNN can be included in our two-step approach to improve model performance** A) Architecture of our deep, time-distributed CNN. B) PSTH prediction for the response of an example cell to checkerboard stimulation. Inset: histogram of model performance for all cells. C) Empirical and model predicted noise correlations versus distance between cells. Inset: scatterplot.

them to generalize across stimulus conditions. To test for this, we run model simulations of one stimulus using its stimulus filter and the coupling filters inferred from the other. For the checkerboard movie (Fig. 4), and compared to the case where couplings are inferred on the same stimulus, with our two-step inference we obtained performances that are almost equal for the PSTH ($\rho = 0.81 \pm 0.05$, versus $\rho = 0.81 \pm 0.05$) and rather good for noise correlations (CoD$= 0.84$, versus CoD$= 0.95$). In addition, we never observed self-excitation (Fig. 4D). By contrast, when we used the couplings inferred by whole log-ℓ maximization, self-excitation happens so often (93% of the time in 100% of the repetitions) that we were not able to estimate the model performance (Fig. 4E).

For the moving bar video (Fig. S2), our two-step inference yielded performances similar to the case where couplings were inferred on the same stimulus (Table 1). Using the couplings inferred by whole log-ℓ maximization instead, the model performance decreased for the PSTH ($\rho = 0.65 \pm 0.12$, versus $\rho = 0.71 \pm 0.10$), and improved for noise correlations (CoD$= 0.80$, versus CoD$= 0.55$). In conclusion, two-step outperforms whole log-ℓ maximization on both stimuli (Table 1).

7 Deep GLM outperforms previous approaches

Our two-step inference decomposes the model training into two independent components, one for the stimulus processing and one for network effects. In the previous experiments we still used a linear convolution to process the stimulus, but thanks to this decomposition, we can also consider any machine capable of predicting the neurons firing rates $\{\lambda^i(t)\}_{i=1}^N$. In order to predict the response to checkerboard stimulation with higher accuracy, we inferred a deep, time-distributed CNN, a special case of CNNs [9] with the additional constraint that the weights of the convolutional layers are shared in time [24]. In our architecture, two time-distributed convolutional layers are followed by a max-pooling and eventually by two dense layers that output the firing rate $\lambda^i(t)$ (Fig. 5A). After training, we included the model in our two-step inference to build a model with both a deep architecture for the stimulus component, and a network of coupling filters.

The model shows higher performance in predicting the PSTH: $\rho = 0.87 \pm 0.04$, versus $\rho = 0.82 \pm 0.05$ and $\rho = 0.81 \pm 0.05$, when compared to our previous models (Fig. .5B). In addition, the model was capable of predicting noise correlations with high accuracy (Fig. .5C, CoD$= 0.93$, versus CoD$= 0.94$ and CoD $= 0.95$). We also did not observe any self-excitation transient. In summary, the model combines the benefits of deep networks with those of the GLM with its neuronal couplings.

We summarise all the different model performances in Table 1.

8 Discussion

In this work we have studied the application of the GLM to the case of retinal ganglion cells subject to complex visual stimulation with strong correlations. We have shown how whole log-ℓ maximization over all model parameters leads to inferring erroneous coupling filters that reflect stimulus correlations (Fig. 1G). This effect introduces spurious noise correlations when the model is simulated (Fig. 1E&F),

	Checkerboard stimulus			Moving bars stimulus		
	PSTH	noise-corr.	self-exc.	PSTH	noise-corr.	self-exc.
whole log-ℓ maximization	0.82 ± 0.05	0.94	0%	0.71 ± 0.10	0.55	10%
two-step approach	0.81 ± 0.05	0.95	0%	0.72 ± 0.10	0.91	0%
coupl. exchange max logℓ	unstable	unstable	93%	0.65 ± 0.12	0.80	0%
coupl. exchange two-step	0.81 ± 0.05	0.84	0%	0.73 ± 0.09	0.91	0%
CNN	0.87 ± 0.04	0.93	0%	—	—	—

Table 1: **Model performance for different inference approaches.** We computed Pearson's correlation coefficients between empirical and model predicted firing rate (PSTH). For noise correlations, we estimated the CoD between model predictions and data. The third and forth rows refer to simulations that use the coupling filters inferred from the other stimulus.

prevents its generalization from one stimulus ensemble to another (Fig. 4E), and increases the chance of having self-excitation in the network dynamics (Fig. 1G). This last issue poses a major problem when the GLM is used as a generative model for simulating spiking activity.

To solve these issues we have proposed a two-step algorithm for inferring the GLM that takes advantage of repeated data to disentangle the stimulus processing component from the coupling network. A similar approach has been proposed in the context of maximum entropy models [25, 26], and here we have fully developed it for the GLM. Our method prevents the rise of large couplings reflecting strong stimulus correlations (Fig. 3D). The absence of these couplings lowers the probability of observing self-excitation (Fig. 3E) and the inferred GLM does not predict spurious noise correlations (Fig. 3B&C). In addition, with our two-step inference the couplings are robust to a change of stimulus, and allows for generalizations (Fig. 4). In particular we showed that a model with the stimulus filter inferred from checkerboard data but couplings inferred from moving bar stimulation predicts with high accuracy the response to checkerboard.

The strongest drawback of using our method is the requirement of repeated data, which are not necessary for whole log-ℓ maximization of GLM. This may limit the application of our inference approach. However we emphasize that only $165s$ of repeated data were needed for inferring the couplings. In addition, another possibility that deserves to be tested is the use of spontaneous activity instead of repeated stimuli. For the retina, this activity can be recorded while the tissue is exposed to a static full-field image (blank stimulus). However, as spontaneous activity is usually very low, these recordings need to be long enough to measure correlations with high precision.

Another important contribution of our work is the possibility to easily include deep CNNs into the GLM to increase its predicting power. Deep CNNs represent today one of the best options for modelling and predicting the mean response of sensory neurons to complex stimuli such as naturalistic ones [9, 16, 17, 18]. However, a generative model for predicting neuronal correlated noise with deep CNN is still lacking: building a deep network that would take as an input both the stimulus and the past activity of the neural population would be very challenging, since it would have to deal with very heterogeneous inputs. Our framework solves this problem by separating the CNN inference from that of coupling and spike-history filters, and can thus be easily added on an already inferred CNN.

The GLM has been used to estimate the impact of correlated noise on information transmission, but mostly for stimuli with low complexity [1, 27]. Future works can apply our method to model the responses to complex stimulations and study its impact on stimulus encoding.

Broader Impact

In this work we present a computational advance to improve the inference and performance of the GLM. As the GLM is one of the most used models in computational neuroscience, we believe that many researchers can benefit from this work to advance in their investigations. The fight against blindness, which affects about 45 millions people worldwide, is one of such possible applications. Retinal prostheses, where an array of stimulating electrodes is used to evoke activity in neurons, are a promising solution currently under clinical investigation. A central challenge for such implants is to improve the information that is sent to the brain. A central challenge for retinal implants is thus to mimic the computations carried out by a healthy retina to optimize information sent to the brain. Modeling retinal processing could thus help optimize vision restoration strategies in the long term [28].

We believe that no one will be put at disadvantage from this research, that there are no consequences of failure of the system. Biases in the data do not apply to the present context.

Acknowledgments

We thank M. Chalk and G. Tkacik for useful discussion. This work was supported by ANR TRAJECTORY and DECORE, by the European Union's Horizon 2020 research and innovation programme under grant agreement No 785219, No. 785907 (Human Brain Project SGA2) and No. 639888, a grant from AVIESAN-UNADEV to OM, by the French State program Investissements d'Avenir managed by the Agence Nationale de la Recherche [LIFESENSES: ANR-10-LABX-65], with the support of the Programme Investissements d'Avenir IHU FOReSIGHT (ANR-18-IAHU-01) and by Agence Nationale de la Recherche grant ANR-17-ERC2-0025-01 "IRREVERSIBLE".

References

[1] J.W. Pillow, J. Shlens, L. Paninski, A. Sher, A.M. Litke, E. J. Chichilnisky, and E.P. Simoncelli. *Spatio-temporal correlations and visual signalling in a complete neuronal population*. Nature , **454**:995–999, 2008.

[2] Alison I Weber and Jonathan W Pillow. Capturing the dynamical repertoire of single neurons with generalized linear models. *Neural computation*, 29(12):3260–3289, 2017.

[3] Baktash Babadi, Alexander Casti, Youping Xiao, Ehud Kaplan, and Liam Paninski. A generalized linear model of the impact of direct and indirect inputs to the lateral geniculate nucleus. *Journal of Vision*, 10(10):22–22, 2010.

[4] Subhodh Kotekal and Jason N MacLean. Recurrent interactions can explain the variance in single trial responses. *PLOS Computational Biology*, 16(1):e1007591, 2020.

[5] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

[6] Il Memming Park, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature neuroscience*, 17(10):1395, 2014.

[7] Caroline A Runyan, Eugenio Piasini, Stefano Panzeri, and Christopher D Harvey. Distinct timescales of population coding across cortex. *Nature*, 548(7665):92–96, 2017.

[8] Rajeev V Rikhye, Aditya Gilra, and Michael M Halassa. Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, 21(12):1753–1763, 2018.

[9] L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus. Deep learning models of the retinal response to natural scenes. In *Advances in Neural Information Processing Systems*, pages 1361–1369, 2016.

[10] A. Heitman, N. Brackbill, M. Greschner, A. Sher, A. M. Litke, and E.J. Chichilnisky. Testing pseudo-linear models of responses to natural scenes in primate retina. *bioRxiv*, 2016.

[11] Ryota Kobayashi, Shuhei Kurita, Anno Kurth, Katsunori Kitano, Kenji Mizuseki, Markus Diesmann, Barry J Richmond, and Shigeru Shinomoto. Reconstructing neuronal circuitry from parallel spike trains. *Nature communications*, 10(1):1–13, 2019.

[12] Il Memming Park, Evan W Archer, Nicholas Priebe, and Jonathan W Pillow. Spectral methods for neural characterization using generalized quadratic models. In *Advances in neural information processing systems*, pages 2454–2462, 2013.

[13] David Hocker and Il Memming Park. Multistep inference for generalized linear spiking models curbs runaway excitation. In *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 613–616. IEEE, 2017.

[14] Felipe Gerhard, Moritz Deger, and Wilson Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. *PLoS computational biology*, 13(2), 2017.

[15] Jose C Principe and Jyh-Ming Kuo. Dynamic modelling of chaotic time series with neural networks. In *Advances in neural information processing systems*, pages 311–318, 1995.

[16] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *BioRxiv*, page 201764, 2018.

[17] Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, and Surya Ganguli. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. In *Advances in Neural Information Processing Systems*, pages 8535–8545, 2019.

[18] Santiago A Cadena, Fabian H Sinz, Taliah Muhammad, Emmanouil Froudarakis, Erick Cobos, Edgar Y Walker, Jake Reimer, Matthias Bethge, Andreas Tolias, and Alexander S Ecker. How well do deep neural networks trained on object recognition characterize the mouse visual system? 2019.

[19] S. Deny, U. Ferrari, E. Mace, P. Yger, R. Caplette, S. Picaud, G. Tkačik, and O. Marre. Multiplexed computations in retinal ganglion cells of a single type. *Nature communications*, 8(1):1964, 2017.

[20] O. Marre, D. Amodei, N. Deshmukh, K. Sadeghi, F. Soo, T. Holy, and M.J. Berry. *Mapping a Complete Neural Population in the Retina*. Journal of Neuroscience , **32(43)**:1485973, 2012.

[21] Pierre Yger, Giulia LB Spampinato, Elric Esposito, Baptiste Lefebvre, Stéphane Deny, Christophe Gardella, Marcel Stimberg, Florian Jetter, Guenther Zeck, Serge Picaud, et al. A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. *ELife*, 7:e34518, 2018.

[22] I.H. Brivanlou, D.K. Warland, and M. Meister. *Mechanisms of Concerted Firing among Retinal Ganglion Cells*. Neuron , **20**:527–539, 1998.

[23] E.J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, 2001.

[24] Francois Chollet et al. Keras. `https://keras.io`, 2015.

[25] E. Granot-Atedgi, G. Tkacik, R. Segev, and E. Schneidman. Stimulus-dependent maximum entropy models of neural population codes. *PLOS Computational Biology*, 9(3):1–14, 2013.

[26] U. Ferrari, S. Deny, M. Chalk, G. Tkačik, O. Marre, and T. Mora. Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons. *Physical Review E*, 98(4):042410, 2018.

[27] F. Franke, M. Fiscella, M. Sevelev, B. Roska, A. Hierlemann, and R.A. da Silveira. Structures of neural correlation and how they favor coding. *Neuron*, 89(2):409–422, 2016.

[28] U. Ferrari, S. Deny, A. Sengupta, R. Caplette, J.-A. Sahel, D. Dalkara, S. Picaud, J. Duebel, and O. Marre. Optogenetic vision restoration with high resolution. *arXiv preprint, arXiv:1811.06866*, 2018. To appear in PLOS Computational Biology.

S1 Empirical data and correlations

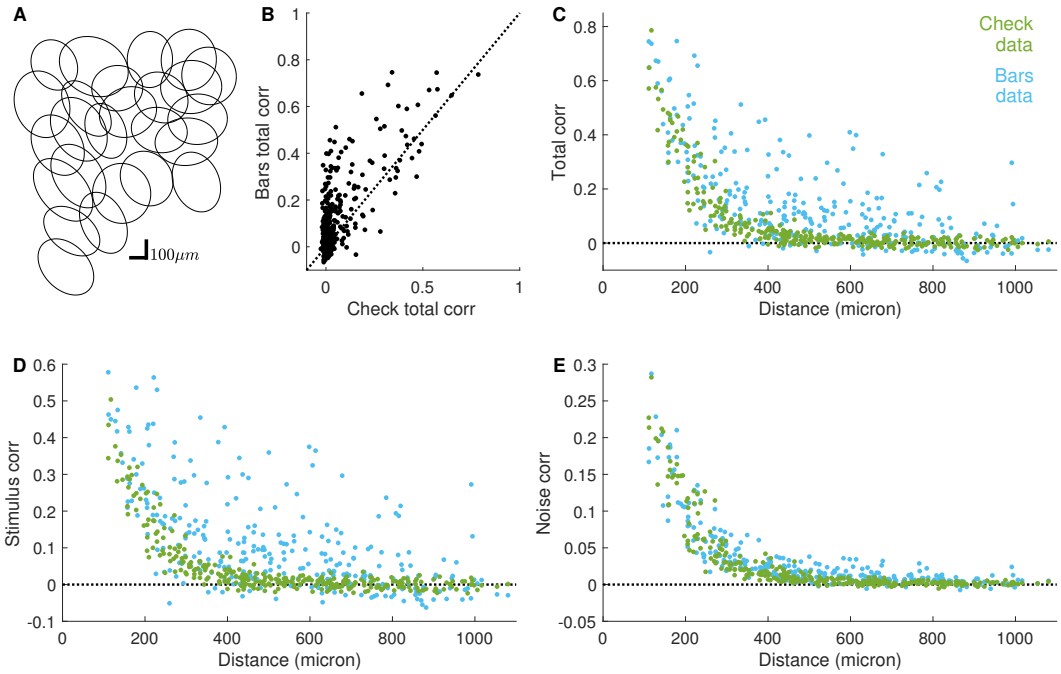

Figure S1: **Stimulus and noise correlation in the retinal response** A) Mosaic for $N = 25$ OFF alpha cells. B) Scatterplot of total pairwise correlation between the spiking activity in response to checkerboard and moving bars video. C) Total pairwise correlation versus cell distance D) Stimulus correlation versus cell distance E) Noise correlation versus cell distance

Responses to checkerboard and moving bars stimuli show different correlation patterns (Fig. S1). The moving bar video induces much stronger and long-ranged stimulus correlations, especially for certain pairs of distant cells. On the contrary, noise correlations decrease smoothly with distance and are of similar magnitude in the two datasets.

S2 Correction for the absolute refractory period

As explained in the main text, when we add the two-step coupling filters to the LNP model, we need to correct the h^i_{int} by its mean, Eq.4. However this correction does not take into account the addition of an absolute refractory period. In fact, if we start with an LNP model with rate $\lambda(t)$, and we prevent the cell to spike if it has spiked in the previous τ^i_{refr} time-bins during simulations, then the model rate will become a random variable itself with an average lower than $\lambda(t)$. In order to correct for this effect, we need first to quantify the mean of $n(t)$, the spike-count at time t:

$$\begin{aligned}
\mathbb{E}\left(\, n(t)\,\right) &= \mathbb{E}\left(n(t) \sim Pois(\lambda(t)) \,\middle|\, \sum_{\tau} n(t-\tau) = 0 \right) \\
&= \mathbb{E}\Big(n(t) \sim Pois(\lambda(t)) \Big) \text{ Prob} \left(\sum_{\tau} n(t-\tau) = 0 \right) \\
&\approx \mathbb{E}\Big(n(t) \sim Pois(\lambda(t)) \Big) \prod_{\tau} \text{Prob}\left(n(t-\tau) = 0 \,\right) \\
&= \lambda(t) \prod_{\tau} \exp\{-\lambda(t-\tau)\} \qquad (5)
\end{aligned}$$

where the approximation is valid under the hypothesis of small λ. By taking the log of Eq. 5, we obtain the correction term $\sum_{\tau} \lambda(t-\tau)$ that needs to be added to $h_{int}(t)$ in order to correct for the addition of the absolute refractory period.

S3 Generalization results for moving bar stimulus

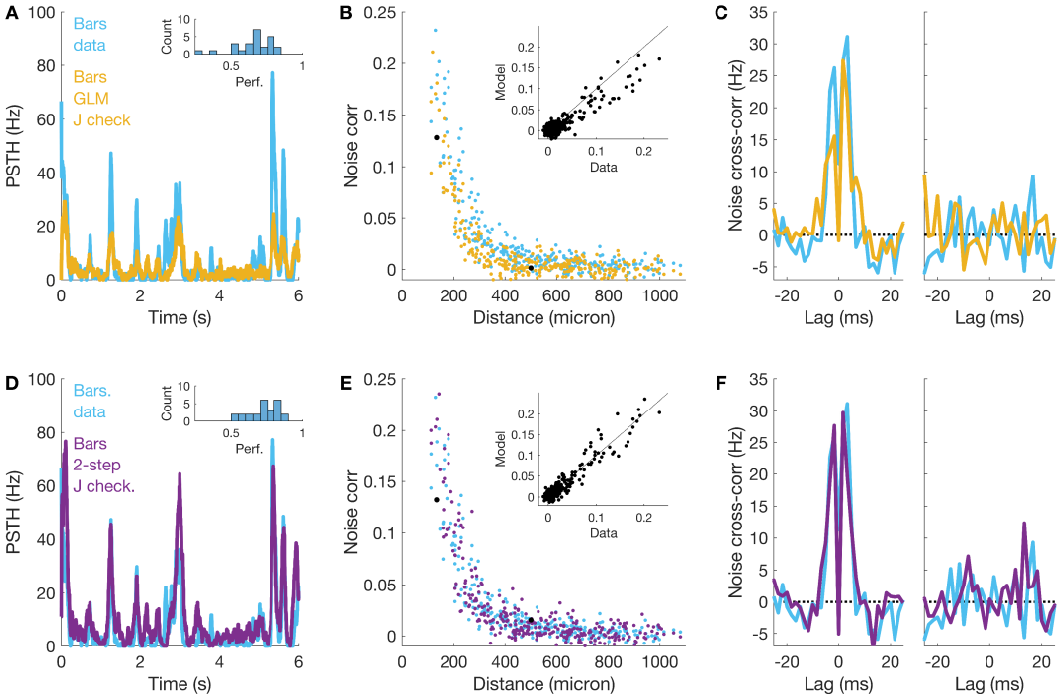

Figure S2: **Generalization results for moving bar stimulus** Simulation of the moving bar responses for a model where stimulus filters were inferred from the response to moving bar and couplings filter were inferred from the checkerboard data (opposite of Fig. 4) with whole log-ℓ maximization (A,B,C) and with our two-step inference (D,E,F). A,D) PSTH predictions. B,E) Noise correlations. C,F) Noise cross-correlation.