Multistep Inference for Generalized Linear Spiking Models Curbs Runaway Excitation

Accepted to 8th International IEEE EMBS Conference on Neural Engineering (NER 2017)

David Hocker and Il Memming Park

***Abstract*— Generalized linear models (GLMs) are useful tools to capture the characteristic features of spiking neurons; however, the predictive power of an autoregressive GLM inferred through maximum likelihood (ML) can be subject to runway self-excitation. We explain here that this runaway excitation is a consequence of the one-step-ahead ML inference used in estimating the parameters of the GLM. Alternatively, inference techniques that incorporate the likelihood of spiking multiple steps ahead in the future can remove this instability. We formulate a multi-step log-likelihood (MSLL) for interpreting spiking data. MSLL is used to infer an autoregressive GLM for individual spiking neurons recorded from the lateral intraparietal (LIP) area of monkeys during a perceptual decision-making task. While ML inference is shown to produce a GLM with poor fits of the neuron's interspike intervals and autocorrelation, in addition to its runaway excitation, MSLL fit models show a substantial improvement in interval statistics and stable spiking.**

I. INTRODUCTION

The generalized linear model (GLM) is a ubiquitous tool in neuroscience to describe spiking neurons, and can capture many important features observed in experimental data. In particular, autoregressive analogs of the Poisson-GLM can characterize self-modulating effects of a spiking neuron, such as refractoriness, self-excitation, and bursting [1]. An autoregressive GLM has the following general form:

$$P(y_t|\vec{Y}_{t-1},\theta) = \text{Poiss}(\lambda_t(\theta,\vec{Y}_{t-1})) \tag{1}$$

$$\lambda_t(\theta,\vec{Y}_{t-1}) = \exp\left(h^T\vec{Y}_{t-1}+a\right), \quad \theta=\{h,a\}, \tag{2}$$

where the probability of a spike occurring in a small time bin at time t is denoted as $P(y_t > 0|\vec{Y}_{t-1},\theta)$, and its mean is the conditional intensity function λ_t with parameters θ, given for the GLM as an autoregressive filter $h \in \mathbb{R}^p$ and a bias a. The filter is convolved with the previous p spikes $\vec{Y}_{t-1} = [Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}]^T$. Maximum likelihood (ML) estimation optimizes θ to maximize the one-step log-likelihood of spiking given the observed data $Y_1, \ldots, Y_T$,

$$\hat{\theta}_{ML} = \underset{\theta}{\text{argmax}} \sum_{t=1}^{T} \log P(y_t = Y_t|\vec{Y}_{t-1},\theta). \tag{3}$$

ML estimation yields a model that is "best" at one-step prediction, however, it is not necessarily the best for long-term prediction when run in the generative mode—feeding generated spikes back as history, recursively. In fact, there can be catastrophic failures where the generated spike trains do not resemble the statistics of the data. This typically manifests in an "unstable" and high firing rate due to positive feedback through a self-excitatory history filter. This is demonstrated in Fig 1A in the case of GLM fit to an individual neuron recorded from the lateral intraparietal (LIP) area of a monkey during a perceptual decision-making task [2]. Figure 1 shows the exponentiated autoregressive filter fit with ML, indicating significant self-excitation at both 3 ms and 60 ms after spiking. Generating spike trains from this GLM, though, can yield spikes trains that entirely saturate the firing rate by spiking as much as possible, which is shown by the raster plot of sampled data in Fig 1B. Moreover, the model fails to capture the long-term temporal structure of the recorded data. The model and data interspike interval (ISI) distribution and autocorrelation function are shown in Fig 2. The fast spiking due to overexcitation significantly shifts the ISI distribution of the GLM, and severely alters the autocorrelation.

There may be several solutions to address this issue with the GLM as a generative model. One proposed solution has been to incorporate a quadratic term into the autoregressive component, capable of reducing the effects of self-excitation [3], [4]. Unfortunately, restrictions on the required structure of such generalized quadratic models may artificially limit the applicability to fit real neural data. More importantly, this is an *ad-hoc* solution of examining new classes of models, while the primary difficulty is a mismatch in the method of inference used versus the predictive qualities desired from the model [5]. In that vein, we instead explore a method of multi-step inference aimed at improving the long-time predictive capabilities of the GLM by construction. This *multistep log-likelihood* (MSLL) inference technique is developed in the following section. We present its performance on capturing the statistics of neural data that were previously ill-described by the GLM with ML inference.

II. MULTISTEP LOG-LIKELIHOOD (MSLL) ESTIMATION

In general, we would like to identify a model that well-predicts spiking many steps ahead in the future. This objective can be encapsulated by the m-step observation likelihood $P(y_{t+m} = Y_{t+m}|\vec{Y}_t,\theta)$, which is the probability of observing y_{t+m} based on $[Y_t, \ldots, Y_{t-p}]$ without the knowledge of intermediate $m-1$ values. Let $\vec{y}_l = [y_{t+1}, \ldots, y_{t+m-1}]$ denote a latent vector for the intentionally missing values.

Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, NY 11794-5230, USA david.hocker@stonybrook.edu & memming.park@stonybrook.edu

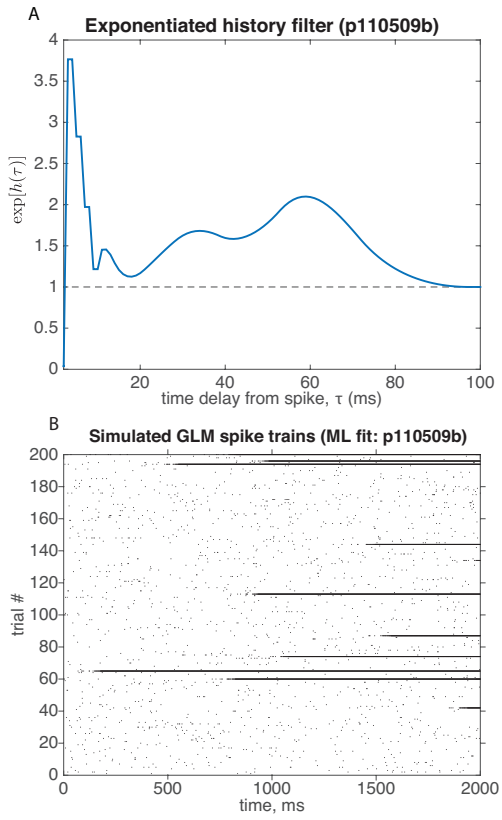

Fig. 1. Characteristic example of a GLM with over-excitation leading to saturated firing for long-time predictions. Model fit from data characterized in reference [2]. **(A)** Exponential gain from the autoregressive filter, showing initial refractoriness, followed by strong self-excitation. **(B)** Long-term sampling from the ML GLM, showing several instances of constant elevated firing rate.

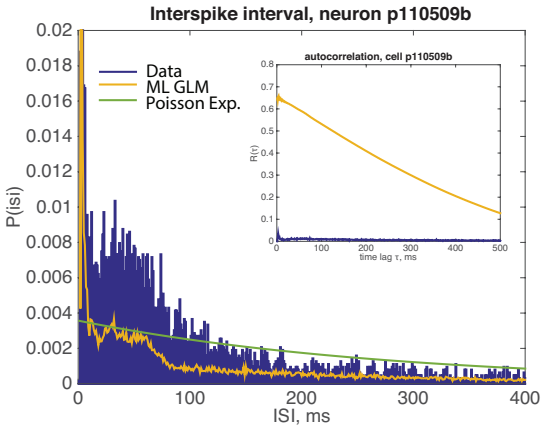

Fig. 2. Comparison of the interspike-interval (ISI) of the experimental and model data for an LIP neuron [2]. A one-parameter model fitting only average firing rate is shown in green (i.e., Poisson exponential) for comparison. Inset, autocorrelation comparison of the GLM and experimental data. In both statistics, the ML GLM fails to capture that temporal structure.

In order to incorporate inference at multiple step sizes, we define a weighted, multi-step log-likelihood,

$$L_{\text{MS}}(\theta, \rho) = \sum_{m=1}^{M} \rho_m \sum_{t=1}^{T} \log P(y_{t+m}|\vec{Y}_t, \theta) \tag{4}$$

where $\rho_m \geq 0$ weights the m-step likelihood. For ease of notation, we will use $P(y_{t+m}|\cdot) \equiv P(y_{t+m} = Y_{t+m}|\cdot)$ to denote the observation likelihood unless otherwise specified. From the definition of conditional probability, we express the expected m-step marginal log-likelihood as,

$$\begin{aligned}\mathcal{F} &= \log P(y_{t+m}|\vec{Y}_t, \theta) = \underset{\mathcal{P}}{\text{E}} \log \left[\frac{P(y_{t+m}, \vec{y}_l|\vec{Y}_t, \theta)}{P(\vec{y}_l|Y_{t+m}, \vec{Y}_t, \theta)}\right] \\ &= \underset{\mathcal{P}}{\text{E}}[\log P(y_{t+m}|\vec{y}_l, \vec{Y}_t, \theta)] + \underset{\mathcal{P}}{\text{E}}[\log P(\vec{y}_l|\vec{Y}_t, \theta)] \\ &\quad - \underset{\mathcal{P}}{\text{E}}[\log P(\vec{y}_l|Y_{t+m}, \vec{Y}_t, \theta)]. \end{aligned} \tag{5}$$

Note that (5) holds for any $\vec{y}_l$, hence the last equality holds for any distribution $\mathcal{P}$ over the latents $\vec{y}_l$.

Eq. (5) contains three respective expected log-likelihood terms: The m-step reconstruction error of the model, a "prior" likelihood of the latent spike patterns $\vec{y}_l$, and a "posterior" likelihood of the latents spikes given both prior spike history and the future spike at Y_{t+m}. In the spirit of the expectation-maximization (EM) algorithm [6], we rewrite the last two log-likelihood terms in (5) explicitly as Kullback-Leibler (KL) divergences,

$$\begin{aligned}\mathcal{F} =& \underset{\mathcal{P}}{\text{E}}[\log P(y_{t+m}|\vec{y}_l, \vec{Y}_t, \theta)] \\ &- D_{KL}[\mathcal{P}||P(\vec{y}_l|\vec{Y}_t, \theta)] - H(\mathcal{P}) \\ &+ D_{KL}[\mathcal{P}||P(\vec{y}_l|Y_{t+m}, \vec{Y}_t, \theta)]P + H(\mathcal{P}) \qquad (6) \\ =& \underset{\mathcal{P}}{\text{E}}[\log P(y_{t+m}|\vec{y}_l, \vec{Y}_t, \theta)] - D_{KL}[\mathcal{P}||P(\vec{y}_l|\vec{Y}_t, \theta)] \\ &+ D_{KL}[\mathcal{P}||P(\vec{y}_l|Y_{t+m}, \vec{Y}_t, \theta)]. \qquad (7)\end{aligned}$$

As in EM inference, we can alternate between (E-step) updating the distribution $\mathcal{P}$ to be the posterior distribution given the parameters,

$$\mathcal{P}(\vec{y}_l) = P(\vec{y}_l|Y_{t+m}, \vec{Y}_t, \theta) \quad \text{(latent distribution)}, \tag{8}$$

which removes the gap $D_{KL}\left[\mathcal{P}||P(\vec{y}_l|Y_{t+m}, \vec{Y}_t, \theta)\right] = 0$, and (M-step) maximizing the parameters with respect to the remaining terms with a fixed $\mathcal{P}$,

$$\mathcal{F}' = \underset{\mathcal{P}}{\text{E}}[\log P(y_{t+m}|\vec{y}_l, \vec{Y}_t, \theta)] - D_{KL}[\mathcal{P}||P(\vec{y}_l|\vec{Y}_t, \theta)]. \tag{9}$$

Equivalently, this amounts to maximizing the expected total data log-likelihood as a function of θ,

$$\mathcal{F}'' = \underset{\mathcal{P}}{\text{E}}[\log P(\vec{y}_l, y_{t+m}|\vec{Y}_t, \theta)]. \tag{10}$$

Unfortunately, (8) is a distribution over 2^m possible binary vectors, and a naive implementation incurs a prohibitive computational cost to evaluate the optimization objective (10) and its gradient. Therefore, we deploy a sampling based approximation of the expectation over the posterior distribution. We replace the expectation in (10) with a Monte Carlo

integration using L samples from $\mathcal{P}$,

$$L_{\text{MS}}(\theta,\rho) = \sum_{t,i,m=1}^{T,L,M} \frac{\rho_m}{L}[\log P(y_{t+m}|\vec{y}_{(i),l},\vec{Y}_t,\theta)+$$
$$\text{H}(m-2)\sum_{k=1}^{m-1}\log P(y_{(i),t+k}|(\vec{y}_{(i),t+k-1},\vec{Y}_t,\theta)] \quad (11)$$
$$= \sum_{t,i,m=1}^{T,L,M}\frac{\rho_m}{L}\left[Y_{t+m}\log(\lambda_{t+m}) - \lambda_{t+m} + \right.$$
$$\left.\text{H}(m-2)(M-m-1)\left[y_{t+m-1}\log(\lambda_{t+m-1}) - \lambda_{t+m-1}\right]\right].$$

λ_{t+m} is the conditional intensity function utilizing a sample of latent spikes from the posterior distribution $\mathcal{P}$,

$$\lambda_{t+m} = \exp\left(h^T[\vec{Y}_{t-1},\vec{y}_{(i),l}] + a\right), \quad (12)$$

$\vec{y}_{(i),l}$ is the i-th sample from the latent distribution with fixed θ, and $\text{H}(m-2)$ is a discrete Heaviside function. Eq. (11) can be used for stochastic gradient ascent if sampling from the posterior distribution is accessible. Here we turn to Gibbs sampling in order to draw samples from $\mathcal{P}$ [6]. The posterior can be decomposed using Bayes rule,

$$P(\vec{y}_l|Y_{t+m},\vec{Y}_t,\theta) \propto P(Y_{t+m}|\vec{y}_l,\vec{Y}_t,\theta)P(\vec{y}_l|,\vec{Y}_t,\theta), \quad (13)$$

where we may sample each successive latent spike y_l by determining its probability of being either a spike ($y_l = 1$) or not ($y_l = 0$) given the later observation Y_{t+m}. We initialize the sampling with the originally observed data, and then perform stochastic optimization using Adam gradient ascent [7] with $L = 1$ posterior sample per gradient step.

III. RESULTS

We first show the results of fitting MSLL models from LIP neurons that had previously shown unstable and constant firing in GLM sampling, as shown in Fig. 1. Figure 3A shows the exponentiated, autoregressive history filters for varying-length MSLL inference. In these model fits, $\rho_m = 1,\ \forall m$ to equally weight all step sizes in the inference. The initial, sharp excitation seen in the ML GLM history filter dissipates for longer-step inference, while the longer timescale excitation remains at a lower magnitude. However, the absolute refractory period seen in the GLM additionally disappears for longer inference. Figure 3B shows the ISI distributions for the MSLL fit models, which perform substantially better compared to the ML GLM, with models using larger M values providing similar fits to data as summarized by the Kolmogorov–Smirnov (KS) statistic of the ISI distribution shown in the inset. While the absolute refractory period is not as intensely encoded in the filter as in the ML model, lower spike timings are still reasonably captured (see zoomed-in distribution inset).

In addition to capturing some of the temporal structure of the observed LIP data, they importantly constitute a generative model that stabilizes the unstable firing seen in the GLM. This is demonstrated by comparing a raster plot of original LIP data (Fig. 4A) to raster plots of sampling from the MS 20 model, shown in Fig. 4B.

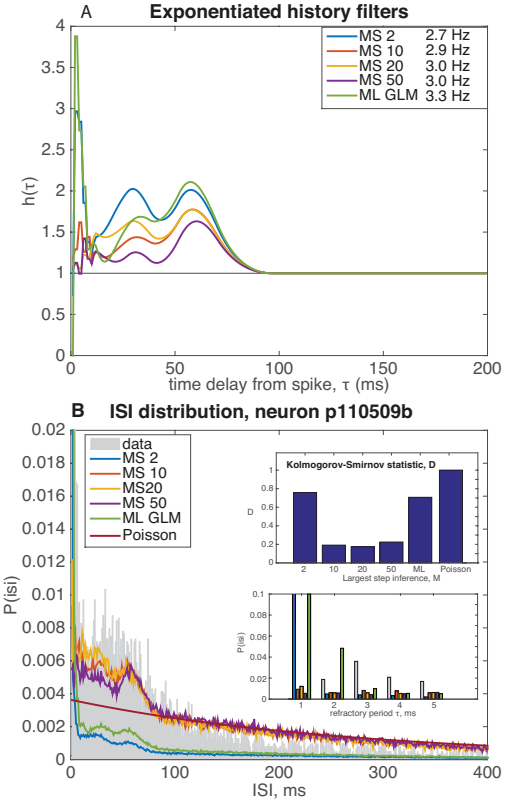

Fig. 3. **(A)** Exponential gain of history filters for MSLL inference at varying step sizes. As longer timestep inference is included in the models, refractoriness and the intensity of the initial self-excitation is diminished in the filter while maintaining the longer-time scale excitation near 60 ms. Bias parameter of each model given in legend. **(B)** Interspike interval (ISI) distributions for MSLL inference. Inset distribution: a zoomed-in view of low ISI. MSLL models with large M show substantial qualitative improvement upon capturing the ISI as compared the ML solution. Bar graph inset: the two-sample KS statistic comparing the data ISI distribution to model distributions, again showing improved fit with larger M value multi-step models.

We also investigated how the prediction performance of multi-step inference models changed with different step sizes. Plotted in Figure 5 is the m-step-ahead observation log-likelihood for models inferred from four-fold cross validation for different step sizes. This m-step log likelihood is given by

$$L_{\text{mstep}}(m) = \sum_{t}^{T}\log P(y_{t+m}|\vec{Y}_t,\theta). \quad (14)$$

The maximum likelihood model showed higher log-likelihood for small step sizes, but failed quite poorly for long-step inference. Lower-step models (MS 2) also had quite poor long-step prediction quality, while longer-step models performed reasonably well at low inference (see Fig. 5 inset), and were superior at long-step prediction.

IV. DISCUSSION

Here we demonstrated that multi-step log-likelihood (MSLL) inference for autoregressive spiking models can be a useful tool to identify parameters of sequential generative models that are more appropriate for sampling longer

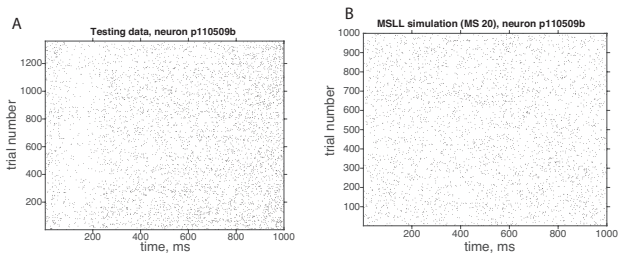

Fig. 4. **(A)** Raster plot of observed LIP data. **(B)** Raster plot of samples taken from the MS 20 model fit to the LIP neuron. No instances of elevated firing rates were observed from the MS 20-step model, as was observed in the ML model.

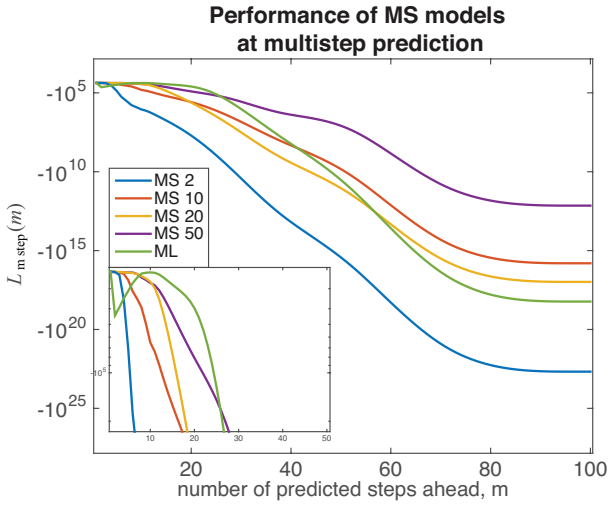

Fig. 5. Multi-step observation log-likelihood for varying step sizes with LIP testing data. Inset: zoomed-in view of small step log-likelihood. The ML solution performs well for small-step predictions, while multi-step models show better performance for longer predictions where ML solutions fail.

series than models fit with the conventional single-step log-likelihood. A set of MSLL fit models with sufficiently large multi-step inference, $M \geq 10$ in this work, generated stable firing rates and fit the interspike interval statistics of a neuron recorded from LIP during a perceptive decision making task, in contrast to an ML fit of the GLM that caused runaway self-excitation.

The computational demands of this method can grow quickly for large step-size predictions. Since Gibbs sampling is used to generate instances of these latent spikes, considerable computational demand is placed upon sampling from the latent posterior, which motivated our use of only one sample per optimization step.

Despite the computational considerations, this mode of inference is an attractive alternative to single-step log-likelihood, although it is not without its own caveats. The models fit to LIP data failed to capture the absolute refractory period for long time-step inference, which may be an important qualitative feature of a neuron. There is a trade-off between better long-term predictions and maintaining short-time scale temporal structure. Investigating the importance of the weighting factor ρ will be an important matter in that regard, as substantial importance should likely be placed on smaller-step inference when the refractory character of a neuron is desired. Crucially, though, the formalism presented here highlights the importance of the likelihood of spike trains generated from the model as a consideration when performing multi-step inference.

Finally, it is important to note that other approaches are tackling similar inference/sampling mismatching. Professor forcing [8] is a recent development meant to overcome a similar problem of generative models using recurrent neural networks for sequence generation. It is an improvement upon a paradigm called scheduled sampling [9], [10], in which ground truth observations as well as periodically introduced samples from the model itself are utilized in training. Rather than combining the samples into training a single model, though, professor forcing uses a generative adversarial network formalism to attempt to fool a discriminative model being given both recursively generated samples and the ground truth observations [11]. In this vein, professor forcing attempts to identify a model with reasonable long-term spike train generation, similar to MSLL. Future development of multistep inference will benchmark against professor forcing.

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*, vol. 454, no. 7207, pp. 995–999, Aug. 2008. [Online]. Available: http://dx.doi.org/10.1038/nature07140

[2] I. M. Park, M. L. R. Meister, A. C. Huk, and J. W. Pillow, "Encoding and decoding in parietal cortex during sensorimotor decision-making," *Nature Neuroscience*, vol. 17, no. 10, pp. 1395–1403, Oct. 2014. [Online]. Available: http://dx.doi.org/10.1038/nn.3800

[3] I. M. Park, E. Archer, N. Priebe, and J. W. Pillow, "Spectral methods for neural characterization using generalized quadratic models," in *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[4] D. Hocker and I. M. Park, "Instability of the generalized linear model for spike trains," in *Computational and Systems Neuroscience (COSYNE)*, submitted 2016.

[5] J. C. Príncipe and J.-M. Kuo, "Dynamic modelling of chaotic time series with neural networks," in *NIPS*, 1994.

[6] C. M. Bishop, *Pattern Recognitiion and Machien Learning*. Springer: Springer, 2006.

[7] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv e-prints*, Dec. 2014.

[8] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. Courville, and Y. Bengio, "Professor forcing: A new algorithm for training recurrent networks," in *Advances in Neural Information Processing Systems (NIPS)*.

[9] F. Huszár, "How (not) to Train your Generative Model: Scheduled Sampling, Likelihood, Adversary?" *ArXiv e-prints*, Nov. 2015.

[10] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks," *ArXiv e-prints*, June 2015.

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *ArXiv e-prints*, June 2014.