We thank the reviewers for their very useful comments.We are encouraged by reviewers thinking our work "is a novel contribution" (R1), its topic "is very relevant to GLMs" (R2) and "seems like an important advance" (R3). We are glad they thought our work has general significance in the neuroscience community at NeurIPS (R1) and many will find it useful (R2). We address the reviewers' comments below and we will incorporate all the feedback. Rebuttal Figure 1 shows some selected panels (for space reasons) from Figs. 3 and 4 that we hope illustrate how changes proposed by the reviewers will be incorporated. We will incorporate error bars by repeating the optimization procedure multiple times and we will add L2 regularization on the history filter to compare with our method.

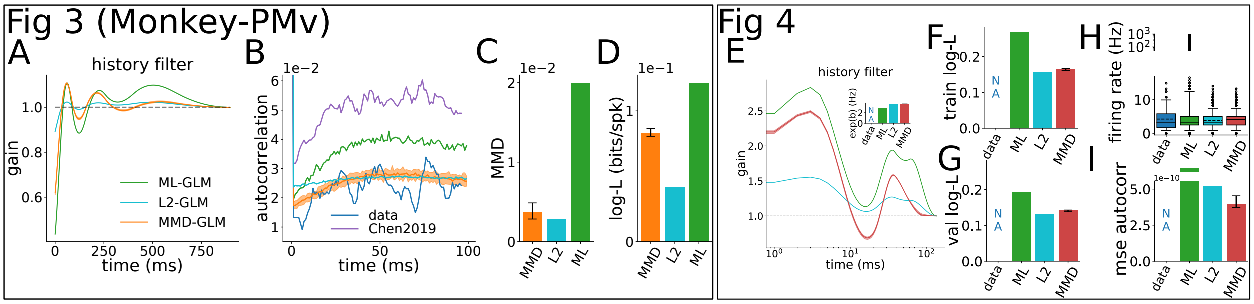
R2 suggested regularizing the history filter. As R2 noted, shrinkage of the history filter can lead to improved stability of the model. Rebuttal Figure 1 shows results for L2 regularization on the history filter. Results suggest that increasing α in L2 regularization to match the data firing rate or stabilize the model (see next response for the choice of regularization strength) in general leads to worse general performance than MMD regularization. In contrast, MMD regularization can be used to explicitly match any quantity and not just reduce the size of the history filter coefficients as in L2 regularization. MMD regularization yielded parameters with higher likelihood values (Rebuttal Figure 1 D, F) that also captured better the data autocorrelation (Rebuttal Figure 1 B, I) (data autocorrelation was smoothed to improve visuals). Note that the (model-based) MMD in Rebuttal Figure 1C is smaller for L2 regularization than for MMD regularization, but this is simply due to the L2 regularized model matching its free-running and data-conditioned distributions.

Choosing regularizer weight (α in NLL + αMMD) (R2). There are multiple possible strategies for choosing the hyperparameter α—one could focus on matching one of the key statistics of the free-running model (e.g. mean firing rate) with the data, or minimize the runaway excitation. For Figure 3 Monkey-PMv dataset we performed a grid search and chose the smallest α value such that the mean firing rate of 8000 samples differed less than 10% from the mean data firing rate. We used the same criteria to determine the regularizing weight in the added L2 regularization (Rebuttal Figure 1 A-D). For Figure 3 Human Ctx dataset we chose the smallest α such that out of 2400 samples none showed a diverging firing rate. For Rebuttal Figure 1 E-I we chose the smallest α such that out of 8000 samples none showed a diverging firing rate. We would state in the final version how we choose α for all the models in Figure 4.

Statistical significance and error bars (R1, R2). As pointed out by R1and R2, our estimation procedure is stochastic. Rebuttal Figure 1 shows panels from Figures 3 (Rebuttal Figure 1 A-D) and 4 (Rebuttal Figure 1 E-I) with the results of repeating the optimization procedure 20 times. For Figure 4 panels, we illustrate variability for one choice of kernel (same model-based kernel as in Figure 3, autocorrelation of the history filter convolved with spikes). The shading in the curves and the error bars in the barplots represent the 95% percentiles of the distribution obtained by repeating the optimization. The lines and bar heights represent the mean over optimization repetitions. We can see that for the examples shown here the stochastic optimization with the proposed MLE initialization is robust and similar parameters are found consistently. Similar robustness was observed when repeating the optimization procedure multiple times for the Human Cortex dataset of Figure 3 (not shown but would be incorporated in the final version).

Scaling (R3). Our procedure is definitely more expensive than MLE as it involves sampling and computing MMD during the optimization—in fact, we initialize the optimizer at the MLE. The time and space complexity varies with many factors; e.g. score function or model-based MMD, biased or unbiased estimator and the choice of the kernel—for some kernels, it is quadratic in the number of spikes, and for some other kernels, it only depends on the duration of recording. All the optimizations in the submission take as much as a minute to five in a desktop computer (compared with seconds for MLE) and there is room for improvement in the current implementation of the procedure. The authors have experiences with designing computationally efficient kernel and optimization tricks, and we plan to continue to improve the practicality of this novel concept.

Generating from ML fit discarding run-away excitation trials (R2). Removing "outlier" trials from the generative model has several disadvantages: (1) it can be very inefficient, (2) ML cannot guarantee that the truncation will result in a "good" distribution, and (3) interpretation of the model parameters is no longer straightforward.



Rebuttal Figure 1: **(A) Corresponds to 3B.. (B) Corresponds to 3C. (C, D) Correspond to 3D. (E) Corresponds to 4A. (F) Corresponds to 4B. (G) Corresponds to 4C. (H) Corresponds to 4D. (I) Corresponds to 4H.**