Probabilistic models for neural data
Session 3: Linear models and generalized linear models #1

To do before this session:

- **Read Park & Pillow (2011) and prepare assigned presentation**
- **Read up on this session's statistical concepts**
- **Complete pre-session quiz**

In this session we will first discuss Park & Pillow (2011) which applies linear Gaussian models to inferring neural receptive fields. Then, we will start looking at generalized linear models (GLMs). The previously discussed linear models assume a Gaussian likelihood, which implicitly assumes that the neural responses (e.g., the spike counts) are continuous, and can be negative. This is certainly not the case if these observations are indeed spike counts, which are non-negative integer values, and become binary values (spike/no spike, 1/0) if we consider very small time bins. Furthermore, responses might be non-linear functions of the stimulus. Adequate response likelihoods, together with non-linearities lead us to generalized linear models (not to be confused with general linear models, also abbreviated as GLMs).

To understand GLMs, we will first focus on likelihoods of interest to model neural responses, including the Bernoulli and the Poisson distribution (unfortunately, PRML does not introduce the Poisson distribution, but the discussed paper does). To introduce GLMs, we can initially treat them as linear classification models that classify small time bins by the absence or presence of a spike. These models can be interpreted as linearly separating the stimulus space into areas where the neuron is likely to spike, and areas where it is not. This can in turn be generalized to an inhomogeneous Poisson process (i.e., a Poisson process whose rate varies across time) that allows for a wider range of spike counts within larger time bins. (Advanced) For those mathematically inclined, likelihood functions can be further generalized to the exponential family of distributions that are tightly linked to the activation function used in GLMs.

As GLMs are a central concept in neural data analysis, we will spend two sessions on them. In this session we focus on modeling responses of single neurons, and an associated theoretical paper. In the next session, we focus on populations and their use for decoding, and discuss a paper that applies GLMs to neural population responses. All topics marked as "(Advanced)" are optional.

Paper: Park & Pillow (2011) Receptive field inference with localized priors

When reading the paper, please focus on the following:

- How/why does regularization improve test-set performance?
- Why did they use empirical Bayes rather than find the MAP estimate across both parameters and hyperparameters simultaneously, or use cross-validation to tune the hyperparameters?
- What were the features of the used prior? In particular,

- Why did they choose a Gaussian prior (rather than, e.g., a Student's T prior)?
    - What are the assumptions underlying the chosen prior? Are they realistic?
    - Why does a smaller a-priori variance induce stronger shrinkage?
    - Why did they tune the prior's parameters (i.e., the hyperparameters?)
- Which measures did they use to compare different models?
- What benefits do they gain from using empirical Bayes rather than the full Bayesian approach?

Don't worry too much about the exact mathematical details about the fixed point methods for ridge regression and ARD, how the ALDf and ALDsf are implemented, how the credible intervals are found, and about the MCMC approach underlying the full Bayesian approach. However, please make sure to understand the conceptual difference between the empirical Bayes and full Bayesian approach.

Presentations:

1. The neural encoding model and empirical Bayes inference (Fig. 1 & Eqs.)
    a. Using Fig. 1a & Eq. (1), explain the assumptions that the encoding model makes about how neural activity y_i depends on the stimulus vector x_i.
    b. Using Fig. 1b & Eqs. (2) & (8), explain the hierarchical Bayesian model. What is the prior and what does this prior assume? Assuming a diagonal $C(\theta)$, how would large/small values along its diagonal impact the receptive field estimates? What is the impact of the hyperparameters?
    c. Using Fig. 1c & Eqs. (7) & (10), explain how full Bayesian inference is approximated by empirical Bayes inference.
2. Previous methods (Fig. 2, left & center columns)
    a. Explain ridge regression and discuss why it gives better estimates than maximum likelihood.
    b. How does the ARD prior differ from that of ridge regression and how does it impact the estimates? Why can't the hyperparameters be estimated by cross-validation?
    c. What is the form of the Lasso prior, and how does its form impact the estimates? Why can, in contrast to ARD, the prior be estimated by cross-validation?
    d. What is the form of the ASD prior, and how does it impact the estimates? How does it qualitatively differ from all previously discussed priors?
3. Automatic locality determination (ALD; Fig. 2 right column & Fig. 3)
    a. Conceptually, what does the ALDs prior, Eq. (11) achieve? What is the effect of large/small C_{ii} values on the receptive field estimates? How is this reflected in the receptive field estimates in Figs. 2 & 3?
    b. Without worrying too much about the exact mathematical details of ALDf and how its hyperparameters are estimated, how does it qualitatively differ from ALDs? Under which circumstances would this lead to better estimates? How is this reflected in the receptive field estimates in Figs. 2 & 3?

    c. Again, without worrying too much about the math details of ALDsf, what are its qualitative features, and how do those impact the receptive field estimates in Figs. 2 & 3?

4. Application to simulated data (Figs. 4 & 5)
    a. Explain the simulations performed in Fig. 4, and interpret their results.
    b. For the white noise receptive field in Fig. 4F, why does the ‘correct’ ridge regression prior not outperform ASD & ALDf?
    c. Why is the estimation error expected to decrease with more training data (Figs. 5 A/C)? At which point would we expect the different prior types to achieve similar estimation errors?
5. Application to neural data (Figs. 6 & 7)
    a. Describe the results in Fig. 6. Why was it important to compute the estimate’s error on a hold-out dataset?
    b. Describe the estimates in Fig. 7. Can you identify parts of the ALDsf estimates in which the prior might ‘overregularize’ these estimates?
6. Empirical Bayes vs. full Bayesian approach (Figs. 9 & 10)
    a. Without going through the details of the math, describe the qualitative difference between empirical Bayes and the full Bayesian approach. Don’t worry too much about how the authors compute the credible intervals or perform full Bayesian inference by MCMC.
    b. Describe the results shown in Fig. 9. What are credible intervals and why are they of interest? What does the close match between full Bayesian and empirical Bayes credible intervals indicate? How do they differ and why?
    c. In Fig. 10, how do the estimates of the two approaches differ? Under which circumstances should we care about this difference and when can we ignore it?

Statistical concepts: generalized linear models (GLMs)

Probability distributions over binary random variables (PRML 2.0-2.1)
Binary random variables are essential to model spike trains, as, within short time bins, a neuron can either emit no spike, or a single spike. This behavior can be captured by the Bernoulli distribution, which is the simplest distribution for binary random variables, and is the basis for understanding the inhomogeneous Poisson process (see below). The beta distribution forms the conjugate prior to a Bernoulli likelihood, and is described in PRML 2.1.1 (you can skip the math details in this section, and revisit them later, if needed).

(Advanced) *The exponential family of probability distributions* (PRML 2.4)
The exponential family of probability distributions is an immensely useful family of distributions that includes the Gaussian, Bernoulli, Poisson, and many other frequently occurring probability distributions. They share some properties, and understanding this family provides a more general view on their properties. It also makes clear the generality of conjugate priors for likelihoods in this family. This section is a bit more technical - feel free to skip it.

Generalized linear models (PRML 4.0 & 4.3)
Generalized linear models generalize linear models to non-Gaussian likelihood functions, which show up when trying to model spike trains as Bernoulli random variables (spike/no spike) within each time bin, or inhomogeneous Poisson processes (when using larger time bins). For Bernoulli random variables, they can be thought of as binary classification models, that, for each time bin, model the probability of the presence (class 1) or absence (class 0) of a spike. Under certain assumptions (i.e., the commonly used exponential activation function), these models become a special case of logistic regression models, which are discussed in PRML 4.3.2. Most important in this section is to understand the difference between generative and discriminative models, which is described in PRML 4.0, as this distinction becomes relevant once we use GLMs for decoding (Session 4). You should also read through the different discriminant models (PRML 4.3), but can focus mostly on logistic regression (PRML 4.3.2). (Advanced) Note that IRLS (PRML 4.3.3) is the algorithm underlying most GLM implementations. (Advanced) The canonical link function (PRML 4.3.6) explains the relationship between the likelihood and activation function, and generalizes across different likelihood assumptions, including standard linear models with Gaussian likelihoods.

Unfortunately, PRML does not introduce the Poisson distribution or the homogeneous/inhomogeneous Poisson process. The next session's paper does so, but not in much detail. For more information on these topics, see Chapter 1 of Dayan & Abbott (2001) (see Session 1 notes for a link to the pdf).