Evaluation of Linear Optimization Algorithms for Estimating Spectro-Temporal Receptive Fields

F. E. THEUNISSEN[1], M. SCHACHTER[2], M. D. OLIVER [2]
[2]HWNI,[1]UC Berkeley, Berkeley, CA

Abstract

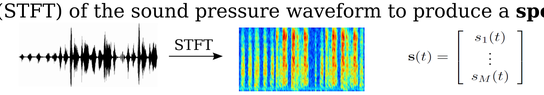
Stimulus-response functions are commonly described by a linear spatio-temporal receptive field, the STRF. The STRF is easily interpreted and can be estimated efficiently.

Here we analyze three methods for fitting STRFs: Ridge Regression, Threshold Gradient descent, and Least Angle Regression with Elastic Net (LARS-EN). Threshold gradient descent is a regularized gradient descent method and Least-Angle Regression is similar to linear regression with a Lasso penalty and L2 regularization.

Hyperparameters for the methods were fit using cross validation, and the integrated coherence ("normal mutual information") was used as a metric to judge goodness of fit for prediction datasets. Although a work-in-progress, LARS and Threshold Gradient Descent seem to outperform Ridge Regression.

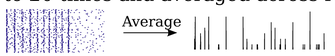
Introduction to STRF Fitting

This study focuses on the song system of the Zebra Finch. Sound pressure waveforms are converted into a time-frequency representation in the avian cochlea. We simulate this conversion by taking the Short-Time Fourier Transform (STFT) of the sound pressure waveform to produce a **spectrogram**:



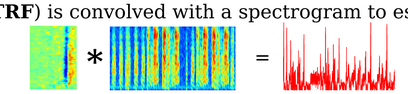
$$\mathbf{s}(t) = \begin{bmatrix} s_1(t) \\ \vdots \\ s_M(t) \end{bmatrix}$$

We present natural sounds (bird songs) while recording spikes. The stimulus is presented up to 20 times and averaged across responses to produce a **PSTH**:



$r(t)$

Given stimulus/response pairs, we want a model that predicts responses to novel stimuli. We use a linear model where a **spatio-temporal receptive field** (**STRF**) is convolved with a spectrogram to estimate the PSTH:



$$\hat{r}(t) = \mathbf{w} * \mathbf{s}(t) \qquad \mathbf{w} = \begin{bmatrix} w_{11} & \cdots & \cdots & w_{1P} \\ \vdots & \ddots & \ddots & \vdots \\ w_{M1} & \cdots & \cdots & w_{MP} \end{bmatrix}$$

To learn optimal STRFs we minimize a **sum-of-squares error function**:

$$E(\mathbf{w}) = \sum_{i=1}^{N} \int_t (r(t)^2 - \hat{r}(t)^2) dt$$

We make sure not to fit the noise, a phenomenon called **overfitting**, by using **regularization**. We add penalties to the error function, such as the **lasso (L1)**:

$$P_1(\mathbf{w}) = \sum_{i,j} |w_{ij}|$$

or **ridge regression (L2)**:

$$P_2(\mathbf{w}) = \sum w_{ij}^2$$

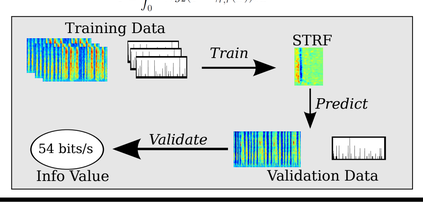
Lasso tends to shrink STRF weights to zero, while the L2 penalty tends to push STRF weights to common values. The penalty is multiplied by a user-set constant called a **hyperparameter**. We fit hyperparameters using **cross-validation**. The dataset is broken into K chunks, and the model is trained K times, each time leaving a chunk out to quantify the prediction goodness.

We use **coherence** to judge model goodness, equal to the cross-correlation between the Fourier transform of response and prediction:

$$|\gamma^2_{r,\hat{r}}(\omega)| = \frac{\langle R(\omega)\hat{R}^*(\omega)\rangle \langle R^*(\omega)\hat{R}(\omega)\rangle}{\langle R(\omega)R^*(\omega)\rangle \langle \hat{R}(\omega)\hat{R}^*(\omega)\rangle}$$

We integrate this quantity to get the **normal mutual information**:

$$I = \int_0^\infty log_2(1 - \gamma^2_{r,\hat{r}}(\omega)) d\omega$$

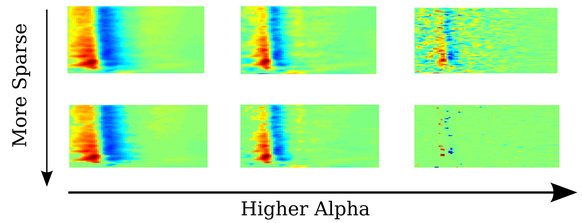


Direct Fit

DirectFit is an implementation of a technique called Ridge Regression. Ridge regression adds an L2 penalty to the normal least-squares error function:

$$E(\mathbf{w}) = \sum_{i=1}^{N} \int_t (r(t)^2 - \hat{r}(t)^2) dt + \alpha P_2(\mathbf{w})$$

Nonzero values of **alpha** push STRF weights to zero and prevent overfitting. After the STRF is found, we apply a sigmoidal mask controlled by a **sparseness** hyperparameter to reduce noise. Cross-validation is used to estimate the average prediction goodness across all sparseness+alpha combinations.



Threshold Gradient Descent (TG)

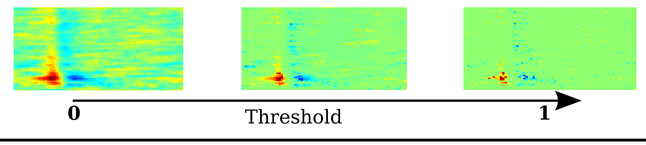
Gradient descent is an interative algorithm that updates the weights at every step by adding the negative of the error function's gradient. The **gradient** is a vector with element as the derivative of the error function with respect to a different STRF weight:

$$\nabla E_i = \frac{d}{dw_i} E(\mathbf{w})$$

The update step looks like this:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \xi \nabla E$$

A **threshold** hyperparameter between 0 and 1 is chosen. At each step, elements of the gradient which are less than threshold*max(gradient) are set to zero. Traditional gradient descent has threshold=0. **Coordinate Descent** is when threshold=1. Values between 0 and 1 produce results that are intermediate between ridge (0) and lasso (1). We use cross-validation to find an optimal value for the threshold hyperparameter.

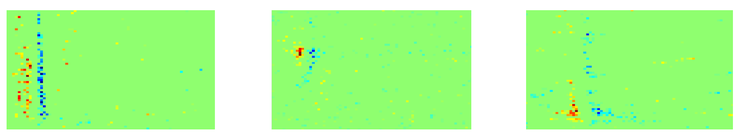


Least-Angle Regression (LARS)

LARS with Elastic Net (LARS-EN) combines ridge regression and lasso penalties, minimizing an error that looks similar to this:

$$E(\mathbf{w}) = \sum_{i=1}^{N} \int_t (r(t)^2 - \hat{r}(t)^2) dt + \lambda P_1(\mathbf{w}) + (1-\lambda) P_2(\mathbf{w})$$

LARS is an iterative algorithm, and on each step it maintains an **active set** of weights that are most correlated with the PSTH. Each step is taken in a direction that is equi-angular between weights in the active set, which prevents "greedy" solutions that overfit the data. The **lambda** hyperparameter controls the balance between lasso and ridge penalties. All of the STRFs produced by LARS are relatively sparse:

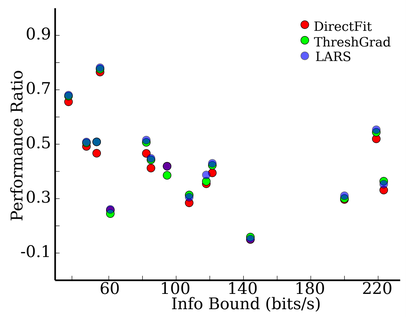


Results-in-Progress

LARS and Threshold Gradient Descent Outperform Direct Fit

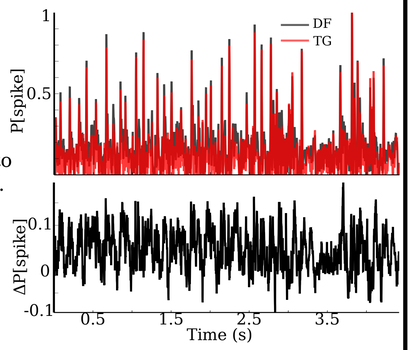
Coherence techniques allow us to put an upper bound on the information values for a spiking response. Here we plot the info captured for each method in the form of a ratio between the upper bound and model info.
These results show threshold gradient descent and LARS outperforming ridge regression.

DF: avg perf=0.42 +/- 0.16
TG: avg perf=0.43 +/- 0.16
LARS: avg perf=0.44 +/- 0.16



LARS and Threshold Gradient Descent Produce Less Noise

We have begun to investigate why LARS and TG outperform direct fit. The first clue is the difference in predictions between the methods. The right figure shows an example prediction (upper panel) for direct fit (black) and TG (red). TG seems to produce less low-probability spikes. When we subtract the TG response from direct fit response (lower), most of the difference is positive, clearly showing the TG STRF produces less noise in the low-probability range of the PSTH.



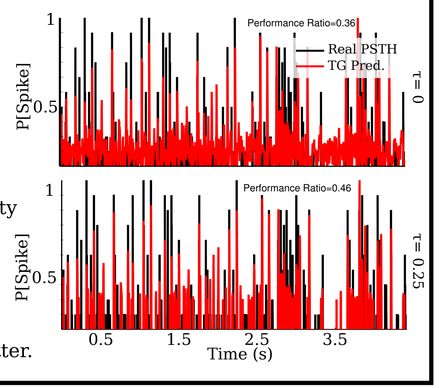
Simple Output Nonlinearity Improves Predictions

Knowledge that low-probability noise in the PSTH worsens our predictions suggests a simple output nonlinearity to improve predictions:

$$\hat{r}(t) = f(\mathbf{w} * \mathbf{s}(t))$$

$$f(x) = \begin{cases} x & if \quad x > \tau \\ 0 & otherwise \end{cases}$$

This threshold output nonlinearity significantly improves response predictions. We're currently searching for more analytically friendly (continuous, invertible, and parameterized) output non-linearities that perform even better.



References

Friedman J.H., Popescu B.E., (2004) Gradient Directed Regularization for Linear Regression and Classification http://www-stat.stanford.edu/~jhf/ftp/pathlite.pdf

Theunissen F.E., David S.V., Singh N.C., Hsu A., Vinje W.E., Gallant J.L. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli Network: Computation in Neural Systems 2001, Vol. 12, No. 3 : Pages 289-316

Theunissen, F. E., Amin, N., Shaevitz, S. S., Woolley, S. M. N., Fremouw, T. and Hauber, M. E. (2004), Song Selectivity in the Song System and in the Auditory Forebrain. Annals of the New York Academy of Sciences, 1016: 222–245. doi: 10.1196/annals.1298.023

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67: 301–320. doi: 10.1111/j.1467-9868.2005.00503.