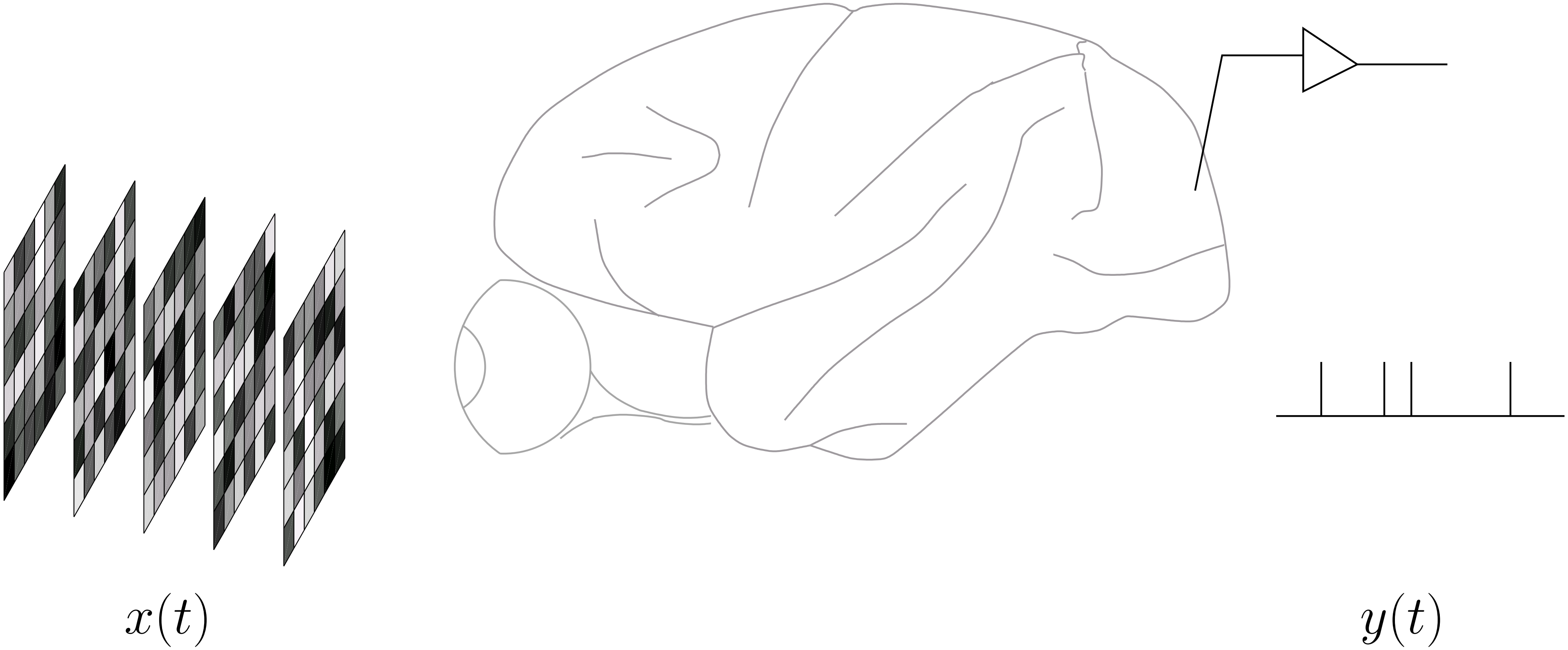
Neural Encoding Models

Maneesh Sahani
`maneesh@gatsby.ucl.ac.uk`

Gatsby Computational Neuroscience Unit
University College London

Term 1, Autumn 2011

Studying sensory systems



Decoding: $\hat{x}(t) = G[y(t)]$ (reconstruction)

Encoding: $\hat{y}(t) = F[x(t)]$ (systems identification)

General approach

Goal: Estimate $p(\text{spike}|x, H)$ [or $\lambda(t|x[0,t), H(t))$] from data.

- Naive approach: measure $p(\text{spike}, H|x)$ directly for every setting of x.
  - too hard: too little data and too many potential inputs.
- Estimate some functional $f(p)$ instead (e.g. mutual information)
- Select stimuli efficiently
- Fit models with smaller numbers of parameters

Spikes, or rate?

Most neurons communicate using action potentials — statistically described by a point process:

$$P\big(\text{spike} \in [t, t + dt)\big) = \lambda(t|H(t), \text{stimulus}, \text{network activity})dt$$

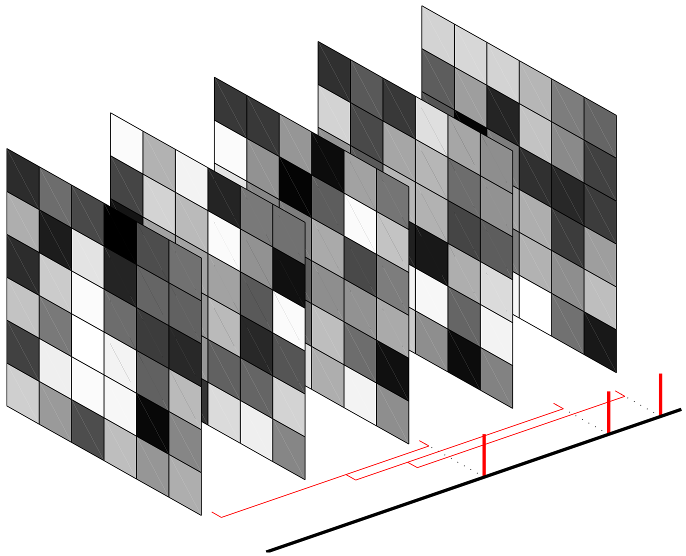
To fully model the response we need to identify λ. In general this depends on spike history $H(t)$ and network activity. Three options:

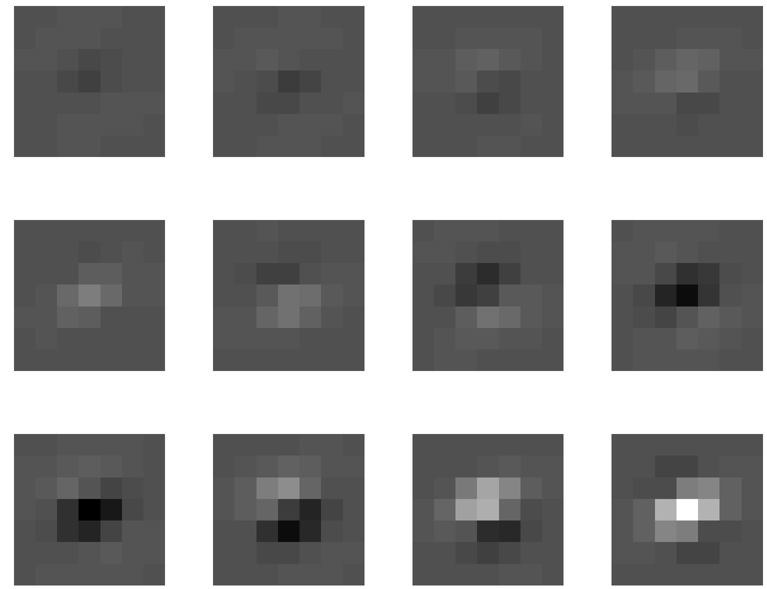
- Ignore the history dependence, take network activity as source of "noise" (i.e. assume firing is inhomogeneous Poisson or Cox process, conditioned on the stimulus).
- Average multiple trials to estimate

  $$\overline{\lambda}(t, \text{stimulus}) = \lim_{N\to\infty} \frac{1}{N} \sum_n \lambda(t|H_n(t), \text{stimulus}, \text{network}_n)$$

  the mean intensity (or PSTH), and try to fit this.
- Attempt to capture history and network effects in simple models.

Spike-triggered average





Decoding: mean of P $(x \mid y = 1)$

Encoding: predictive filter

Linear regression

$$y(t) = \int_0^T x(t-\tau)w(\tau)d\tau$$

$$\boxed{x_1 \;\; x_2 \;\; x_3 \;\; \ldots \;\; x_T \;\; x_{T+1} \;\; \ldots}$$

$$\begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_T \\ x_2 & x_3 & x_4 & \ldots & x_{T+1} \\ & & \vdots & & \\ & & & & \end{bmatrix} \times \begin{bmatrix} w_t \\ \vdots \\ w_3 \\ w_2 \\ w_1 \end{bmatrix} = \begin{bmatrix} y_T \\ y_{T+1} \\ \vdots \\ \\ \end{bmatrix}$$

$$XW = Y$$

$$W(\omega) = \frac{X(\omega)^* Y(\omega)}{|X(\omega)|^2}$$

$$W = \underbrace{(X^\mathsf{T} X)}_{\Sigma_{SS}}{}^{-1} \underbrace{(X^\mathsf{T} Y)}_{\text{STA}}$$

Linear models

So the (whitened) spike-triggered average gives the minimum-squared-error linear model.

Issues:

- overfitting and regularisation
  - standard methods for regression
- negative predicted rates
  - can model deviations from background
- real neurons aren't linear
  - models are still used extensively
  - interpretable suggestions of underlying sensitivity
  - may provide unbiased estimates of cascade filters (see later)

How good are linear predictions?

We would like an absolute measure of model performance.

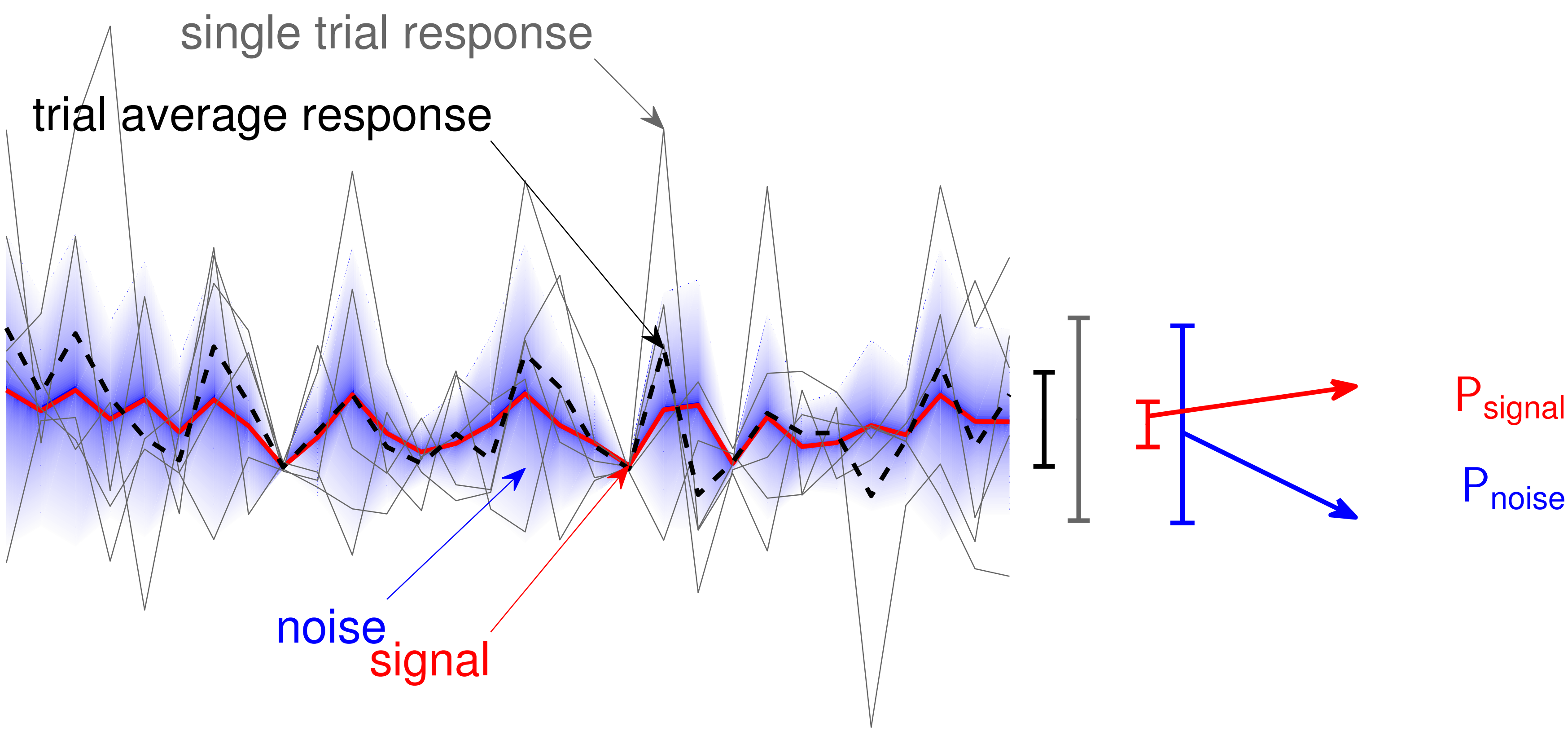
Measured responses can never be predicted perfectly:

- The measurements themselves are noisy.

Models may fail to predict because:

- They are the wrong model.
- Their parameters are mis-estimated due to noise.

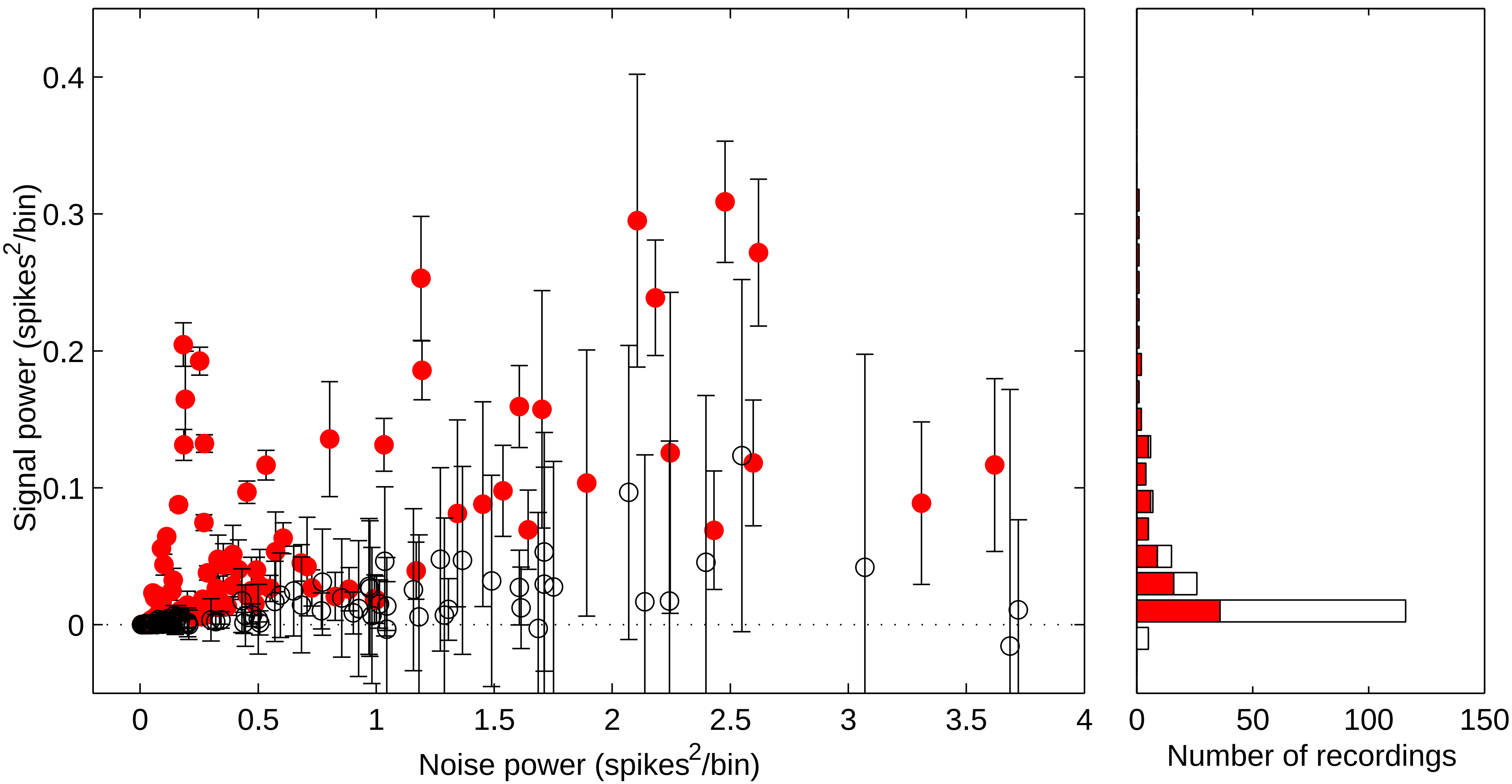
Estimating predictable power



$$\underbrace{\text{response}}_{\mathbf{r}^{(n)}} = \text{signal} + \text{noise}$$

$$\left.\begin{aligned} \overline{\mathsf{P}(\mathbf{r}^{(n)})} &= \mathsf{P}_{\text{signal}} + \mathsf{P}_{\text{noise}} \\ \mathsf{P}(\overline{\mathbf{r}^{(n)}}) &= \mathsf{P}_{\text{signal}} + \frac{1}{N}\,\mathsf{P}_{\text{noise}} \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} \widehat{\mathsf{P}}_{\text{signal}} &= \frac{1}{N-1}\left(N\mathsf{P}(\overline{\mathbf{r}^{(n)}}) - \overline{\mathsf{P}(\mathbf{r}^{(n)})}\right) \\ \widehat{\mathsf{P}}_{\text{noise}} &= \overline{\mathsf{P}(\mathbf{r}^{(n)})} - \widehat{\mathsf{P}}_{\text{signal}} \end{aligned}\right.$$

Signal power in A1 responses



Testing a model

For a perfect prediction

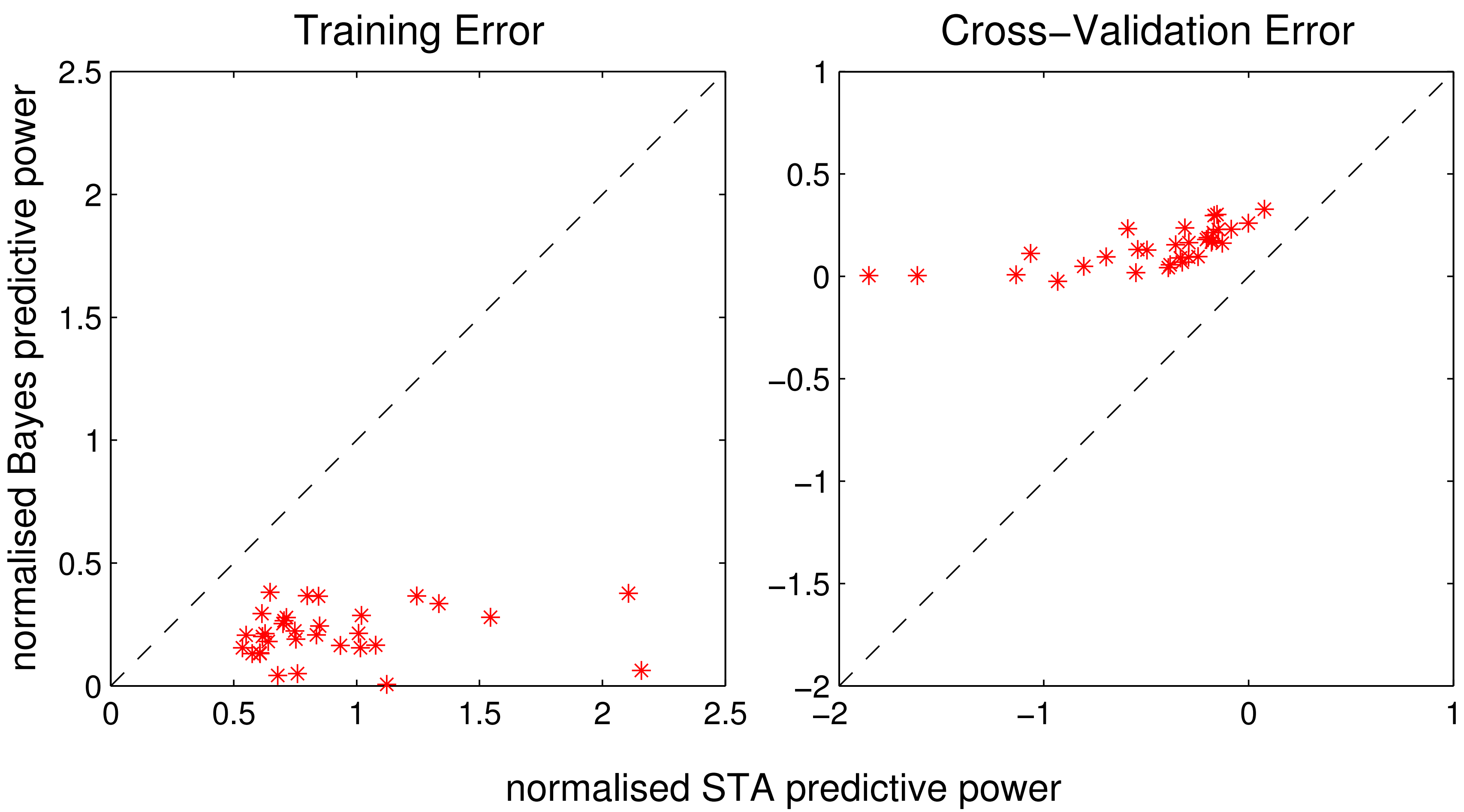
$$\left\langle \mathrm{P}(\overline{\mathrm{trial}}) - \mathrm{P}(\mathrm{residual}) \right\rangle = \mathrm{P}(\mathrm{signal})$$

Thus, we can judge the performance of a model by the normalized predictive power

$$\frac{\mathrm{P}(\overline{\mathrm{trial}}) - \mathrm{P}(\mathrm{residual})}{\widehat{\mathrm{P}}(\mathrm{signal})}$$

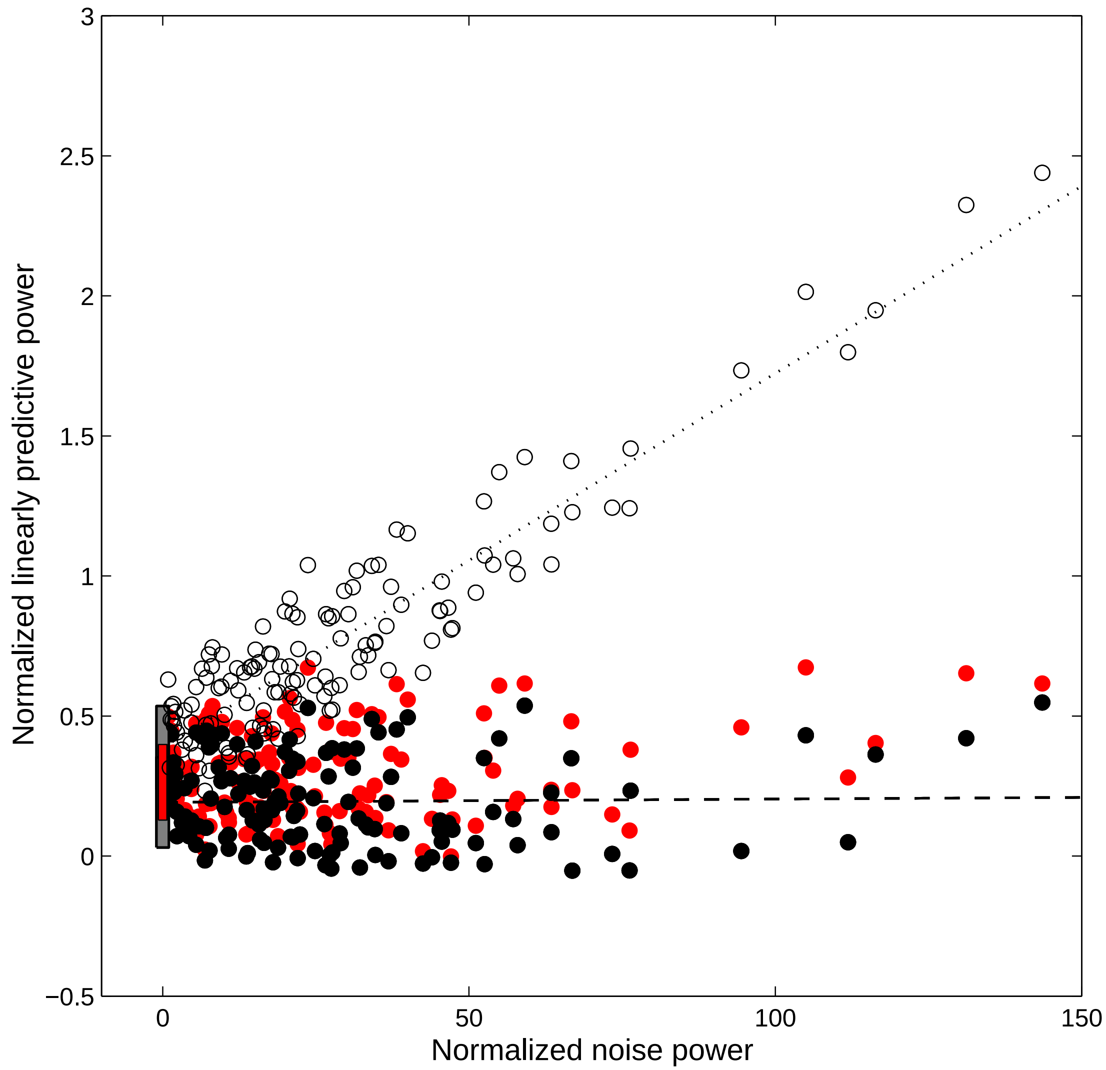
Similar to coefficient of determination (r^2), but the denominator is the predictable variance.

Predictive performance

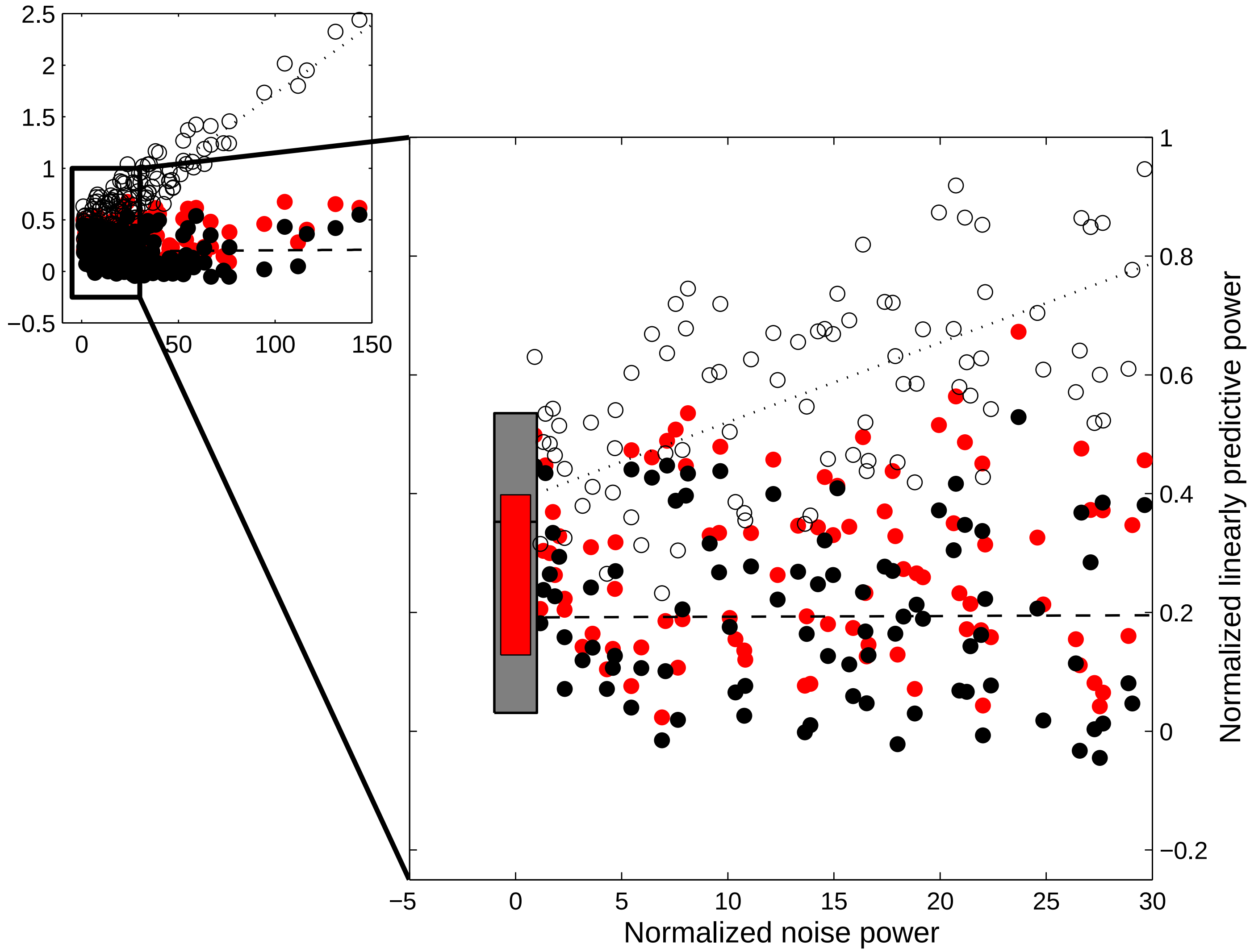


Extrapolating the model performance

Jackknifed estimates

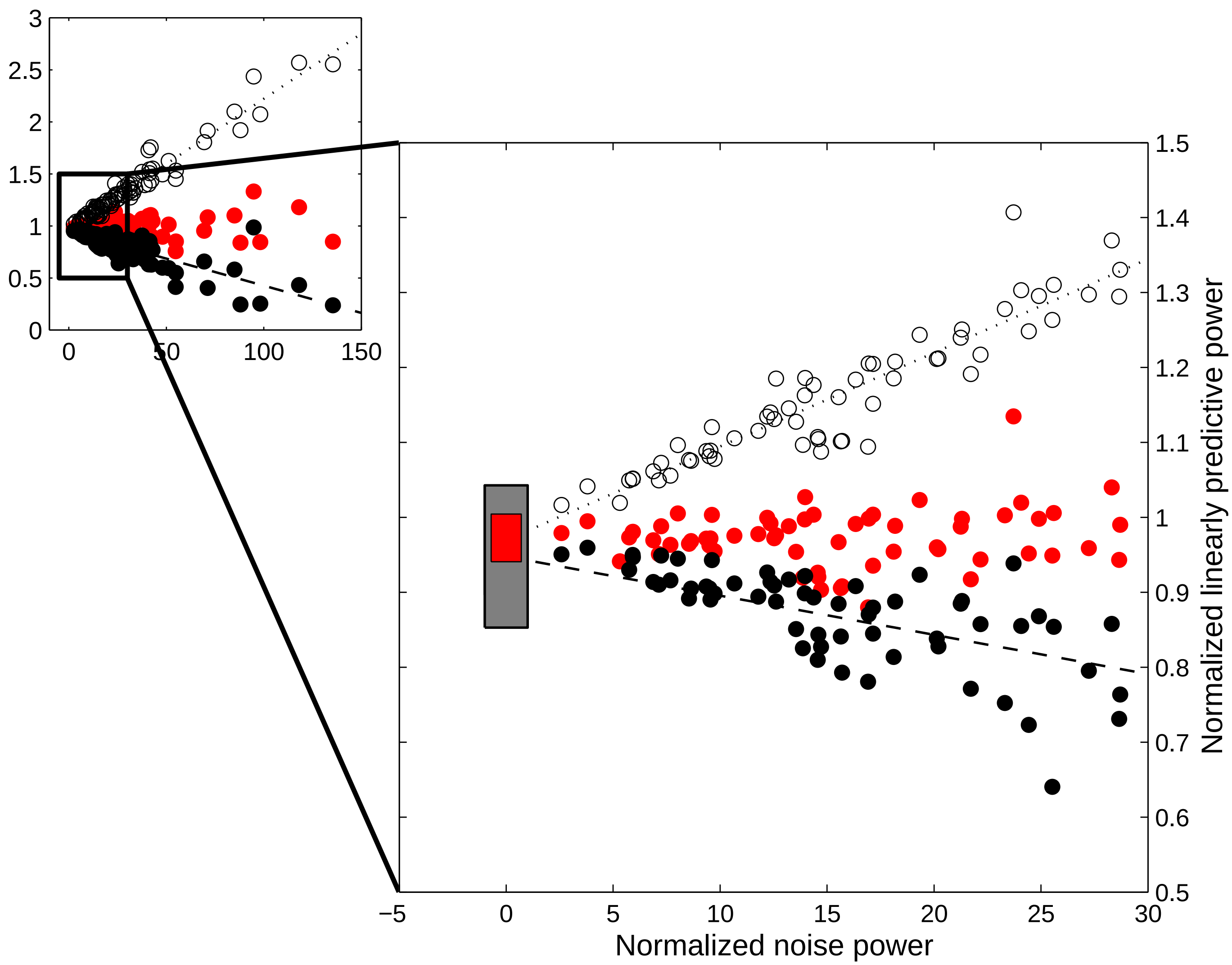


Extrapolated linearity



[extrapolated range: (0.19,0.39); mean Jackknife estimate: 0.29]

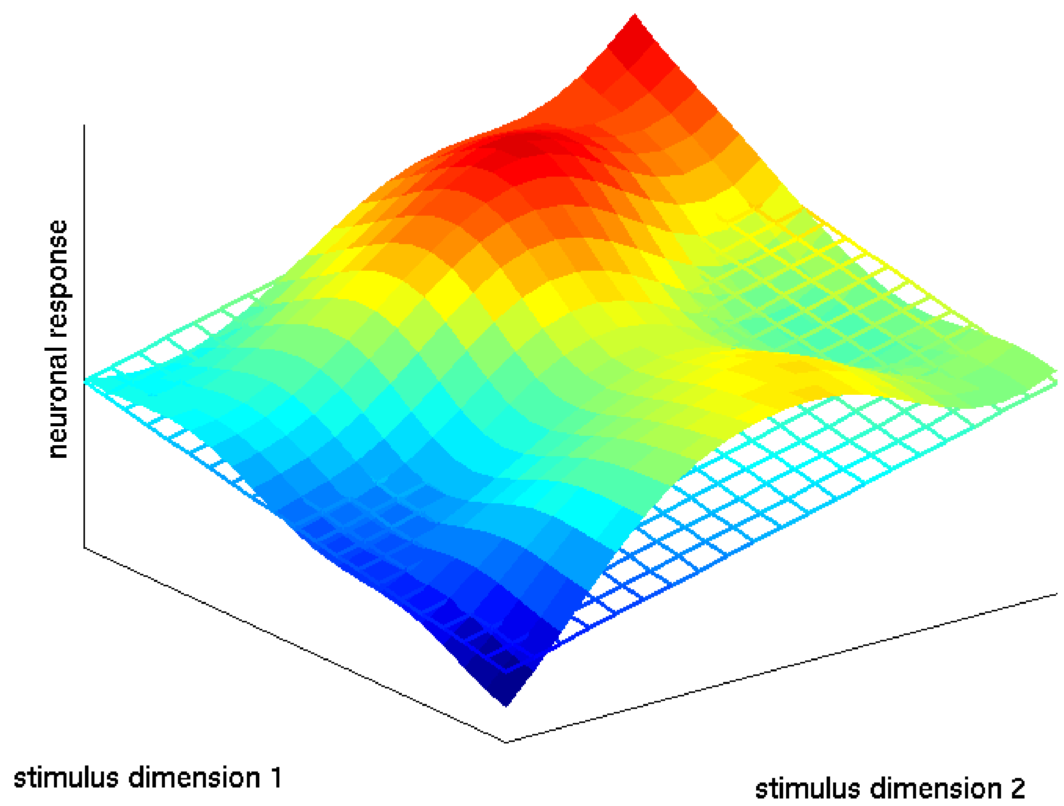
Simulated (almost) linear data



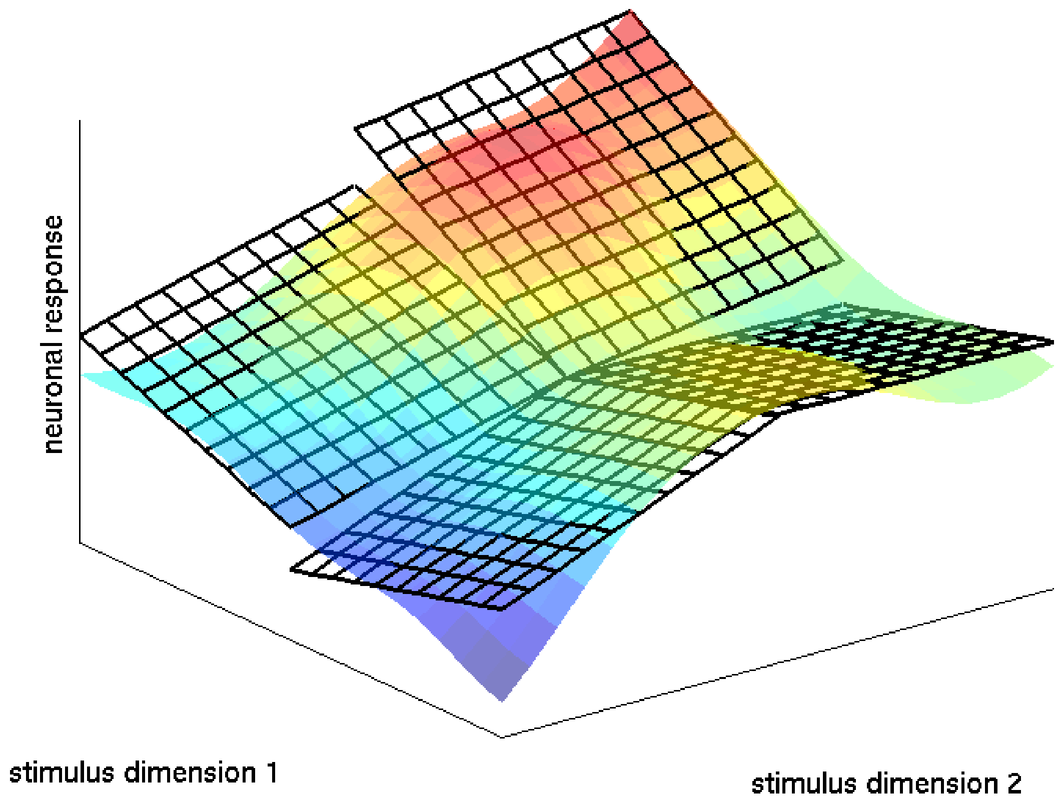
[extrapolated range: (0.95,0.97); mean Jackknife estimate: 0.97]

Linear fits to non-linear functions

Linear fits to non-linear functions



Approximations are stimulus dependent



(Stimulus dependence does not always signal response adaptation)

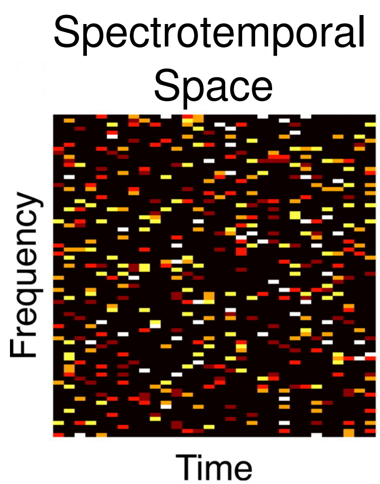
Consequences

Local fitting can have counterintuitive consequences on the interpretation of a “receptive field”.

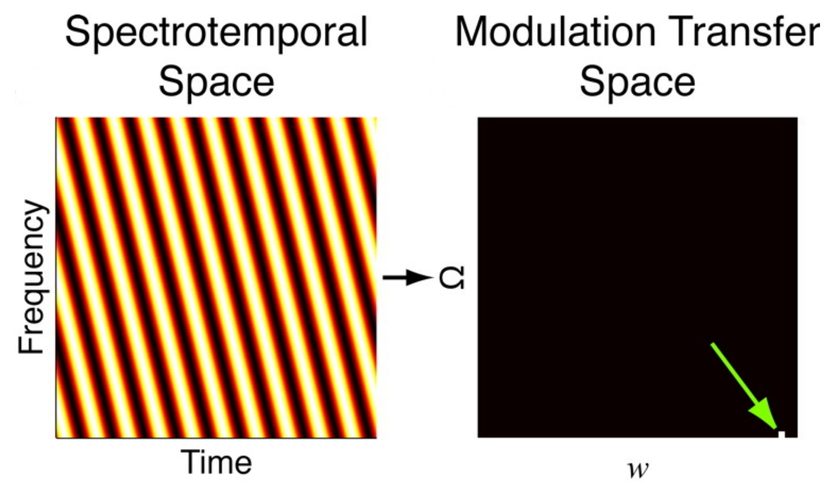
“Independently distributed” stimuli

Knowing stimulus power at any set of points in analysis space provides noinformation about stimulus power at any other point.

DRC:

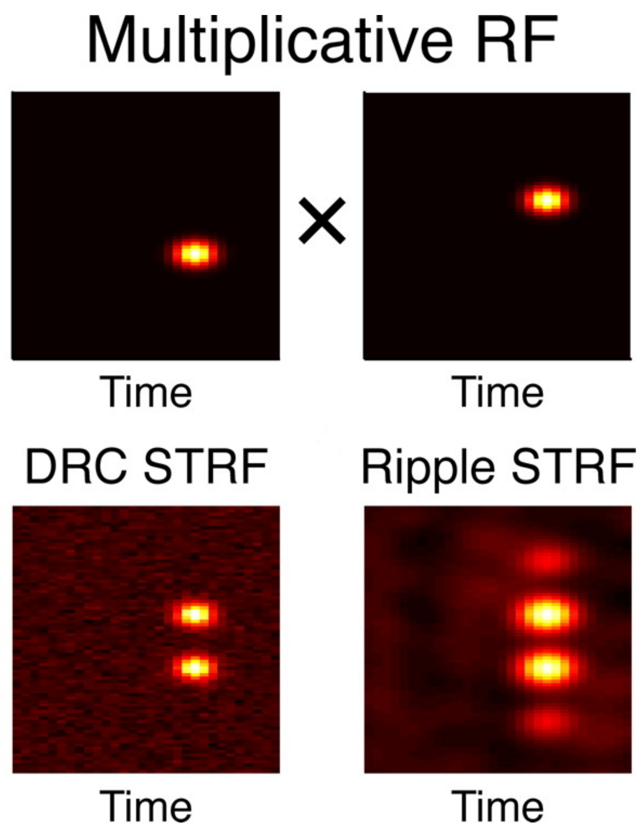


Ripple:



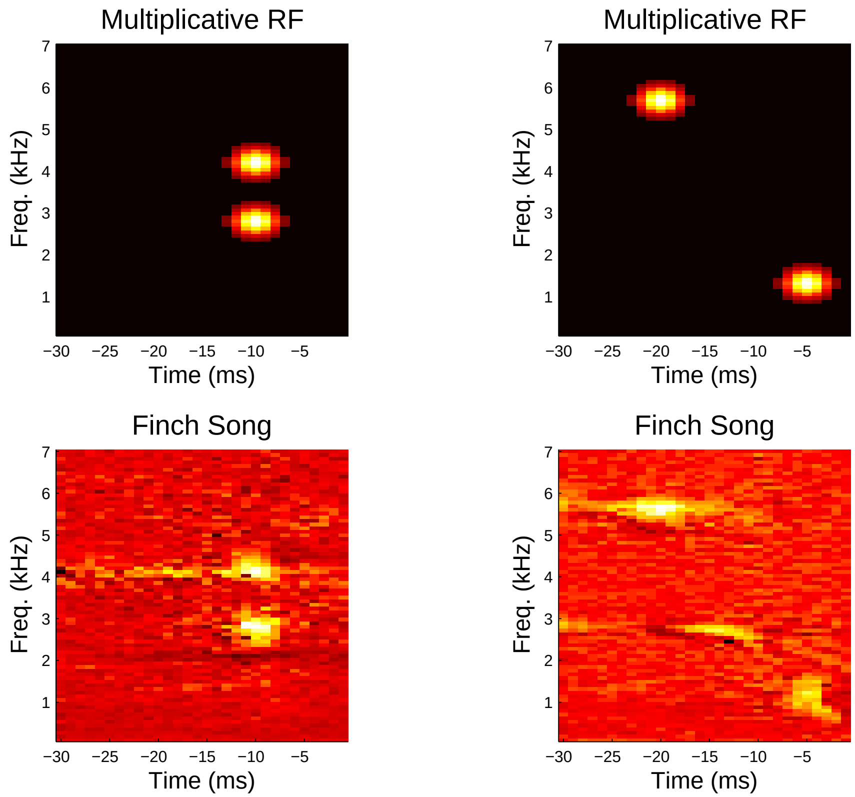
Independence is a property of stimulus *and* analysis space.

Nonlinearity & non-independence distort RF estimates



Stimulus may have higher-order correlations in other analysis spaces — interaction with nonlinearities can produce misleading “receptive fields.”

What about natural sounds?



Usually not independent in any space — so STRFs may not be conservative estimates of receptive fields.

Beyond linearity

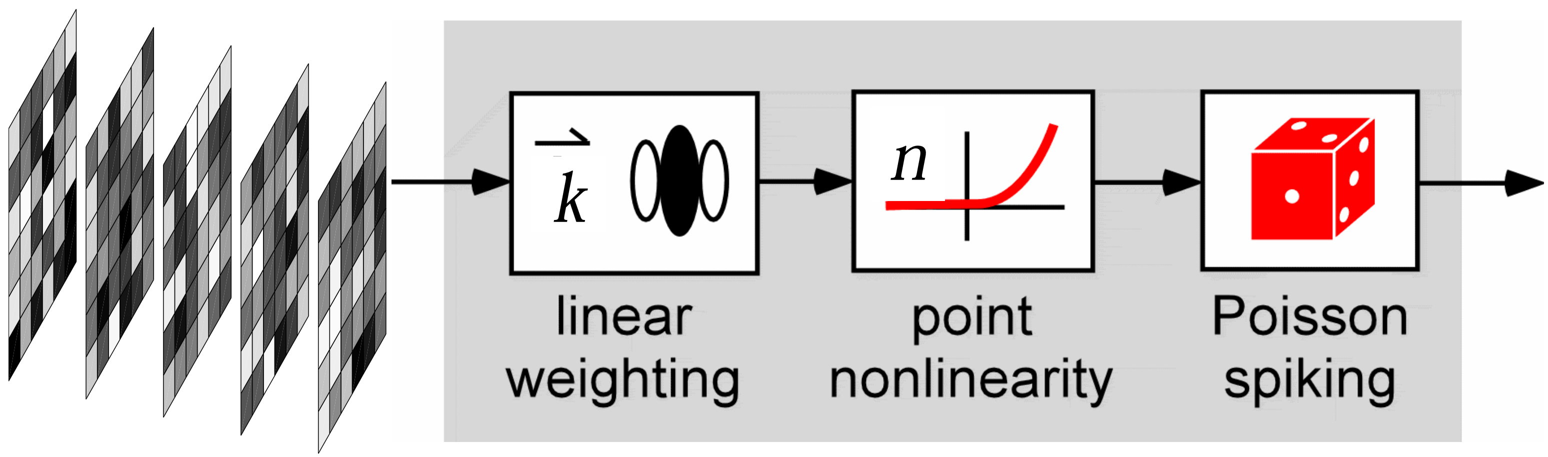
Beyond linearity

Linear models often fail to predict well. Alternatives?

- Wiener/Volterra functional expansions
  - M-series
  - Linearised estimation
  - Kernel formulations
- LN (Wiener) cascades
  - Spike-trigger covariance (STC) methods
  - "Maximimally informative" dimensions (MID) $\Leftrightarrow$ ML nonparametric LNP models
  - ML Parametric GLM models
- NL (Hammerstein) cascades
  - Multilinear formulations

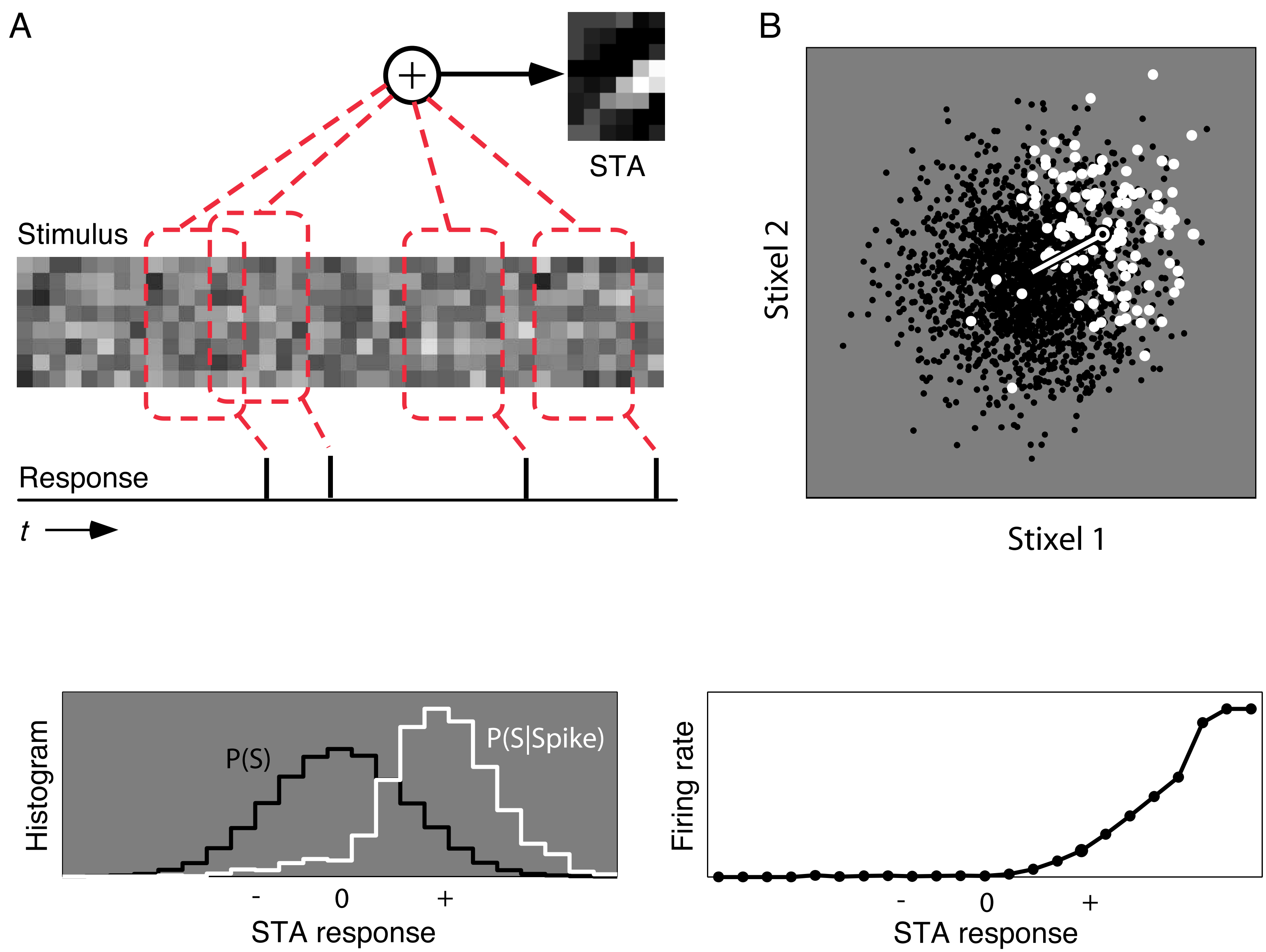
Non-linear models

The LNP (Wiener) cascade



Rectification addresses negative firing rates. Possible biophysical justification.

LNP estimation – the Spike-triggered ensemble



Single linear filter

STA.

Non-linearity.

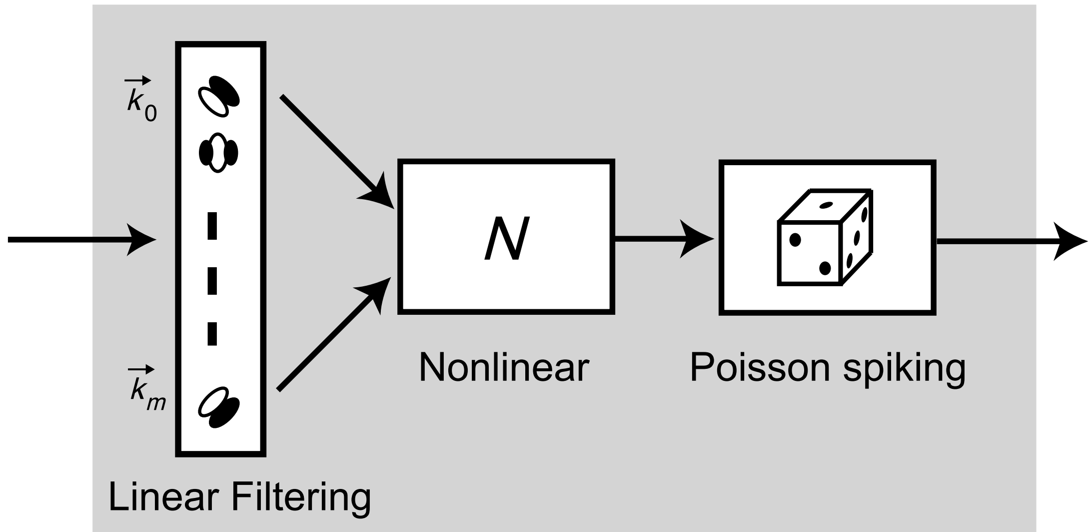
STA unbiased for spherical (elliptical) data.

Bussgang.

Non-spherical inputs.

Biases.

Multiple filters

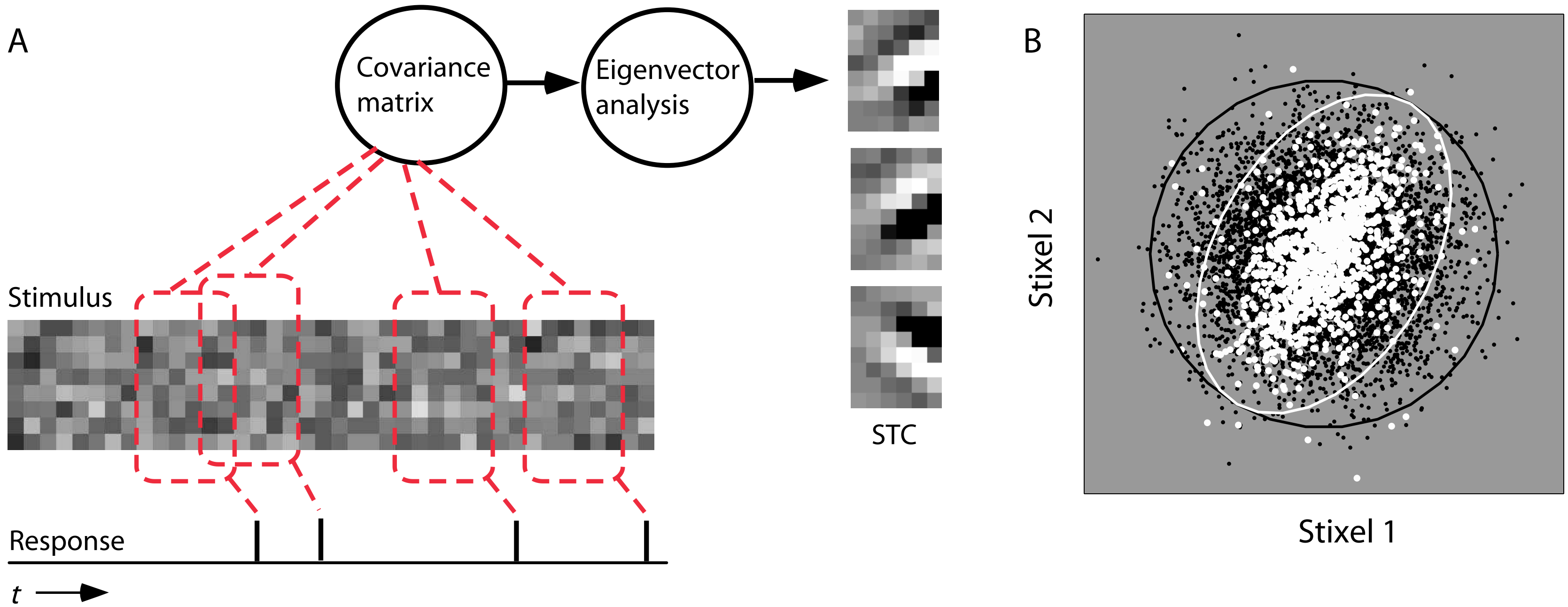


Distribution changes along relevant directions (and, usually, along all linear combinations of relevant directions).

Proxies for distribution:

- mean: STA (can only reveal a single direction)
- variance: STC
- binned (or kernel) KL: MID “maximally informative directions” (equivalent to ML in LNP model with binned nonlinearity)

STC



Project out STA:

$$\widetilde{X} = X - (X\mathbf{k}_{\text{sta}})\mathbf{k}_{\text{sta}}^{\mathsf{T}}; \quad C_{\text{prior}} = \frac{\widetilde{X}^{\mathsf{T}}\widetilde{X}}{N}; C_{\text{spike}} = \frac{\widetilde{X}^{\mathsf{T}}\text{diag}(Y)\widetilde{X}}{N_{\text{spike}}}$$

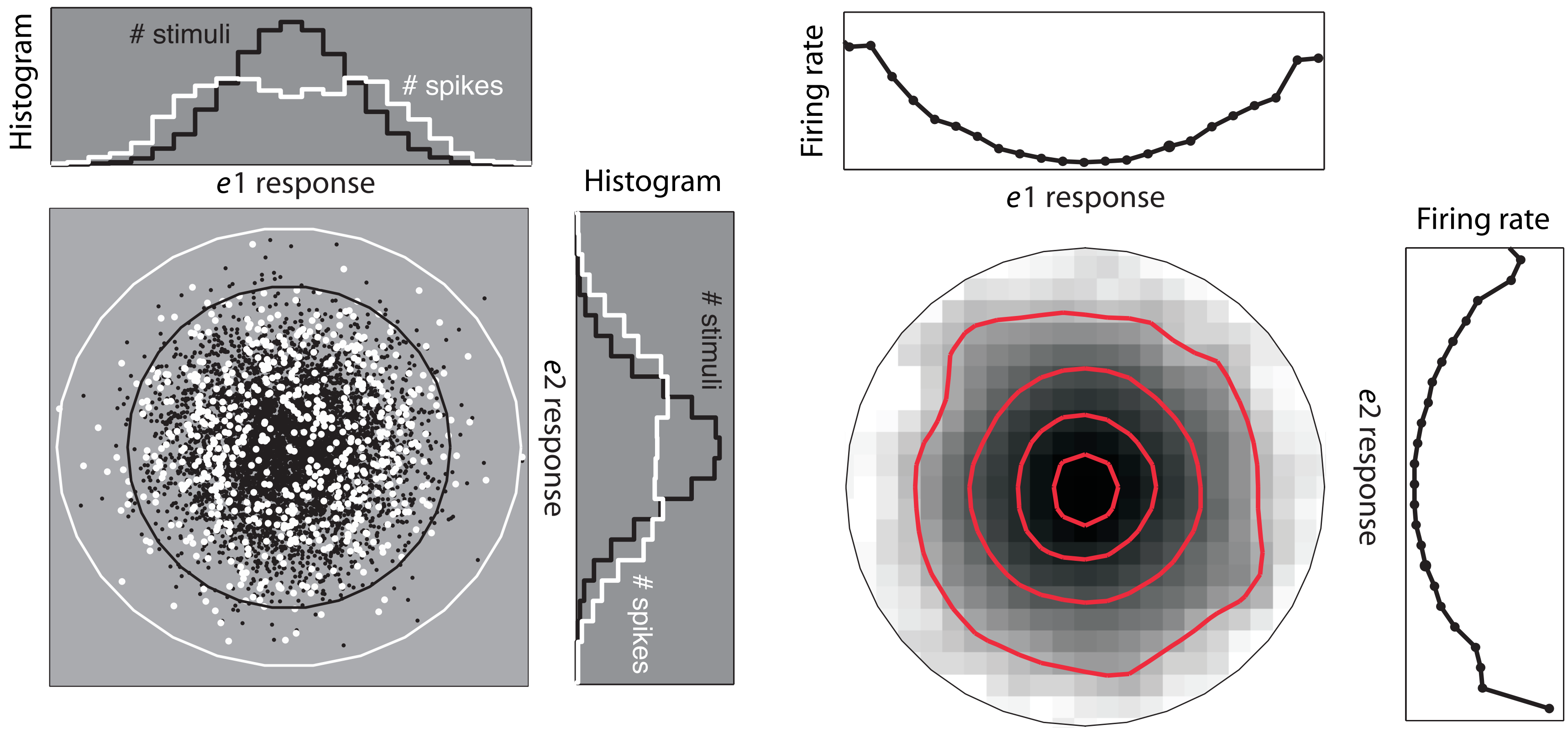
Choose directions with greatest change in variance:

$$\text{k-} \operatorname*{argmax}_{\|\mathbf{v}\|=1} \mathbf{v}^{\mathsf{T}}(C_{\text{prior}} - C_{\text{spike}})\mathbf{v}$$

$\Rightarrow$ find eigenvectors of $(C_{\text{prior}} - C_{\text{spike}})$ with large (absolute) eigvals.

STC

Reconstruct nonlinearity (may assume separability)



Biases

STC (obviously) requires that the nonlinearity alter variance.

If so, subspace is unbiased if distribution

- radially (elliptically) symmetric
- AND independent

$\Rightarrow$ Gaussian.

May be possible to correct by transformation, subsampling or weighting (latter two at cost of variance).

More LNP methods

- Non-parametric non-linearities: "Maximally informative dimensions" (MID) $\Leftrightarrow$ "non-parametric" maximum likelihood.
  - Intuitively, extends the variance difference idea to arbitrary differences between marginal and spike-conditioned stimulus distributions.
$$\mathbf{k}_{\text{MID}} = \underset{\mathbf{k}}{\operatorname{argmax}}\, \mathbf{KL}[P(\mathbf{k}\cdot\mathbf{x})\|P(\mathbf{k}\cdot\mathbf{x}|\text{spike})]$$
  - Measuring KL requires binning or smoothing—turns out to be equivalent to fitting a non-parametric nonlinearity by binning or smoothing.
  - Difficult to use for high-dimensional LNP models.

- Parametric non-linearities: the "generalised linear model" (GLM).

Generalised linear models

LN models with specified nonlinearities and exponential family noise.

In general (for monotonic g):

$$y \sim \text{ExpFamily}[\mu(\mathbf{x})]; \qquad g(\mu) = \beta\mathbf{x}$$

For our purposes easier to write

$$y \sim \text{ExpFamily}[f(\beta\mathbf{x})]$$

(Continuous time) point process likelihood with GLM-like dependence of λ on covariates is approached in limit of bins $\to 0$ by either Poisson or Bernoulli GLM.

Mark Berman and T. Rolf Turner (1992) Approximating Point Process Likelihoods with GLIM

Journal of the Royal Statistical Society. Series C (Applied Statistics), 41(1):31-38.

Generalised linear models

Poisson distribution $\Rightarrow f = \exp()$ is *canonical* (*natural params* $= \beta\mathbf{x}$).

Canonical link functions give concave likelihoods $\Rightarrow$ unique maxima.

Generalises (for Poisson) to any f which is convex and log-concave:

$$\text{log-likelihood} = c - f(\beta\mathbf{x}) + y \log f(\beta\mathbf{x})$$

Includes:

- threshold-linear
- threshold-polynomial
- "soft-threshold" $f(z) = \alpha^{-1}\log(1 + e^{\alpha z})$.

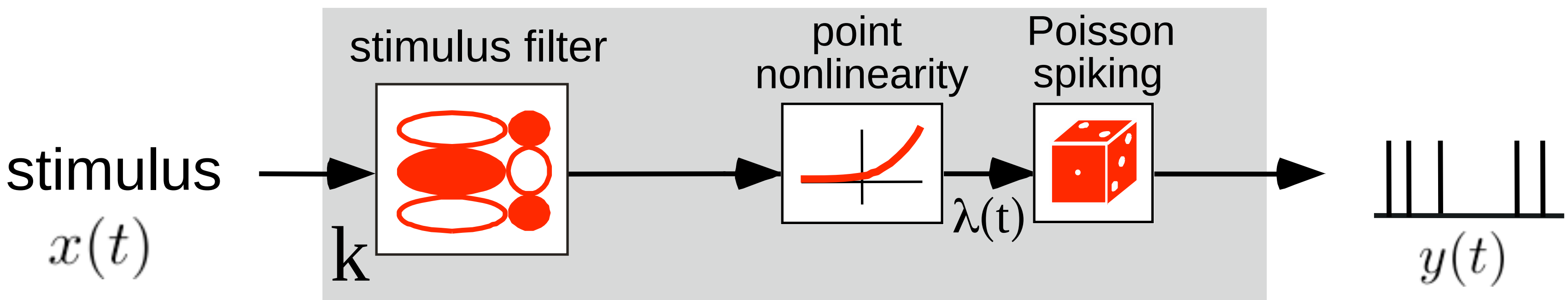
Generalised linear models

ML parameters found by

- gradient ascent
- IRLS

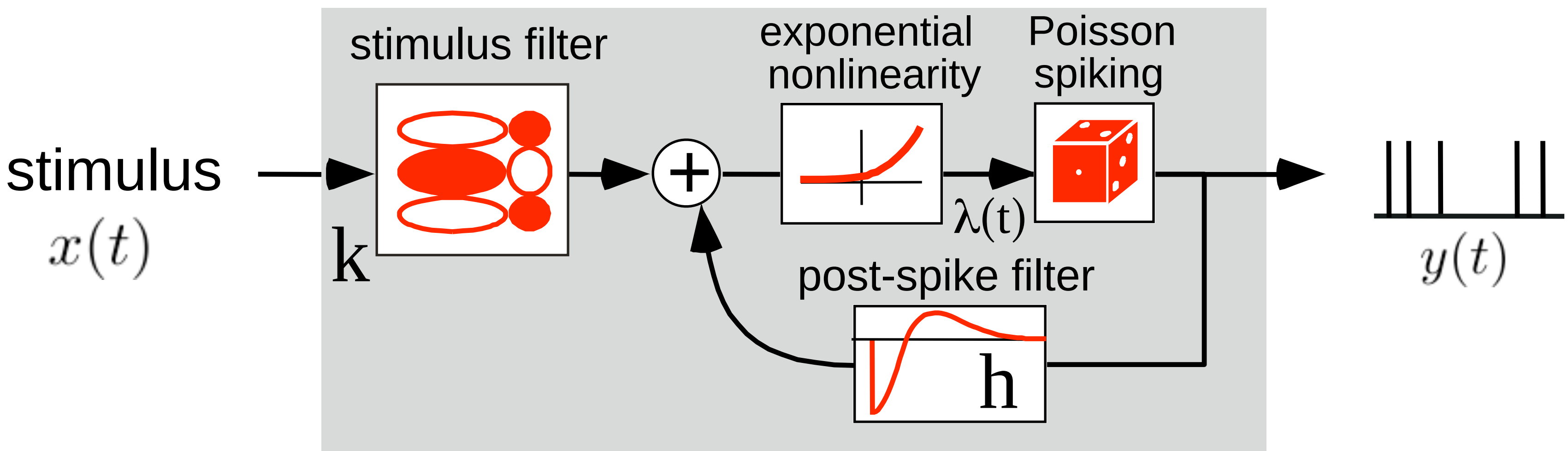
Regularisation by L_2 (quadratic) or L_1 (absolute value – sparse) penalties (MAP with Gaussian/Laplacian priors) preserves concavity.

Linear-Nonlinear-Poisson (GLM)



GLM with history-dependence

(Truccolo et al 04)



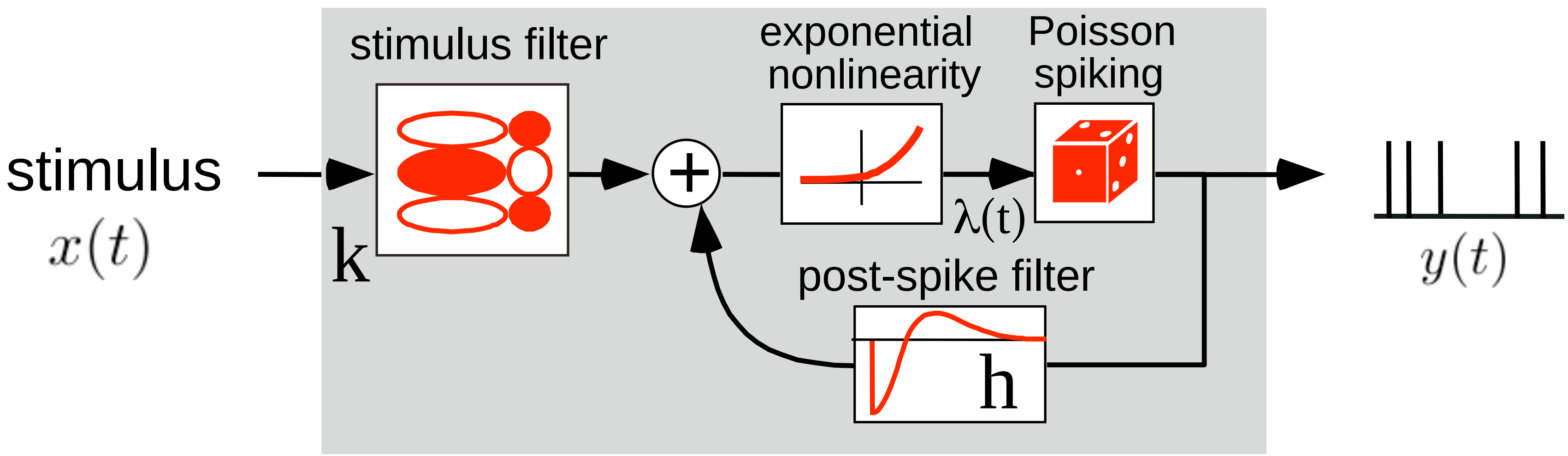
conditional intensity
(spike rate)

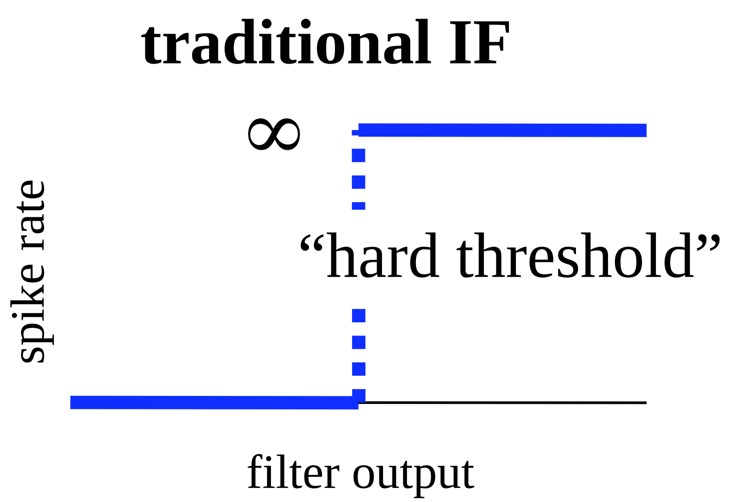
$$\lambda(t) = f(k \cdot x(t) \ + \ h \cdot y(t))$$

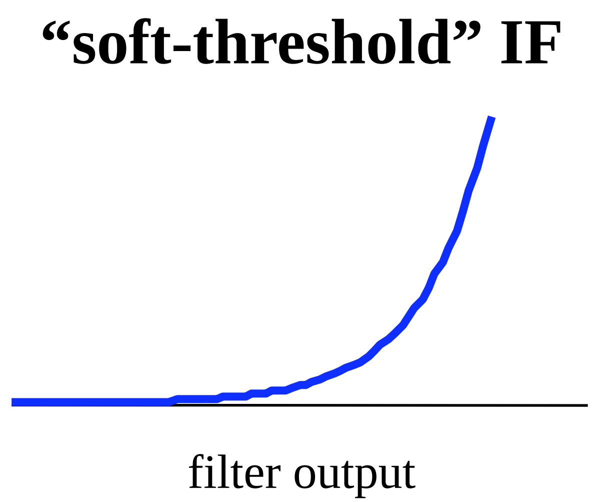
$$= e^{k \cdot x(t)} \ \cdot \ e^{h \cdot y(t)}$$

- rate is a product of stim- and spike-history dependent terms
- output no longer a Poisson process
- also known as “soft-threshold” Integrate-and-Fire model

GLM with history-dependence

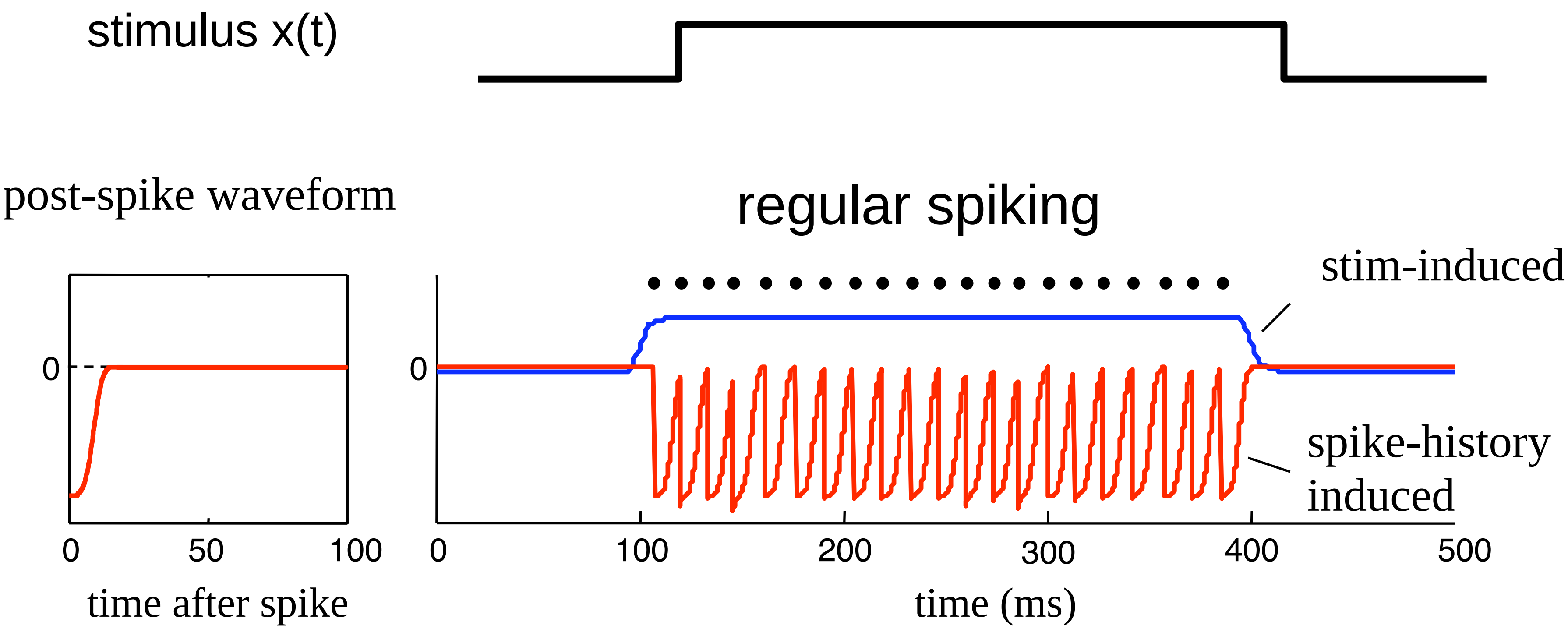




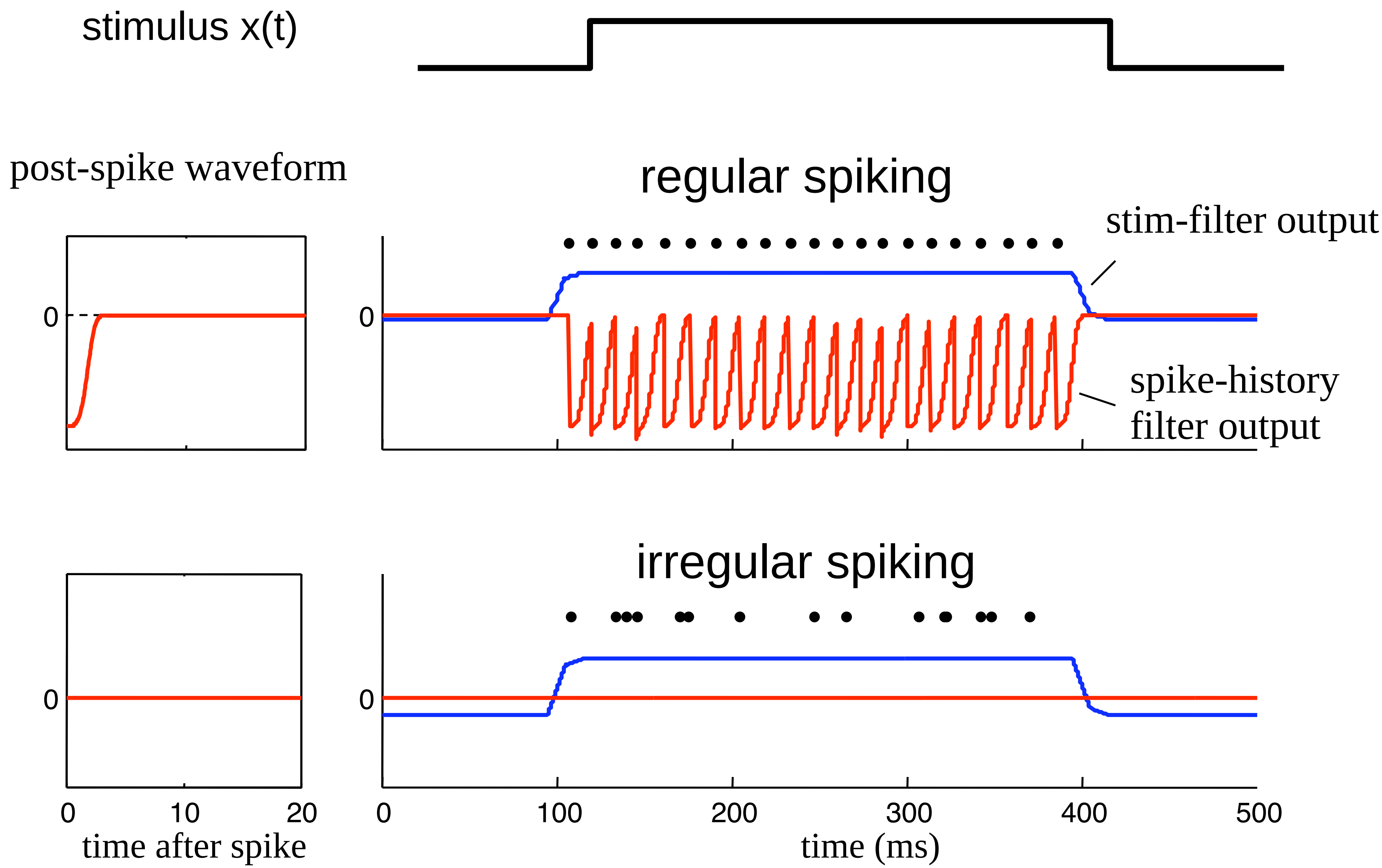


- "soft-threshold" approximation to Integrate-and-Fire model

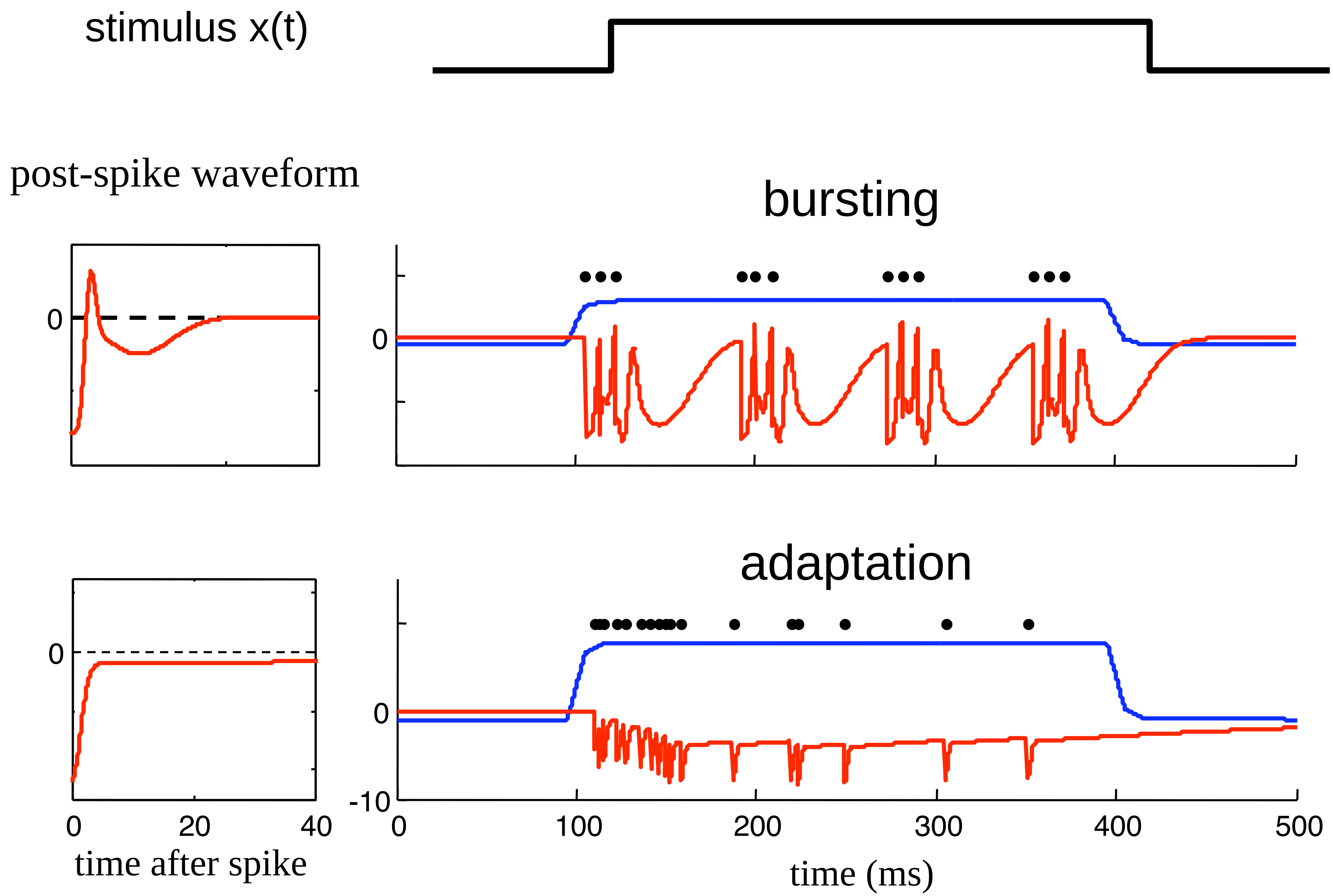
GLM dynamic behaviors



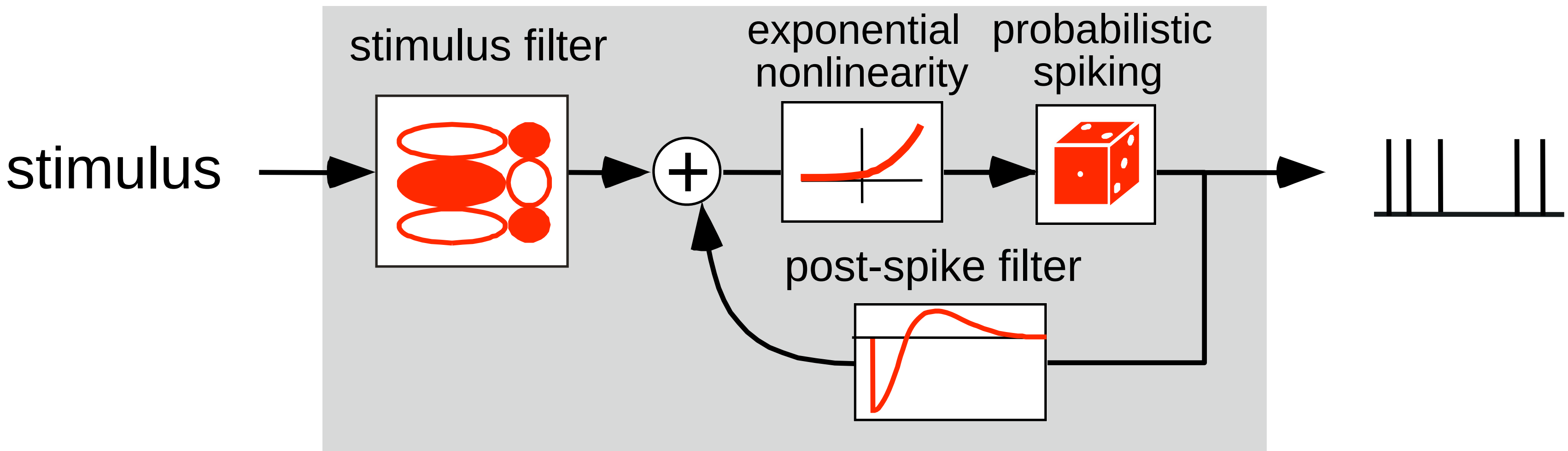
GLM dynamic behaviors



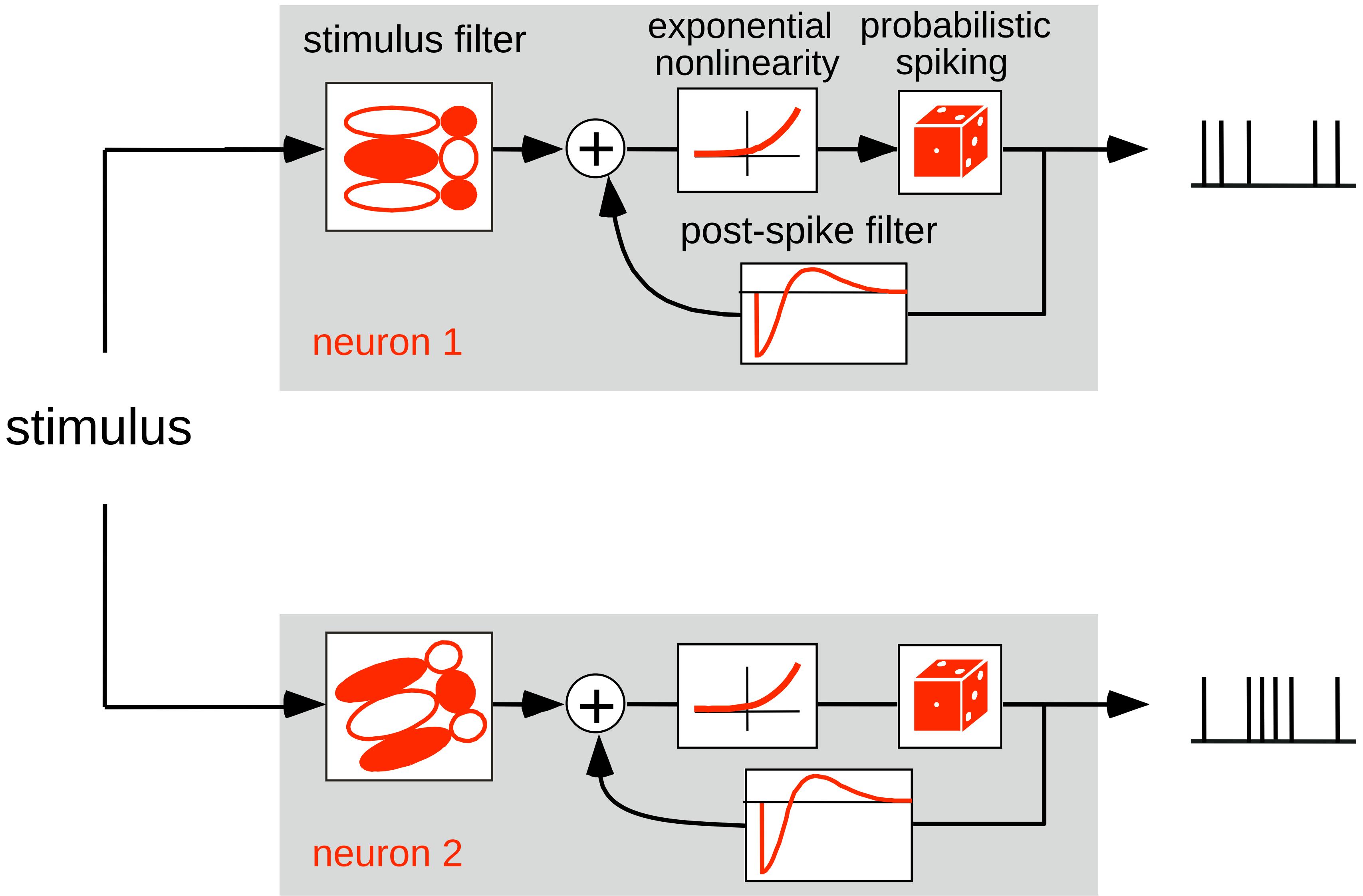
GLM dynamic behaviors



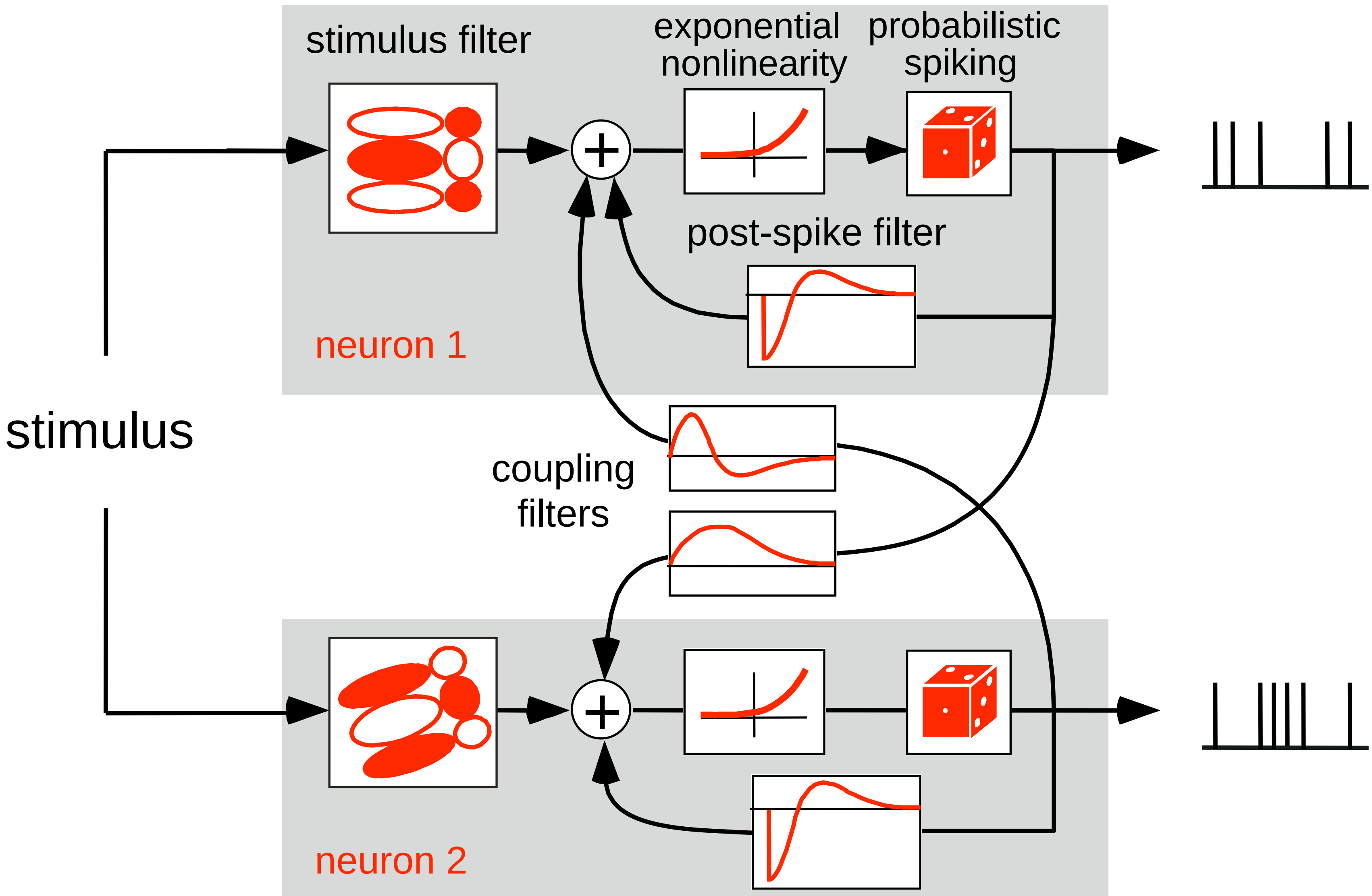
Generalized Linear Model (GLM)



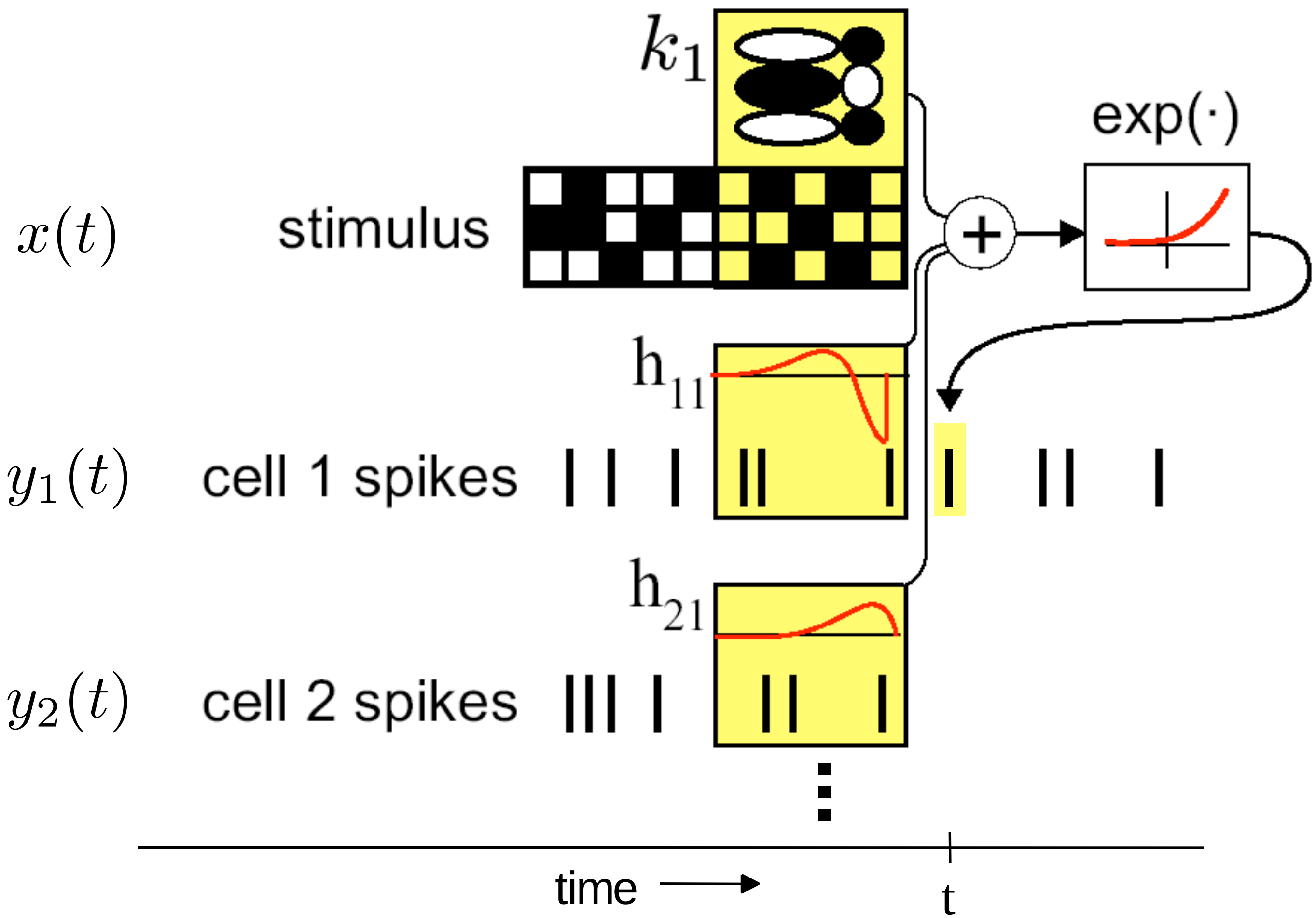
multi-neuron GLM



multi-neuron GLM



GLM equivalent diagram:



conditional intensity
(spike rate)

$$\lambda_i(t) = \exp(k_i \cdot x(t) \; + \; \textstyle\sum_j h_{ij} \cdot y(t))$$

Multilinear models

Input nonlinearities (Hammerstein cascades) can be identified in a multilinear (cartesian tensor) framework.

Input nonlinearities

The basic linear model (for sounds):

$$\underbrace{\widehat{r}(i)}_{\text{predicted rate}} = \sum_{jk} \underbrace{w^{\textbf{tf}}_{jk}}_{\text{STRF weights}} \underbrace{s(i-j,k)}_{\text{stimulus power}} ,$$

How to measure s? (pressure, intensity, dB, thresholded, ...)

We can *learn* an optimal representation $g(.)$:

$$\hat{r}(i) = \sum_{jk} w^{\textbf{tf}}_{jk} g(s(i-j,k)).$$

Define: basis functions $\{g_l\}$ such that $g(s) = \sum_l w^{\textbf{l}}_l g_l(s)$
and stimulus array $M_{ijkl} = g_l(s(i-j,k))$. Now the model is

$$\hat{r}(i) = \sum_j w^{\textbf{tf}}_{jk} w^{\textbf{l}}_l M_{ijkl} \quad \text{or} \quad \widehat{\mathbf{r}} = (\mathbf{w}^{\textbf{tf}} \otimes \mathbf{w}^{\textbf{l}}) \bullet \mathbf{M}.$$

Multilinear models

Multilinear forms are straightforward to optimise by alternating least squares.

Cost function:

$$\mathcal{E} = \left\| \mathbf{r} - (\mathbf{w}^{\mathbf{tf}} \otimes \mathbf{w}^{\mathbf{l}}) \bullet \mathbf{M} \right\|^2$$

Minimise iteratively, defining *matrices*

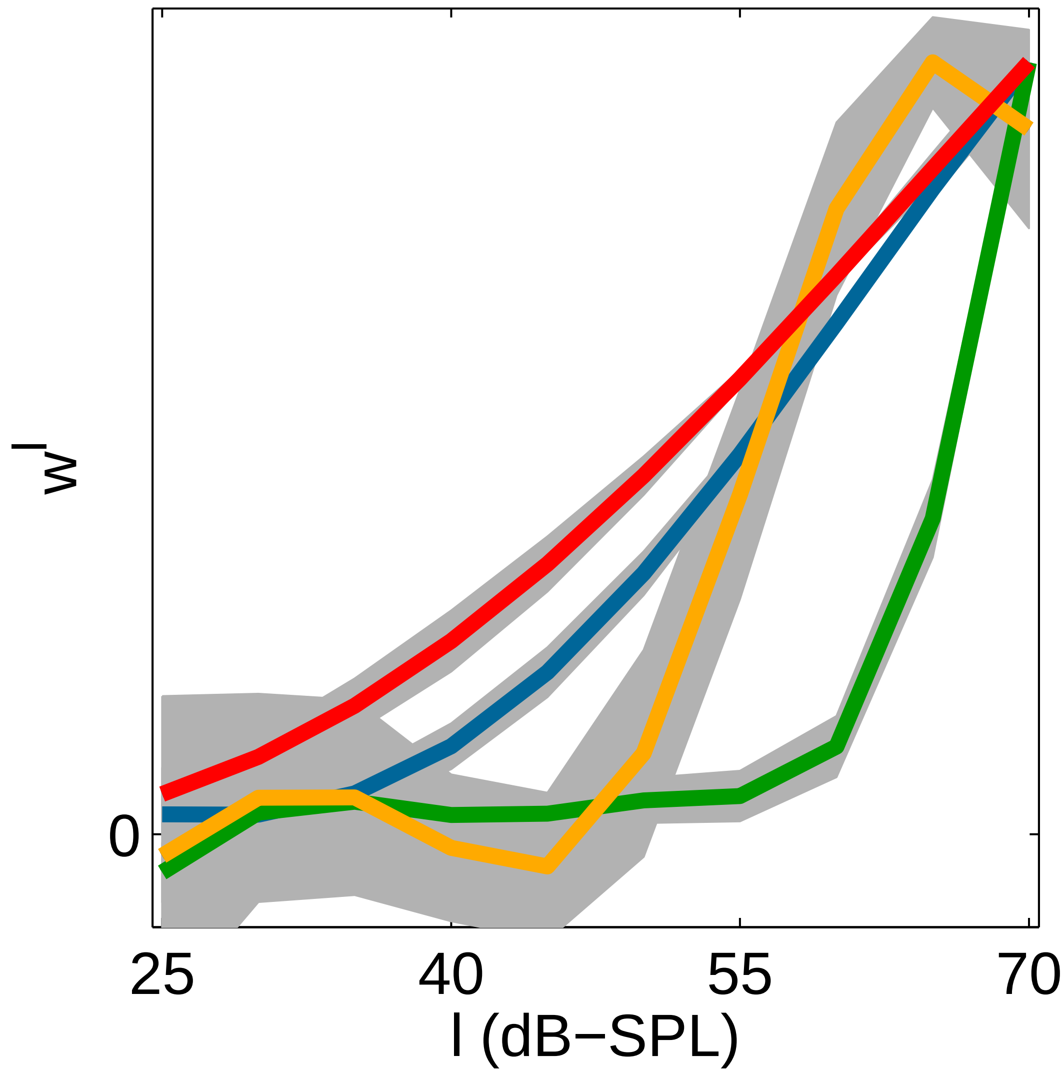
$$\mathbf{B} = \mathbf{w}^{\mathbf{l}} \bullet \mathbf{M} \qquad \text{and} \qquad \mathbf{A} = \mathbf{w}^{\mathbf{tf}} \bullet \mathbf{M}$$

and updating

$$\mathbf{w}^{\mathbf{tf}} = (\mathbf{B}^{\mathsf{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{r} \qquad \text{and} \qquad \mathbf{w}^{\mathbf{l}} = (\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{r}.$$

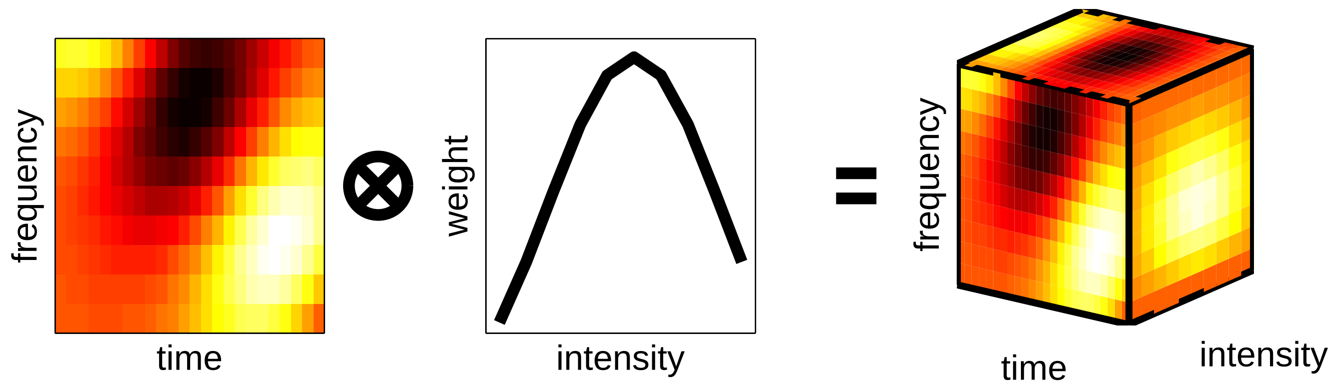
Each linear regression step can be regularised by evidence optimisation (suboptimal), with uncertainty propagated approximately using *variational* methods.

Some input non-linearities



Parameter grouping

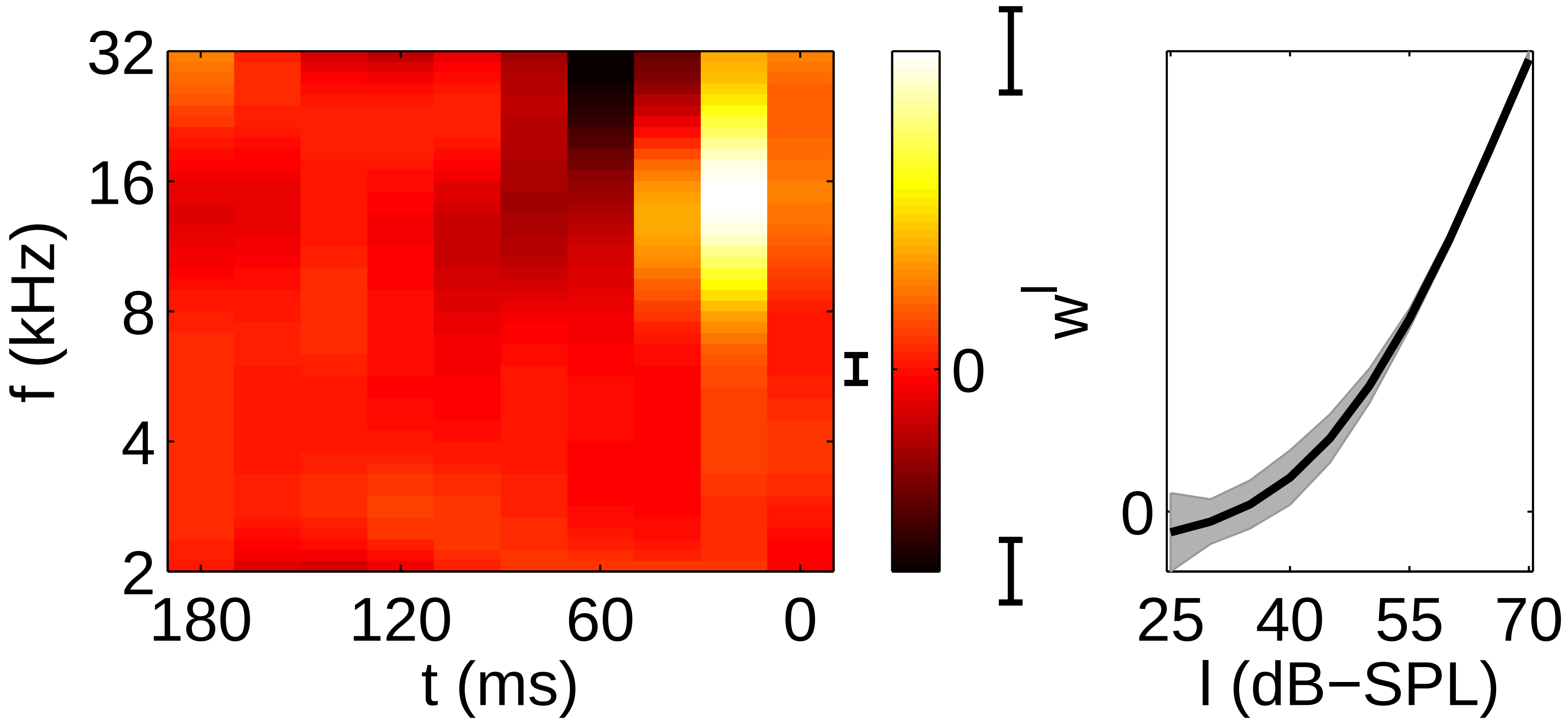
Separable STRF: (time) $\otimes$ (frequency). The input nonlinearity model is separable in another sense: (time, frequency) $\otimes$ (sound level).



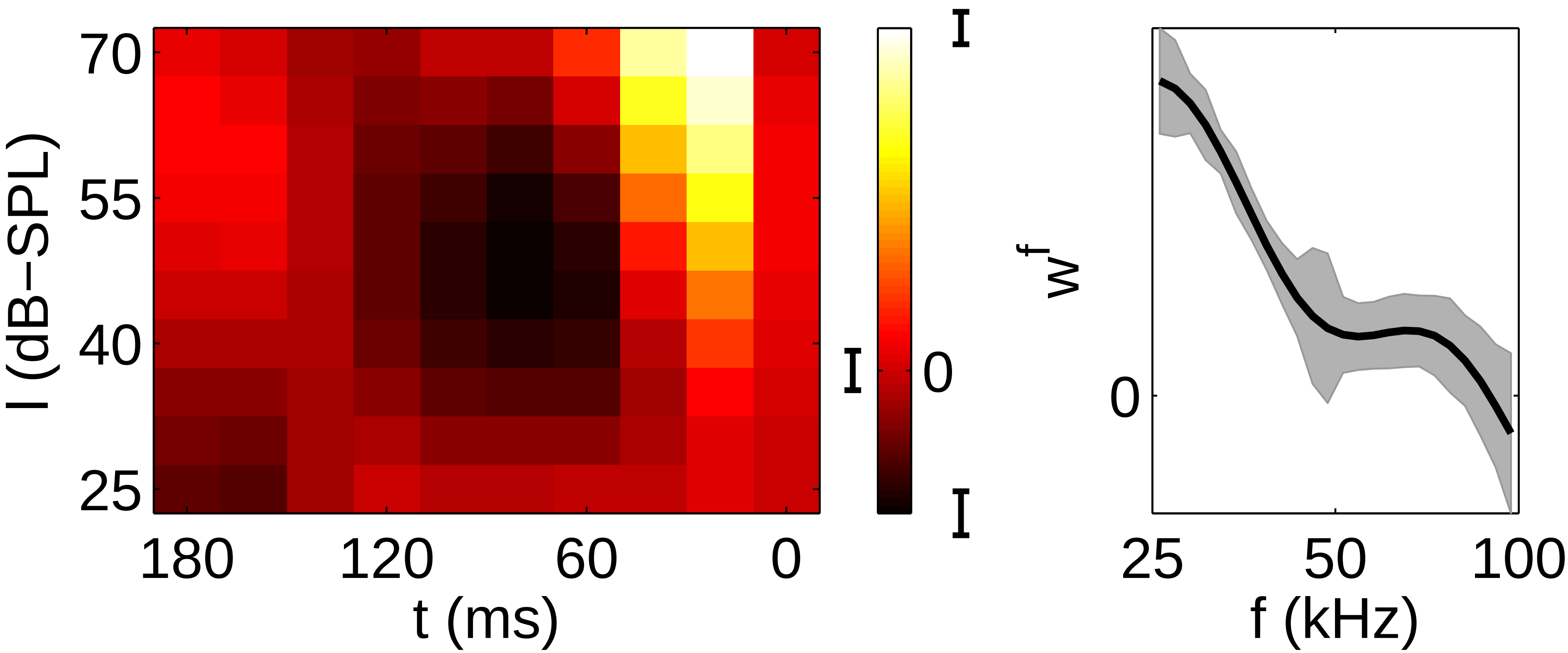
Other separations:

- (time, sound level) $\otimes$ (frequency): $\hat{\mathbf{r}} = (\mathbf{w}^{\mathbf{tl}} \otimes \mathbf{w}^{\mathbf{f}}) \bullet \mathbf{M}$,
- (frequency, sound level) $\otimes$ (time): $\hat{\mathbf{r}} = (\mathbf{w}^{\mathbf{fl}} \otimes \mathbf{w}^{\mathbf{t}}) \bullet \mathbf{M}$,
- (time) $\otimes$ (frequency) $\otimes$ (sound level): $\hat{\mathbf{r}} = (\mathbf{w}^{\mathbf{l}} \otimes \mathbf{w}^{\mathbf{f}} \otimes \mathbf{w}^{\mathbf{l}}) \bullet \mathbf{M}$.

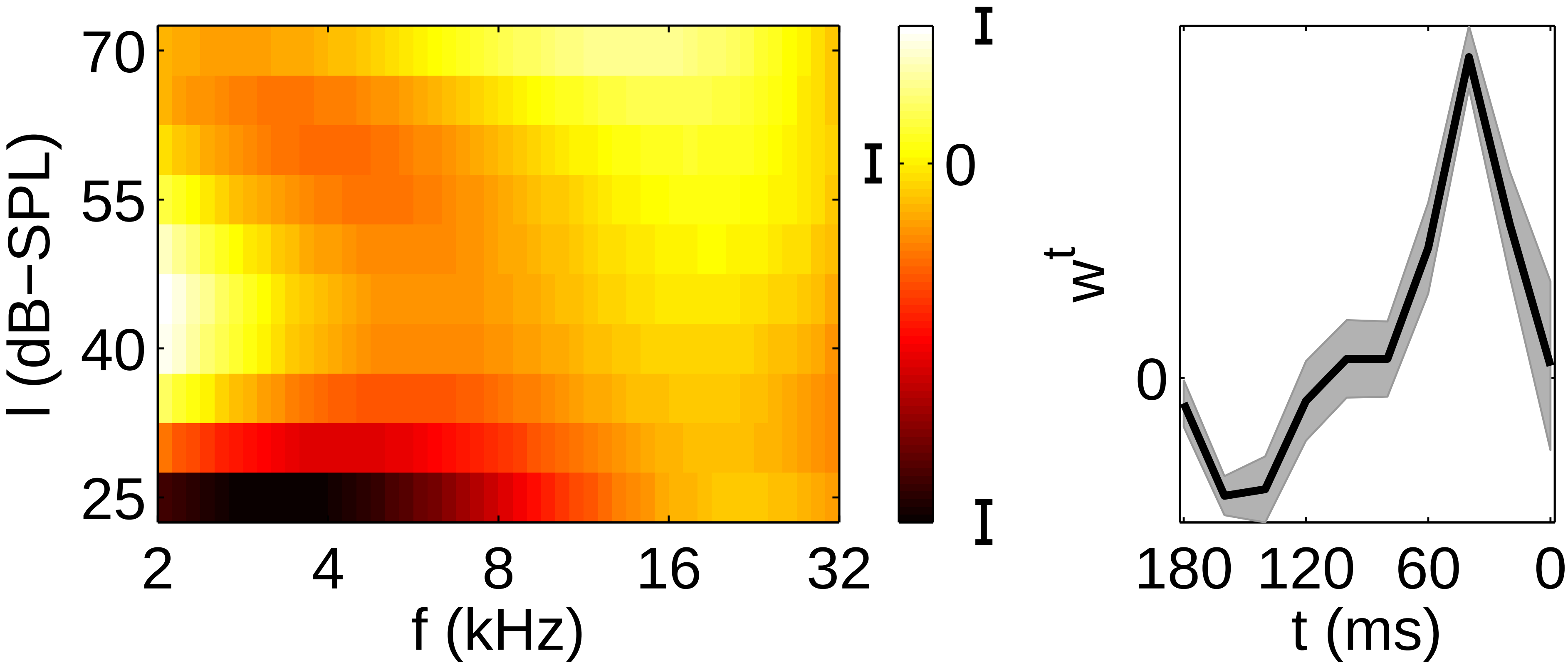
(time, frequency) ⊗ (sound level)



(time, sound level) ⊗ (frequency)

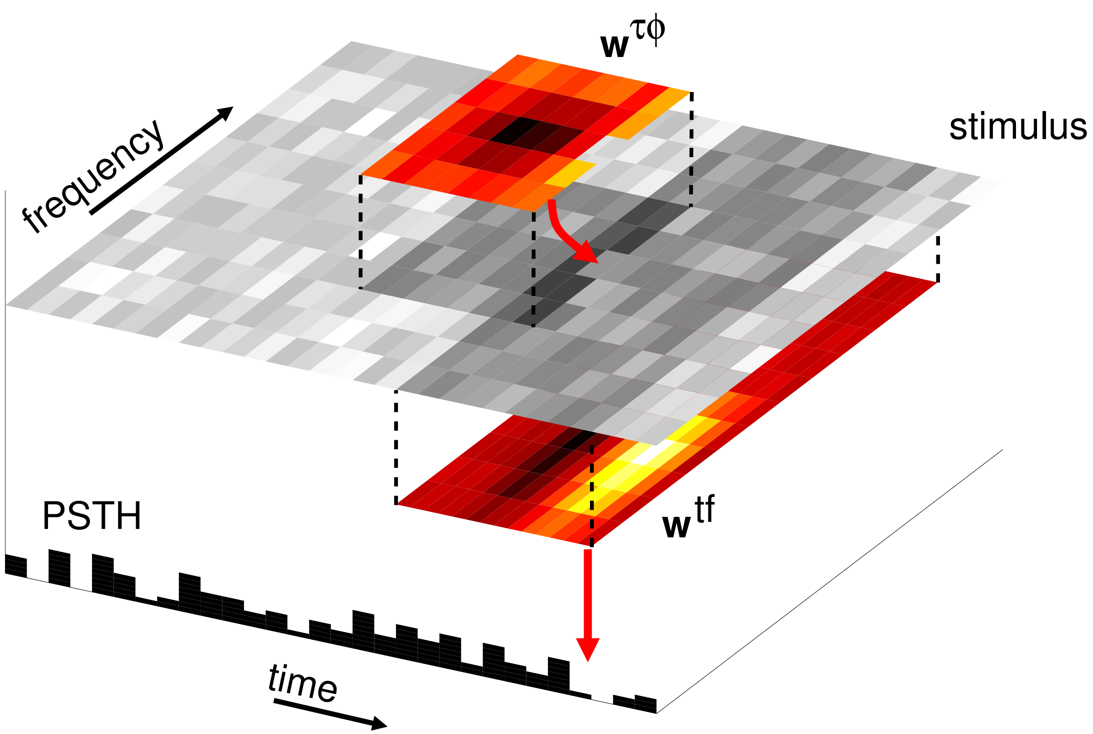


(frequency, sound level) ⊗ (time)



Contextual influences

$$\text{rate}(i) = c + \sum_{jkl} w_j^{\mathbf{t}} w_k^{\mathbf{f}} w_l^{\mathbf{l}}\, g_l(\text{sound}_{i-j,k})$$
$$\Big(c_2 + \sum_{mnp} w_m^{\boldsymbol{\tau}} w_n^{\boldsymbol{\phi}} w_p^{\boldsymbol{\lambda}}\, g_p(\text{sound}_{i-j-m,k+n})\Big)$$



Contextual influences

Introduce extra dimensions:

- $\boldsymbol{\tau}$: time difference between contextual and primary tone,
- ϕ: frequency difference between contextual and primary tone,
- $\boldsymbol{\lambda}$: sound level of the contextual tone.

Model the *effective* sound level of a primary tone by

$$\text{Level}(i,j) \rightarrow \text{Level}(i,j) \cdot (\text{const} + \text{Context}(i,j))\,.$$

and the context by

$$\text{Context}(i,j) = \sum_{m,n,p} w_m^{\boldsymbol{\tau}} w_n^{\boldsymbol{\phi}} w_p^{\boldsymbol{\lambda}} h_p(s(i-m, j+n))$$

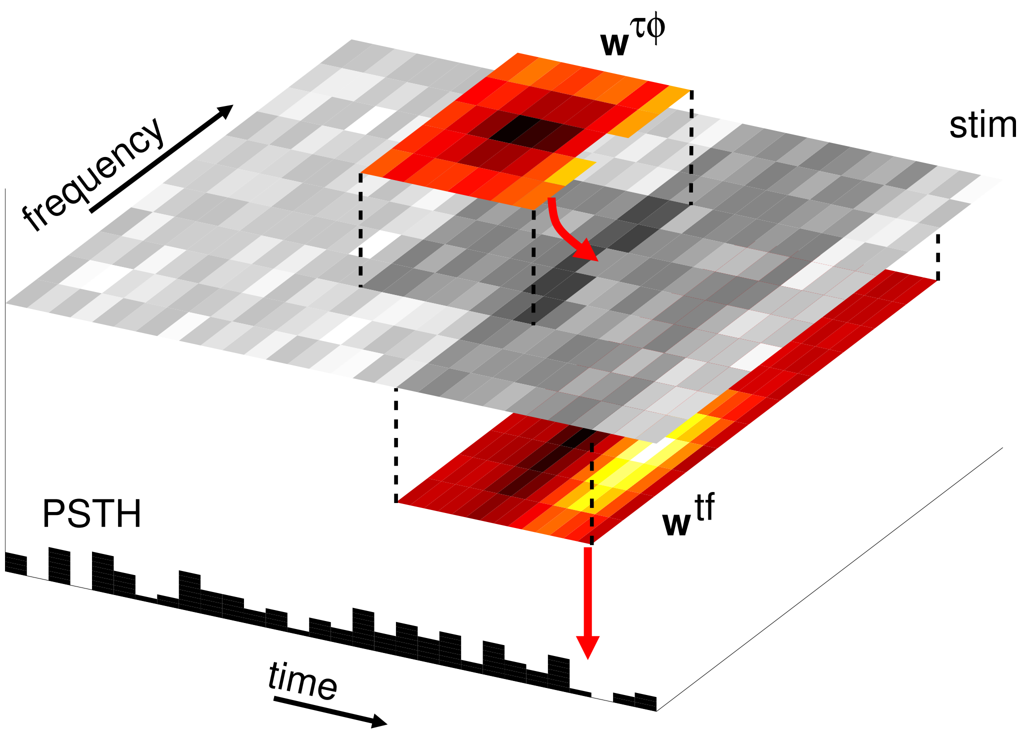
This leads to a multilinear model

$$\widehat{\mathbf{r}} = (\mathbf{w}^{\mathbf{t}} \otimes \mathbf{w}^{\mathbf{f}} \otimes \mathbf{w}^{\mathbf{l}} \otimes \mathbf{w}^{\boldsymbol{\tau}} \otimes \mathbf{w}^{\boldsymbol{\phi}} \otimes \mathbf{w}^{\boldsymbol{\lambda}}) \bullet \mathbf{M}.$$

Inseparable contexts

We can also allow *inseparable* contexts (and principal fields), dropping the level-nonlinearity to reduce parameters.

$$r(i) = c + \sum_{jk} w^{\mathbf{tf}}_{jk}\, \mathsf{sound}_{i-j,k} \Big(1 + \sum_{mn} w^{\boldsymbol{\tau\phi}}_{mn}\, \mathsf{sound}_{i-j-m,k+n}\Big)$$



Performance