Neural Decoding

Mark van Rossum
School of Informatics, University of Edinburgh
January 2012

Acknowledgements: Chris Williams and slides from Gatsby Liam Paninski.
Version: January 15, 2018

Why decoding?
Understanding the neural code.
- Given spikes, what was the stimulus?
- What aspects of the stimulus does the system encode? (capacity is limited)
- What information can be extracted from spike trains:
  - By “downstream” areas? Homunculus.
  - By the experimenter? Ideal observer analysis.
- What is the coding quality?
- Design of neural prosthetic devices

Related to encoding, but encoding does not answer above questions explicitly.

Decoding examples
- Hippocampal place cells: how is location encoded?
- Retinal ganglion cells: what information is sent to the brain? What is discarded?
- Motor cortex: how can we extract as much information as possible from a collection of M1 cells?

Overview
- Stimulus reconstruction (single spiking neuron, dynamic stimuli)
- Spike train discrimination (spike based)
- Stimulus discrimination (single neuron, rate based, static stimulus $s = \{s_a, s_b\}$)
- Population decoding (multiple neurons, rate based, static stimulus $s \in \mathbb{R}$)
- Dynamic population decoding ($s(t) \in \mathbb{R}$)
1. Spike train decoding

Dayan and Abbott §3.4, Rieke Chap 2 and Appendix

Estimate the stimulus from spike times \( t_i \) to minimize e.g.
\[
\langle s(t) - s_{est}(t) \rangle^2
\]

First order reconstruction:
\[
s_{est}(t - \tau_0) = \sum_{t_i} K(t - t_i) - \langle r \rangle \int d\tau K(\tau)
\]

The second term ensures that \( \langle s_{est}(t) \rangle = 0 \)

Delay \( \tau_0 \) can be included to make decoding easier: predict stimulus at time \( t - \tau_0 \) based on spikes up to time \( t \) (see causal decoding below)

Or use Fourier space
\[
\tilde{K}(\omega) = \frac{\tilde{Q}_{rs}(\omega) \exp(i\omega \tau_0)}{\tilde{Q}_{rr}(\omega)}
\]

Note, one can design the stimulus (e.g. Gaussian white noise), but one can not design the response \( r(t) \).

If \( Q_{rr}(\tau) \approx \langle r \rangle \delta(\tau) \) (tends to happen at low rates, hence not very relevant) then \( K \) is the STA, so decoder equals encoder
\[
K(\tau) = \frac{1}{\langle n \rangle} \left\langle \sum_{i=1}^{n} s(t_i + \tau - \tau_0) \right\rangle
\]

Note, for constant Poisson process \( Q_{rr}(\tau) \approx \langle r \rangle \delta(\tau) \)

Acausal Minimization

Let \( r(t) = \sum \delta(t - t_i) \)

Minimizing squared error (similar to Wiener kernels) gives implicit equation for optimal \( K \)
\[
\int_{-\infty}^{\infty} d\tau' Q_{rr}(\tau - \tau') K(\tau') = Q_{rs}(\tau - \tau_0)
\]

where
\[
Q_{rr}(\tau - \tau') = \frac{1}{T} \int_0^T dt((r(t - \tau) - \langle r \rangle)(r(t - \tau') - \langle r \rangle))
\]
\[
Q_{rs}(\tau - \tau_0) = \langle r \rangle C(\tau_0 - \tau)
\]

where \( C(\tau) = \langle 1/n \sum_i s(t_i - \tau) \rangle \) is STA from encoding slides.

Quality of reconstruction

Define reconstruction quality as :
\[
\gamma = 1 - \frac{\langle (s_{est} - s)^2 \rangle}{\sigma_s}
\]

An I&F transmits more information than Poisson (cf. encoding).
Causal decoding

- H1 neuron of the fly
- Solid line is reconstruction using acausal filter
- Note, reconstruction quality will depend on stimulus


- Requires \( K(t - t_i) = 0 \) for \( t \leq t_i \).

\[
\hat{s}(t - \tau_0) = \sum_{t_i} K(t - t_i) - \langle r \rangle \int d\tau K(\tau)
\]

- Delay \( \tau_0 \) buys extra time

Delay \( \tau_0 = 160 \) ms. (C: full (non-causal) kernel)

Finding \textit{optimal} kernel while imposing causality analytically is harder.

- Hope that \( K(\tau) = 0 \) for \( \tau < 0 \) and \( \tau_0 \) sufficiently large.
- Wiener-Hopf method (spectral factorization)
- Expand \( K(\tau) \) using a causal basis
- Use discrete formulation

Causality

- Organism faces causal (on-line) decoding problem.
Higher order reconstruction

Build a library of spike patterns (up to triplets). Measure mean and covariance of $P(s|\{t_0, t_1, \ldots\})$. Reconstruct with weighted sum of means, §A6.

Conclusion stimulus reconstruction

- Stimulus reconstruction similar to encoding problem. But
  - Response is given, can not be chosen to be white
  - Imposing causality adds realism but reduces quality
- The reconstruction problem can be ill-posed. It is not always possible to reconstruct stimulus (cf dictionary). For instance: complex cell.
  - Still, the cell provides information about the stimulus. Could try to read the code, rather than reconstruct the stimulus (e.g. ideal observer)

2. Spike train discrimination

Given two spike trains. How similar are they, or how they compare to template?

Problem: very high dimensional space.

- Cricket auditory neuron in response to 2 songs, 5 repeats/song
- 'Edit distance': two processes
  - Deleting/inserting a spike costs 1
  - Moving a spike costs $\frac{1}{2}[1 - \exp(-|\delta t|/\tau)]$, with parameter $\tau$.

Spike distances

Simpler algorithm:
Convolve (filter) with an exponential $\tilde{f}(t) = \sum_{t_i < t} \exp(-|t - t_i|/t_c)$ and calculate $L_2$ distance

$$D^2 = \frac{1}{t_c} \int_0^T dt [\tilde{f}(t) - \tilde{g}(t)]^2$$

Similar to coherence of between trains
Application to cricket auditory neurons: Play songs repeatedly and discriminate [1]

Optimal discrimination when $\tau$ similar to neural integration time

3. Stimulus Discrimination

- Dayan and Abbott §3.2
- $p(s|r)$, where $r$ is response across neurons and/or time
- In general $s$ can be continuous, e.g. speed
- First, discrimination i.e. distinguishing between two (or more) alternatives (e.g. stimulus or no stimulus)
- For now no time-dependent problems.

SNR and ROC curves

Discriminate between response distributions $P(r_1)$ and $P(r_2)$.

ROC: vary decision threshold and measure error rates.
Larger area under curve means better discriminability
Shape relates to underlying distributions.

For Gaussian distributed responses define single number

$$SNR = 2 \frac{[(r_1) - (r_2)]^2}{\text{var}(r_1) + \text{var}(r_2)}$$

Note, $SNR = 2 \frac{|(r_1) - (r_2)|}{\text{sd}(r_1) + \text{sd}(r_2)}$ is also used, neither is principled when $\text{var}(r_1) \neq \text{var}(r_2)$. 

Spike distances

Using spike distance to measure intrinsic noise:

Discriminate between response distributions $P(r_1)$ and $P(r_2)$.

$D^2$ varies decision threshold and measures error rates.

Larger area under curve means better discriminability
Shape relates to underlying distributions.

For Gaussian distributed responses define single number

$$SNR = 2 \frac{[(r_1) - (r_2)]^2}{\text{var}(r_1) + \text{var}(r_2)}$$

Note, $SNR = 2 \frac{|(r_1) - (r_2)|}{\text{sd}(r_1) + \text{sd}(r_2)}$ is also used, neither is principled when $\text{var}(r_1) \neq \text{var}(r_2)$. 

Readout of a single MT neuron

Some single neurons do as well as animal!
Possibility for averaging might be limited due to correlation?
Population might still be faster [7]

Readout of Object Identity from Macaque IT Cortex

Recording from ~300 sites in the Inferior Temporal (IT) cortex
Present images of 77 stimuli (of different objects) at various locations and scales in the visual field.
Task is to categorize objects into 8 classes, or identify all 77 objects
Predictions based on one-vs-rest linear SVM classifiers, using data in 50 ms bins from 100 ms to 300 ms after stimulus onset

What does this tell us?
Performance of such classifiers provides a lower bound on the information available in the population activity
If neurons were measured independently (paper is unclear), correlations are ignored. Correlation could limit or enhance information...
Distributed representation
Linear classifier can plausibly be implemented in neural hardware

Fig. 1. Accurate readout of object category and identity from IT population activity. (A) Example of multi-unit spiking responses of 3 independently recorded sites to 5 of the 77 objects. Rasters show spikes in the 200 ms after stimulus onset for 10 repetitions (black bars indicate object presentation). (B) Performance of a linear classifier over the entire object set on test data (not used for training) as a function of the number of sites for reading out object category (red, chance ~ 12.5%) or identity (blue, chance ~ 1.3%). The input from each site was the spike count in consecutive 50 ms bins from 100 to 300 ms after stimulus onset [29]. Sequentially recorded sites were combined by assuming independence (Supporting Online Material). In this and subsequent figures, error bars show the SD for 20 random choices of the sites used for training; the dashed lines show chance levels, and the bars next to the dashed lines show the range of performances using the 200 ms before stimulus onset (control). (C) Categorization performance (n = 64 sites, mean ± SD) for different data sources used as input to the classifier: multi-unit activity (MUA) as shown in (B), single-unit activity (UA), and local field potentials (LFP) (Supporting Online Material). (D) This confusion matrix describes the pattern of mistakes made by the classifier (n = 256 sites). Each raw indicates the actual category presented to the monkey [29], and each column indicates the classifier predictions (in color code).
Visual system decoding: independence

Face cells, rate integrated over 500ms, extrapolated to large #stimuli. Extract face identity from population response.

Coding is almost independent! (for these small ensembles)

Dayan and Abbott §3.3

Population encoding uses a large number of neurons to represent information

Advantage 1: reduction of uncertainty due to neuronal variability (Improves reaction time).

Advantage 2: Ability to represent a number of different stimulus attributes simultaneously (e.g. in V1 location and orientation).

4. Population Encoding

Population encoding uses a large number of neurons to represent information

Advantage 1: reduction of uncertainty due to neuronal variability (Improves reaction time).

Advantage 2: Ability to represent a number of different stimulus attributes simultaneously (e.g. in V1 location and orientation).

Cricket Cercal System

At low velocities, information about wind direction is encoded by just four interneurons

\[ \left( \frac{f(s)}{r_{\text{max}}} \right) = [\cos(s - s_a)]_+ \]

Note, rate coding assumed.

Let \( c_a \) denote a unit vector in the direction of \( s_a \), and \( v \) be a unit vector parallel to the wind velocity

\[ \left( \frac{f(s)}{r_{\text{max}}} \right) = [v \cdot c_a]_+ \]

Crickets are Cartesian, 4 directions 45°, 135°, −135°, −45°

Population vector is defined as

\[ v_{\text{pop}} = \sum_{a=1}^{4} \left( \frac{r}{r_{\text{max}}} \right)_a c_a \]
Primary Motor Cortex (M1)

- Certain neurons in M1 of the monkey can be described by cosine functions of arm movement direction (Georgopoulos et al, 1982)
- Similar to cricket cercal system, but note:
  - Non-zero offset rates $r_0$
  $$
  \left( \frac{f(s) - r_0}{r_{\text{max}}} \right) = v \cdot c_a
  $$
  - Non-orthogonal: there are many thousands of M1 neurons that have arm-movement-related tuning curves

Optimal Decoding

Calculate
$$
 p(s|r) = \frac{p(r|s)p(s)}{p(r)}
$$

- Maximum likelihood decoding (ML): $\hat{s} = \arg\max_s p(r|s)$
- Maximum a posteriori (MAP): $\hat{s} = \arg\max_s p(s)p(r|s)$
- Bayes: minimize loss
  $$
  s_B = \arg\min_{s^*} \int_s L(s, s^*) p(s|r) ds
  $$
  - For squared loss $L(s, s^*) = (s - s^*)^2$, optimal $s^*$ is posterior mean, $s_B = \int_s p(s|r)s$. 

For the cercal system, assuming indep. noise

\[ p(r|s) = \prod_a p(r_a|s) \]

where each \( p(r_a|s) \) is modelled as a Gaussian with means and variances

- \( p(s) \) is uniform (hence MAP=ML)
- ML decoding finds a peak of the likelihood
- Bayesian method finds posterior mean
- These methods improve performance over the vector method (but not that much, due to orthogonality...)

**Poisson firing model over time**

\[ \text{count } n_a = r_a T \text{ spikes.} \]

\[
p(r|s) = \prod_{a=1}^{N} \frac{(f_a(s) T)^{n_a}}{n_a!} \exp(-f_a(s) T)\
\]

\[
\log p(r|s) = \sum_{a=1}^{N} n_a \log f_a(s) + \ldots\
\]

Approximating that \( \sum_a f_a(s) \) is independent of \( s \)
ML decoding

- \( s_{ML} \) is stimulus that maximizes \( \log p(r|s) \), determined by

\[
\sum_{a=1}^{N} r_a f'_a(s_{ML}) = 0
\]

- If all tuning curves are Gaussian \( f_a = A \exp\left[-(s - s_a)^2/2\sigma^2_w\right] \) then

\[
s_{ML} = \frac{\sum_a r_a s_a}{\sum_a r_a}
\]

which is simple and intuitive, known as Center of Mass (cf population vector)

Accuracy of the estimator

- Bias and variance of an estimator \( s_{est} \)

\[
b_{est}(s) = \langle s_{est} \rangle - s
\]

\[
\sigma^2_{est}(s) = \langle (s_{est} - \langle s_{est} \rangle)^2 \rangle
\]

\[
\langle (s - s_{est})^2 \rangle = b_{est}^2(s) + \sigma^2_{est}
\]

- Thus for an unbiased estimator, MSE \( \langle (s - s_{est})^2 \rangle \) is given by \( \sigma^2_{est} \), the variance of the estimator

Fisher information

- Fisher information is a measure of the curvature of the log likelihood near its peak

\[
I_F(s) = \left\langle \frac{\partial^2 \log p(r|s)}{\partial s^2} \right\rangle_s = -\int dr p(r|s) \frac{\partial^2 \log p(r|s)}{\partial s^2}
\]

(the average is over trials measuring \( r \) while \( s \) is fixed)

- Cramér-Rao bound says that for any estimator [?]

\[
\sigma^2_{est} \geq \frac{(1 + b'_{est}(s))^2}{I_F(s)}
\]

- **efficient estimator** if \( \sigma^2_{est} = \frac{(1 + b'_{est}(s))^2}{I_F(s)} \).

  - In the bias-free case an efficient estimator \( \sigma^2_{est} = 1/I_F(s) \).

  - ML decoder is typically efficient when \( N \rightarrow \infty \).

- In homogeneous systems \( I_F \) indep. of \( s \).

- More generally Fisher matrix \( (I_F)_{ij}(s) = \left\langle \frac{\partial^2 \log p(r|s)}{\partial s_i \partial s_j} \right\rangle_s \).

- Taylor expansion of Kullback-Leibler

\[
D_{KL}(P(s), P(s + \delta s)) \approx \sum_{ij} \delta s_i \delta s_j (I_F)_{ij}
\]

- Not a Shannon information measure (not in bits), but related to Shannon information in special cases, e.g. [?, ?].
For independent Poisson spikers

\[ I_F(s) = \left\langle -\frac{\partial^2 \log p(r|s)}{\partial s^2} \right\rangle = T \sum_a \langle r_a \rangle \left( \frac{f'_a(s)}{f_a(s)} - \frac{f''_a(s)}{f_a(s)} \right) \]

For dense, symmetric tuning curves, the second term sums to zero. Using \( f_a(s) = \langle r_a \rangle \) we obtain

\[ I_F(s) = T \sum_a \left( \frac{f'_a(s)}{f_a(s)} \right)^2 \]

For dense \( f_a(s) = Ae^{-(s-s_0+a.ds)^2/2\sigma^2} \) with density \( \rho = 1/ds \), sum becomes integral

\[ I_F = \sqrt{2\pi} TA\rho/\sigma_w \]

For Gaussian tuning curves \[Dayan and Abbott (2001)\]

Note that Fisher information vanishes at peak as \( f'_a(s) = 0 \) there

Slope as strategy

From paper on bat echo location \[?\]

Population codes and noise correlations

Noise in neurons can be correlated \( p(r|s) \neq \prod_{a=1}^N p(r_a|s) \) Information in the code can go up or down with correlations depending on details \[?, ?, ?\] ...

From paper on bat echo location \[?\]
Population codes and correlations

Gaussian noise model, with stimulus dep. covariance $Q(s)$:

$$P(r|s) = \frac{1}{\sqrt{(2\pi)^N \det Q}} e^{-[r-f(s)]^T Q^{-1}[r-f(s)]/2}$$

then [?]

$$I_F = f'(s)Q^{-1}(s)f'(s) + \frac{1}{2} \text{Tr}[Q'(s)Q^{-1}(s)Q'(s)Q^{-1}(s)]$$

When $Q'(s) = 0$ and $Q_{ij} = q(|i-j|)$, can use spatial Fourier representation. $I_F$ becomes sum of signal-to-noise ratios

$$I_F = \sum_k |\tilde{f}'(k)|^2 \tilde{q}(k)$$

Thus noise with same correlation length as $f'(s)$ is most harmful [?]

---

Population codes and correlations: Retina

Fit coupled I&F-model (see encoding) to retina data
Optimal receptive field width?

Maximize $I_F = T \sum a (f'(s))^2$ to minimize MSE error [?]
- $(f'(s))^2$ is large for narrow curves
- $I_F$ is increased by including many neurons in the sum, but this is in conflict with narrow tuning: trade-off
- Gaussian tuning curves and replace sum with integral. $D = 1$: accuracy best for infinitely narrow tuning
- For $D = 2$ there is no effect of the width on $I_F$.
- For $D > 2$ $I_F$ increases as tuning curves broaden [?].
- What is $D$ in various brain areas? (93 in IT [?])

Alternative view on optimal coding width

Consider transmission. Maximize $I_{out}^{F}$ wrt connections.

Hippocampal Place Cell Decoding

Encoding: place cells modelled as inhomogeneous Poisson processes
Dynamic model: random walk
Decoding: approximate Kalman filtering
Approach here is to perform inference to invert the encoding process

Minimal loss if output is tuned to input. I.e. RF width depends on input.
Example: Decoding in Hippocampus

Example: Motor decoding

- Rhesus monkey, 43 electrodes in M1
- Monkey controls cursors on a screen using two manipulanda to perform a centre-out reaching task
- Predict hand velocity based on 10 time bins, each of length 100 ms in all 43 neurons.
- Can use linear regression, polynomial regression, Gaussian kernel (support vector regression), spikernel (allows time warping)
- More sophisticated methods outperform linear regression, but linear is already decent
- State-of-the-art w. Kalman filters
- Reconstruction of temporal stimulus
- Spike distances
- Discrimination task
- Population decoding: vector method and “optimal” decoding methods
- Specialist applications using domain knowledge