86-631/42-631 Neural Data Analysis
Lecture 6: Information theory and neural coding I

1) Failures of correlation
    a. Emphasis that *any* change in distribution is potentially informative!
2) Entropy
    a. Quantifying randomness
    b. Properties of Entropy
    c. Examples: entropy of Bernoulli
    d. Joint entropy
    e. Conditional entropy
    f. Chain rule for entropy
3) KL distance
4) Mutual information?
    a. MI as difference of entropies
    b. Equivalence relations
5) ***Slides: surprise***

Big neuro questions:

- What is the neural code?
- How do neurons convey info?
- How do spikes represent intrinsic/extrinsic variables?

} How do we quantify info? What does it mean?

- Use spiking to infer what is happening in the external world
- Spikes are noisy (don't know distribution)
- Narrow ↓ uncertainty

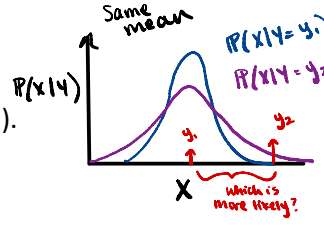
Correlation = Covariance normalized by standard deviation
→ captures linear dependencies b/t 2 random variables

Failures of correlation

Recall last week we talked about how correlation captures the *linear* dependence of two random variables, but doesn't capture non-linear relationships that can cause neurons not to be independent.

Draw pictures on the board that talk about P(Y) vsP(Y|X=x1) vs P(Y|X=x2).

Not independent = ability to share info even though uncorrelated



Mean hasn't changed at all.

Are they independent?
No.

If two variables are independent, then conditioning doesn't Δ the distribution

Point out that really, any time the distribution P(Y|X) is different from the distribution of P(Y), there is the potential for information. That is, any time the variables are not *independent* there can be *information*: intuitively, informativeness has to do with how much the two distributions differ from each other. This is an important point.

Talk about

(1) non-linear changes in mean not captured by correlation
(2) Even if you don't have *any* mean changes, can still get information transfer.

Entropy

Now, if we are going to talk about how informative one variable is about another variable, we need to talk about how random a particular variable is. That is, how random is a random variable?

Which is more random – rolling a dice or flipping a coin?

Obviously, it seems that rolling a dice should be more random than flipping a coin.

What about flipping two coins?

Still, rolling a dice seems more random.

What about flipping three coins?

Now rolling a dice seems less random.

- Shannon wanted a measure that satisfies these conditions:
  1. It should be maximal with p(X) is uniform, and increase with the number of possible values of X.
  2. It should remain the same if we reorder the probabilities assigned to the different values of X.
  3. The uncertainty about two independent random variables should be the sum of the uncertainties of each of them.

Which seems more random, a binomial distribution (5,1/2), or a Poisson distribution(2):

Binomial(5,1/2): $P(K=k)_{Bino}$ = [0.031 0.156 0.313 0.313 0.156 0.03].
Poisson(2): $P(K=k)_{Poiss}$ = [0.135 0.271 0.271 0.180 0.090 0.036 0.012 0.003 …]
Poisson(1): $P(K=k)_{Poiss}$ = [0.368 0.368 0.184 0.061 0.015 0.003 0.001 …]

We need some way of quantifying the "randomness" of a random variable. We do this with the *entropy*.

H(X)

Definition of Entropy:

$$H(X) \equiv -\sum_{x \in X} p(x)\log(p(x)) = E\left[\log\left(\frac{1}{p(x)}\right)\right]$$

expectation; all possible values; weighted sum

Expectation of probabilities of x, not the values of x

(We follow the convention that 0log(0)=0. The limit of xlog(x) as x->0 is 0.)

Note that Entropy is really an expectation – but it's a function of the probabilities, not the values of the random variable! Shifting the random variable (say, adding a constant) doesn't make any difference.

Now, when the logarithm is base 2, we say that the entropy has units of *bits*.

This agrees with our intuition about randomness in a couple of ways. First, note that H(X)≥0. This is because 1≥p(x)≥0.

Base of log = Units of entropy

Some properties:

(i) H(X)≥0 since $0 \le p(x) \le 1$

(ii) H(X)>0 unless p(X) is degenerate (no uncertainty (always heads)) meaning X has a single possible value

(iii) $H_b(X)=\log_b(a)H_a(X)$. Switching the base of the logarithm just multiplies the entropy by a scaling factor. Convention is to use $\log_2$. When ln is used instead, we say the entropy has units of *nats*.

Log base 2 = bits
Log base e = nats

$H(X) = -\sum_{x \in X} p(x)\log(p(x))$

$= -[(1-q)\log(1-q) + q\log(q)]$

$= -q\log(q) - (1-q)\log(1-q)$

$H(0) = 0\log(0) - 1\log(1) = 0$

$0\log(0) = 0 \quad 1\log(1) = 0$

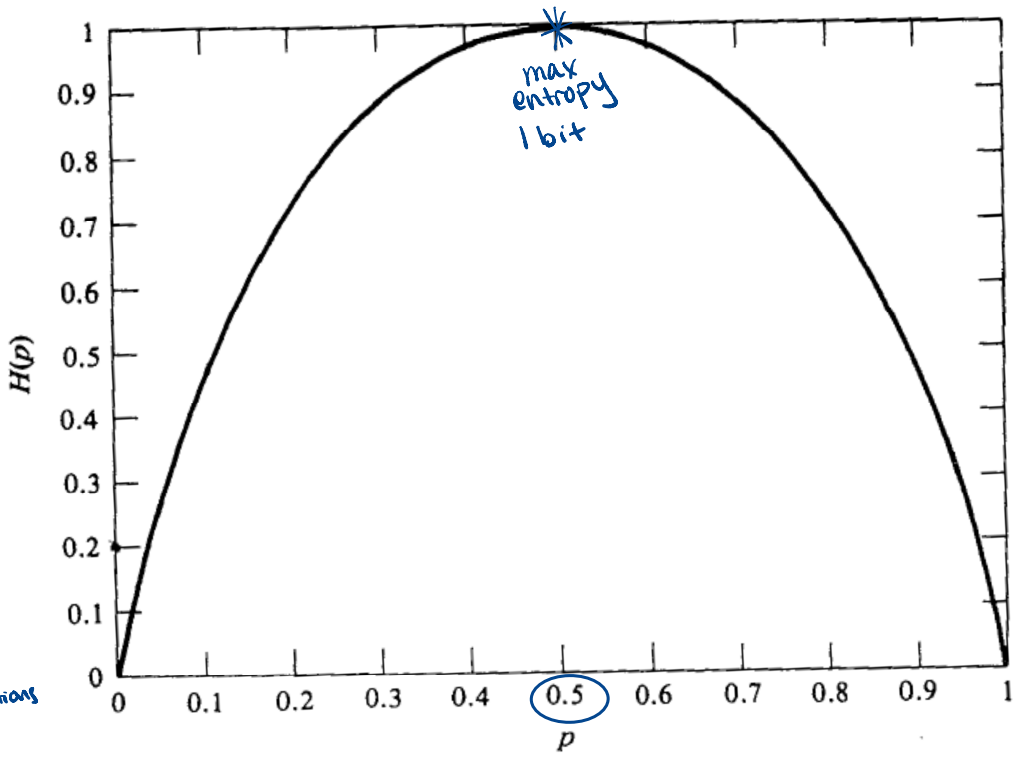
$H(1) = -1\log(1) - 0\log(0) = 0$

Example: Entropy of the Bernoulli distribution. (coin flip)

Assume X~Bernoulli(q). Then H(X) = -qlog(q) – (1-q)log(1-q).

$P(a) = q$

$P(b) = 1-q$



$\log_2(1/2) = -1$

$\log_2(2) = 1$

$\log_2(4) = 2$

$\log_2(8) = 3$

$\log_2(16) = 4$

$\vdots$

$\log_2(2^x) = x$

ex: 6 bits ... on average, would take 6 yes/no questions to pinpoint value

Figure 2.1. $H(p)$ versus p. entropy → on average, # of bits needed to relay outcome of random var (coinflip = need 1 bit)

A couple of things to note.

(1) The entropy goes to zero (degenerate) when q = 0 or 1. This corresponds to our intuition about entropy being a measure of randomness – when q = 0 or 1, there is no randomness to the Bernoulli distribution!

(2) The maximum entropy occurs when q = 0.5, where the entropy is 1 bit. In this case, we have the least *a priori* idea of what will happen when we flip the coin.

Max entropy → $q = 0.5$

$H(0.5) = -(0.5\log(0.5) + 0.5\log(0.5))$

$-((0.5)(-1) + (0.5)(-1))$

$= 1$

Example: entropy of a binomial(4,.5) distribution.

P(k)=choose(4,k)$(½)^k$ $(½)^{n-k}$ = 1/16 * [1,4,6,4,1];

Sum(plogp) = 1/16*[log(16) + 4log(4) + 6log(16/6) + 4log(4) + log(16)] = 1/16*[4+8+8.49+8+4]

H(P) = 32.49/16≅2.03 bits.

H(Poisson(λ=1))~1.88 bits.

Entropy of a joint distribution.

Joint entropy is defined in a very similar manner to regular entropy:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 (p(x,y))$$

Note that if X and Y are independent, then p(x,y) = p(x)p(y). Then…

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x)p(y) \log_2 (p(x)p(y))$$

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x)p(y) (\log_2 (p(x)) + \log_2 (p(y)))$$

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x)p(y) \log_2 (p(x)) - \sum_{x \in X} \sum_{y \in Y} p(x)p(y) \log_2 (p(y))$$

$$H(X,Y) = -\sum_{x \in X} p(x) \log_2 (p(x)) \sum_{y \in Y} p(y) - \sum_{y \in Y} p(y) \log_2 (p(y)) \sum_{x \in X} p(x)$$

$$H(X,Y) = -\sum_{x \in X} p(x) \log_2 (p(x)) - \sum_{y \in Y} p(y) \log_2 (p(y))$$

$$H(X,Y) = H(X) + H(Y)$$

So, if two variables are independent, then their joint entropy is the sum of their individual entropies! This means that the entropy of flipping one coin is 1 bit, and the entropy of flipping 2 coins is 2 bits, and the entropy of flipping n coins is n bits. This corresponds with our intuitive notion of what randomness is.

It should be noted that there is a way to define entropy *axiomatically*. If you start out with the premise that entropy should (1) be non-negative, (2) be zero only when the distribution is degenerate, and (3) add for independent variables, it turns out the only functional form that fits the bill is

$$H(X) = -\sum_{x \in X} p(x) \log_2 (p(x))$$

Also note: H(X,Y) = H(Y,X).

Conditional entropy.

Now, conditional distributions are equivalent to regular distributions, so the definition of entropy for the conditional distribution p(Y|X=x) is simply:

$$H(Y|X=x) = -\sum_{y \in Y} p(y|x) \log_2 \left(p(y|x)\right)$$

See that H(Y|X=x) is a function of x. We define the Conditional Entropy H(Y|X) to be the expected value of H(Y|X=x) over all x:

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x) \qquad H(Y|X) = \mathbb{E}_x\left[H(Y|X=x)\right]$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 \left(p(y|x)\right) \qquad p(x) \cdot p(y|x) \rightarrow \text{joint distribution} = p(x,y)$$

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \left(p(y|x)\right)$$

Some important differences between H(Y|X) and E(Y|X). Remember, E(Y|X) is a random variable! H(Y|X) is **NOT**. It's a **number**.

$\mathbb{E}[Y|X]$ → random variable

$H(Y|X)$ → number

Note: If X and Y are independent, H(Y|X) = H(Y). (because p(y|x) = p(y))

If independent, entropy of X doesn't affect entropy of Y at all. Thus, H(Y) & H(Y|X) are both just H(Y).

Finally, note H(Y|X) ≠ H(X|Y).

$H(Y|X) \neq H(X|Y)$

Chain rule for entropy.

Theorem: H(X,Y) = H(X) + H(Y|X).

joint entropy: $H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(x,y))$

$p(x,y) = p(x)\,p(y|x)$

$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(x)p(y|x))$

$\log_2(p(x)p(y|x)) = \log_2(p(x)) + \log_2(p(y|x))$

$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)(\log_2(p(x)) + \log_2(p(y|x)))$

distribute

$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(x)) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(y|x))$

Pull out just x-stuff w/ x sum

Include negatives!!

$H(X,Y) = -\overbrace{\sum_{x\in X} p(x)\log_2(p(x))}^{H(X)} - \overbrace{\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(y|x))}^{H(Y|X)}$

$$H(X,Y) = H(X) + H(Y|X)$$

Note this also means that H(X,Y) = H(Y) + H(X|Y). Similarly, H(Y)+H(X|Y) = H(X)+H(Y|X).

What does this mean? The entropy in the joint distribution is the entropy of one of the variables, plus the remaining randomness in the other variable after we know the first. It makes intuitive sense, in a way. If X and Y are independent, H(Y|X) = H(Y), and we get the result we proved before. Further, if Y is completely specified by X, then p(y|x) = 1 in only one place and 0 everywhere else. Then H(Y|X)=0, and H(Y,X) = H(X).

KullbackLeibler Distance

The KL Distance, also known as the *relative entropy*, is a measure of the difference between two probability distributions. It is defined in the following way. Say you have two probability mass functions, p(x) and q(x). The KL distance between p and q is:

two distributions defined over same range of values

$$D(p \| q) = \sum_{x \in X} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right)$$

KL dist b/t p & q

· Defined on same outcomes, but what is difference b/t those probability distributions

A couple of things to note. First, this is not a true distance measure: D(p||q) ≠ D(q||p) (and also it turns out it doesn't satisfy the triangle inequality). However, it is often useful to think of the KL distance as measuring the difference between two probability distributions.

We will prove later that the KL distance is non-negative, and is zero iff p(x)=q(x) for all x.

Signal processing theory:

→ send fewest bits possible
→ should only need bits = entropy
→ wrong prob dist for outcome (p instead of q), extra # of bits
→ inefficiency in code

information → uncertainty that gets removed from a random variable when you have knowledge of the other random variable

Mutual Information

Okay, so we now know how to quantify the uncertainty in a given random variable. How do we quantify the information that one variable conveys about another?

What IS information? It's a reduction in uncertainty! And now that we know how to quantify uncertainty, we are ready to quantify information - through a difference in entropies. The information conveyed by one variable about another is defined as the reduction in uncertainty in that variable that you get through conditioning:

remaining uncertainty

$$MI(X;Y) = H(Y) - H(Y|X)$$

$$MI(X;Y) = -\sum_{y\in Y} p(y)\log_2(p(y)) + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(y|x))$$

Replace p(y) with $\sum_x p(x,y)$ $\quad p(y) = \sum_x p(x,y)$

$$MI(X;Y) = -\sum_{y\in Y}\log_2(p(y))\sum_{x\in X} p(x,y) + \sum_{x\in X}\sum_{y\in Y} p(x,y)(\log_2(p(y|x)))$$

$$MI(X;Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(y)) + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2(p(y|x))$$

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)(\log_2(p(y|x)) - \log_2(p(y)))$$

$\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2\left(\frac{p(y|x)}{p(y)}\right)$$

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2\left(\frac{p(y|x)p(x)}{p(y)p(x)}\right)$$

$p(x,y) = p(y|x)p(x)$

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(y)p(x)}\right)$$

$MI(X;Y) = MI(Y;X)$
$= H(Y) - H(Y|X)$
$= H(X) - H(X|Y)$

Note that this definition also means: MI(X;Y) = KL(p(x,y) || p(x)p(y)). ***The mutual information is the KL distance between the joint distribution and the distribution that assumes the variables are independent!***

$MI(X;Y) = KL(p(x,y) \,||\, p(x)p(y))$

p(x)p(y): product of marginal distributions

p(x,y): joint distribution; ||: (dist); p(x)p(y): distribution assumes variables are independent

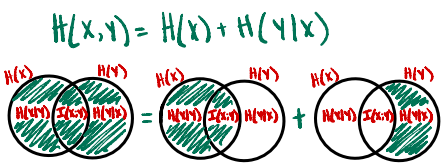
Entropy must be positive, so...

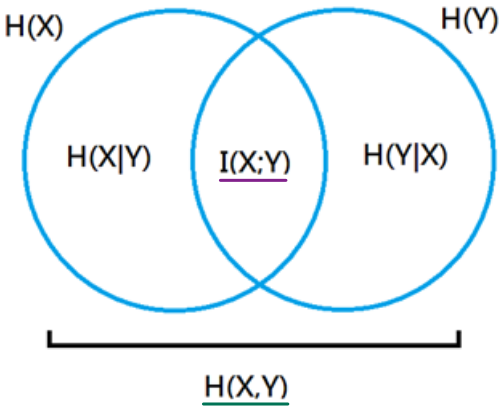
$MI(X;Y) \le H(Y)$
$MI(X;Y) \le H(X)$

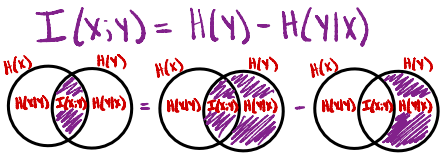
$0 \le MI(X;Y) \le \min(H(X), H(Y))$

(can't give more information than uncertainty in the system)

Some equivalence relations for mutual information.







(i) MI(X;Y) = MI(Y;X). So, if MI(X;Y) = H(Y)-H(Y|X), then MI(X;Y) = H(X)-H(X|Y).
(ii) MI(X;Y) = H(X) + H(Y) – H(X,Y).
(iii) MI(X;X) = H(X). This is why entropy is sometimes called "self-information"

Bounds on MI.

Remember: entropies must be positive. So, what's the maximum value that the mutual information can take?

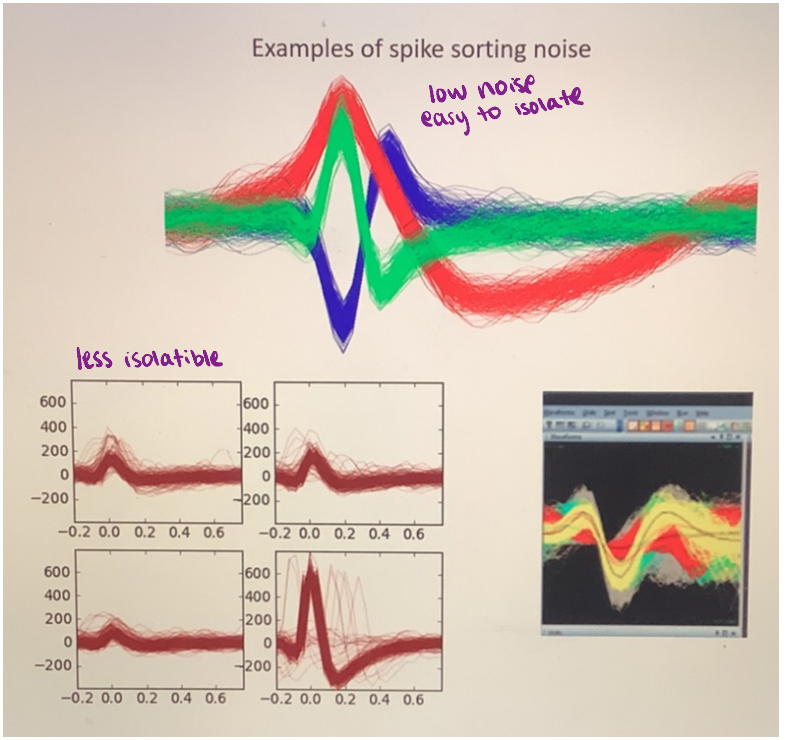
Well, MI(X;Y) = H(X)-H(X|Y). So, clearly, MI≤H(X).

Also, MI(X;Y) = H(Y)-H(Y|X). So, clearly, MI≤H(Y).

Therefore: MI(X;Y) ≤ min(H(X), H(Y))

What's the least it can be?

Motivation for Information theory?



- where electrode is relative to NaVca channels

Noise sources:
→ neurons
→ channels
→ recording tech

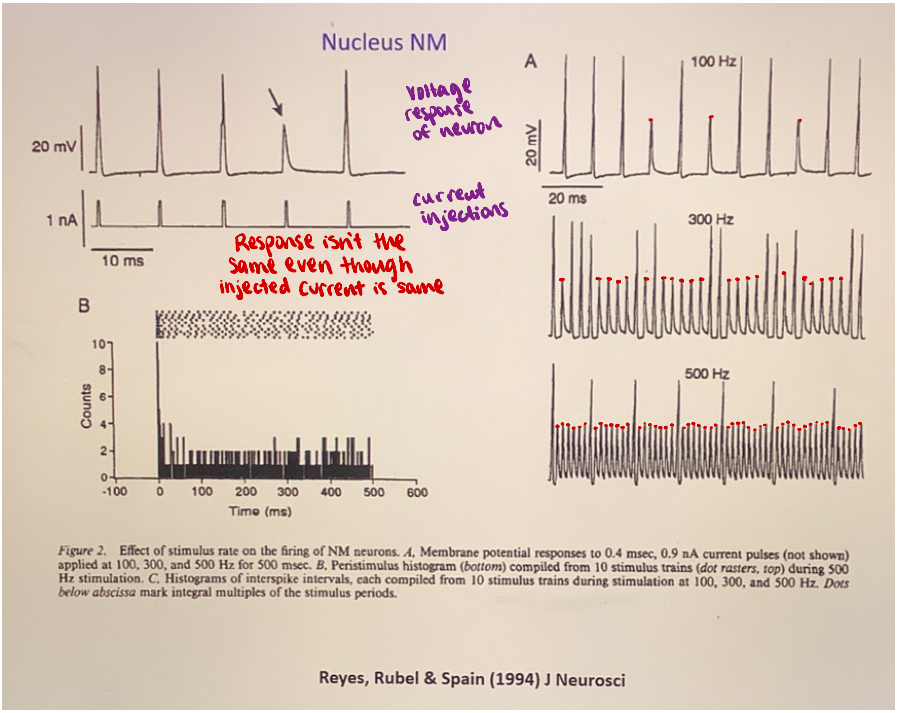
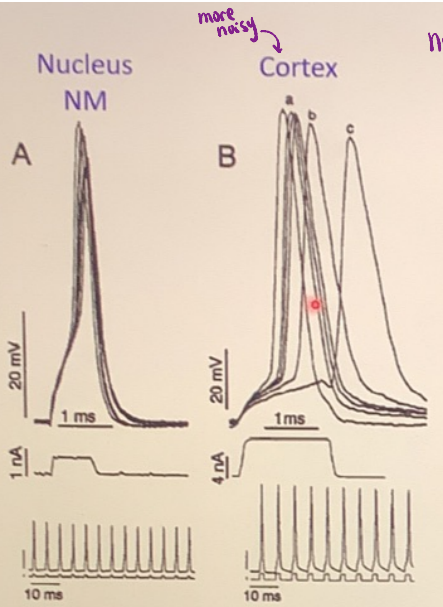


Figure 2. Effect of stimulus rate on the firing of NM neurons. *A*, Membrane potential responses to 0.4 msec, 0.9 nA current pulses (not shown) applied at 100, 300, and 500 Hz for 500 msec. *B*, Peristimulus histogram (*bottom*) compiled from 10 stimulus trains (*dot rasters, top*) during 500 Hz stimulation. *C*, Histograms of interspike intervals, each compiled from 10 stimulus trains during stimulation at 100, 300, and 500 Hz. *Dots below abscissa* mark integral multiples of the stimulus periods.

Reyes, Rubel & Spain (1994) J Neurosci

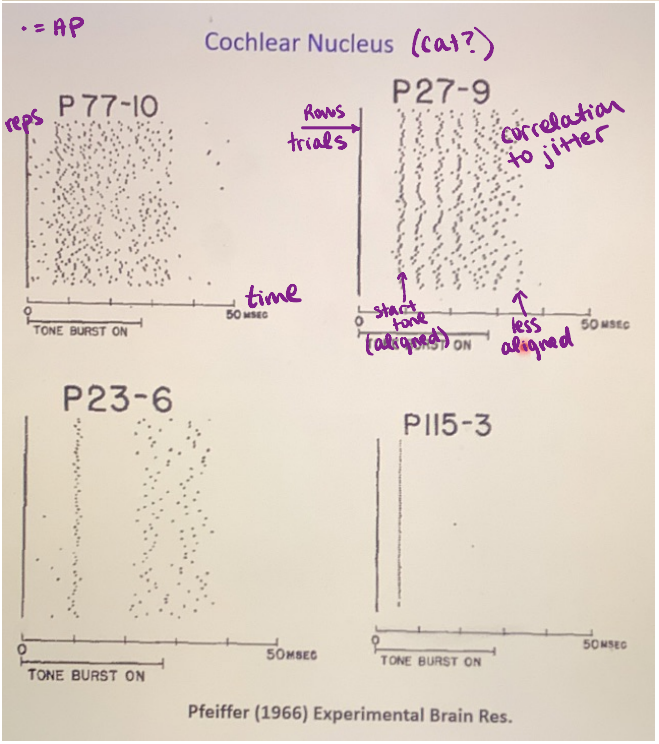
- chicken
- auditory system



Noisiness → part of brain channels

Figure 3. Comparison of phase-locking in NM neurons and cortical neurons. *A*, Superimposed records of the first four, the 20th, and the last action potentials evoked in an NM neuron when current pulses were delivered at 200 Hz for 800 msec. Resting potential was −67 mV. *B*, Superimposed records of the first four (*a*), the 17th (*b*), and the last (*c*) action potentials evoked in a cat layer V pyramidal neuron during stimulation with a 167 Hz, 600 msec pulse train. Resting potential was −70 mV. *Bottom traces* in *A* and *B* show the voltage and current traces during the first 60 msec of the stimulus train. Voltage and current calibration bars are numerically equal to those in *upper traces*.

Reyes, Rubel & Spain (1994) J Neurosci

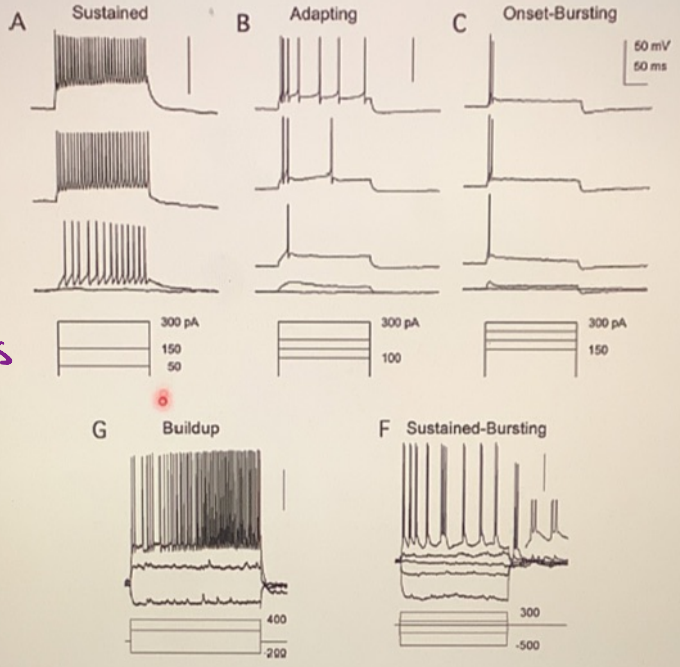


Pfeiffer (1966) Experimental Brain Res.

- Dot rastogram → where do reliable spikes occur?
- 4 different neurons
- response to tone

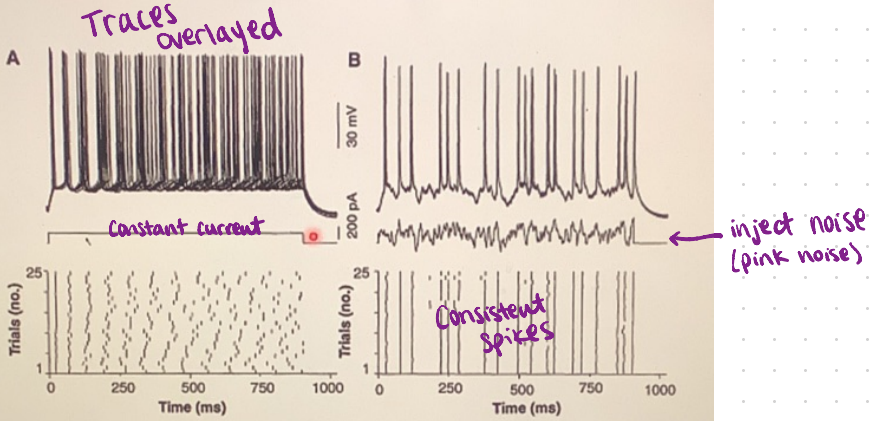
Inferior colliculus

Responses are endogenous



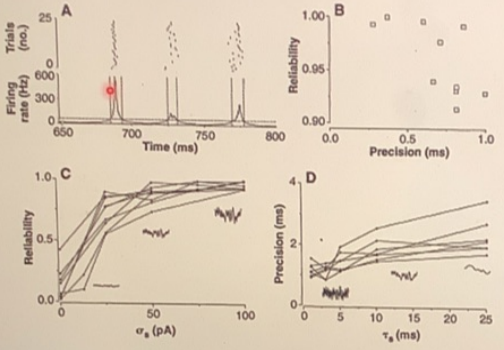
Xie, Gittelman, Li & Pollack (2008) Neuroscience

Cortex



Mainen & Sejnowski (1995) Science.

Cortex

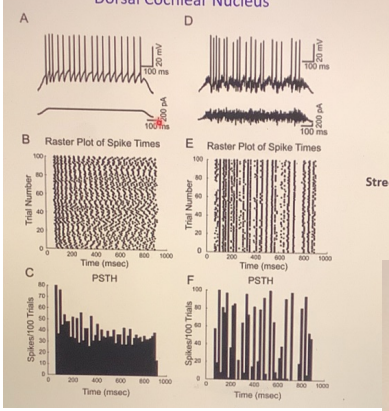


Reliability: fraction of spikes in an "event" (when PSTH crosses threshold)

Precision: std of those spikes

Mainen & Sejnowski (1995) Science.

Dorsal Cochlear Nucleus

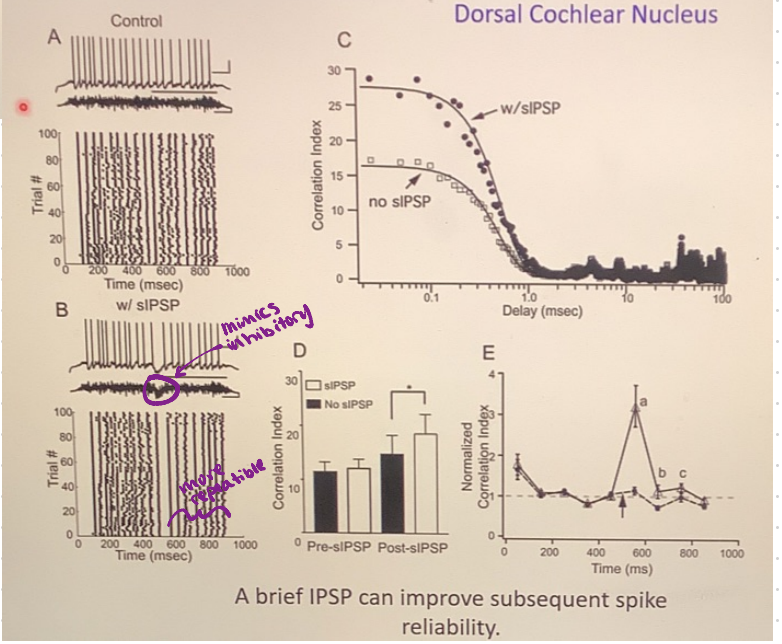


Street & Manis (2007) J Neurophysiol.

If neuron has inputs from upstream, adding noise resets the signals

- Same effect for EPSPs

Dorsal Cochlear Nucleus



A brief IPSP can improve subsequent spike reliability.