86-631/42-6631 Neural Data Analysis
Lecture 7: Information theory and neural coding

1) Recap
2) Mutual information
    a. Conditional information
    b. Joint information
    c. An example.
    d. Jensen's inequality
    e. MI bounds.
    f. Data processing inequality.
3) ***Slides: Changes in whisker deflection coding in the ascending lemniscal pathway (Petersen)***

Recap

Last class, I introduced the concepts of entropy, KL distance, and mutual information, and talked about their interrelationships. I'll briefly recap them here:

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

(randomness)

- always greater than 0
- equal to zero, if p(x)=1 (degenerate)

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

- if X & Y are independent, $p(x,y)=p(x)p(y)$ & $H(X,Y)=H(X)+H(Y)$

2 coin flips → 2 bits
3 coin flips → 3 bits
⋮ ⋮

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

↑ some particular value

KL distance: $D(p \,||\, q) = \sum_{x \in X} p(x) \log_2\left(\frac{p(x)}{q(x)}\right)$ — defined over same x values

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

= $H(Y) - H(Y|X)$ (residual uncertainty)

(Reduction in uncertainty of random variable when told another random variable)

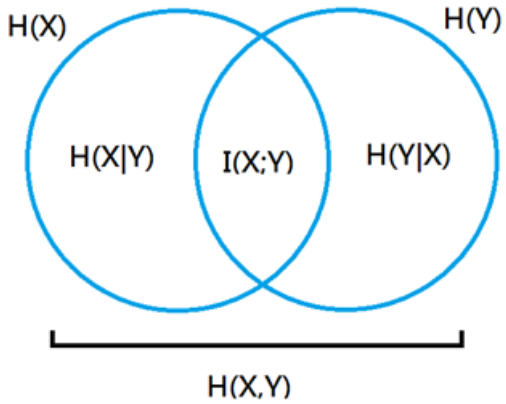
(maximum reduction)

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

$= H(X)$

(self-information)

MI(X;Y) = D(p(x,y)||p(x)p(y)) $MI(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$



Mutual Information

Definition of Conditional Mutual Information:

MI(X;Y|Z) ≡ H(X|Z) – H(X|Y,Z).

Recall that conditional entropy

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

Similarly,

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

Joint Mutual Information

What's the mutual information between one variable and two other variables?

$$\mathrm{MI(S;R)} = \sum_{s\in\{S\}}\sum_{r\in\{R\}} p(s,r)\log_2\left(\frac{p(s,r)}{p(s)p(r)}\right).$$

Suppose S was a vector of parameters, S=[X,Y]. Then

2 random variables — response

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

From this definition, it is not hard to derive the *Chain Rule of Information:*

MI(X,Y;Z) = MI(X;Z) + MI(Y;Z|X). Sketch of proof:

$$\frac{p(x,y,z)}{p(x,y)p(z)} = \frac{p(y,z|x)p(x)}{p(y|x)p(x)p(z)} = \frac{p(y,z|x)p(x)}{p(y|x)p(x)p(z)}\frac{p(z|x)}{p(z|x)} = \frac{p(y,z|x)}{p(y|x)p(z|x)}\frac{p(x,z)}{p(x)p(z)}$$

How do neurons convey information about the external world?

Using Information Theory

Okay, so how do we use mutual information in practice?

Often we are interested in the relationship between a neuron and a particular type of stimuli. Let's take our ubiquitous "oriented bars" example.

Let's say we have 4 bars, at orientations of 0°, 45°, 90°, and 135°. We could present each bar 20 times, and record the number of spikes elicited by the bars over a, say, 250ms period after we present the stimulus.

Now, we could perform a linear regression, and find the tuning curve between spike count and bar orientation. But what if the response is non-linear? Or, what if more subtle features of the response are changing? We really would like to characterize the dependence while making *as few assumptions as possible* about the actual form of the relationship.

So, we compute the MI. Define the stimulus random variable, S, as the probability that each bar was presented. It has 4 values: S_1=0°, S_2=45°, S_3=90°, S_4=135°; each of the 4 values has the same probability: ¼. :: P(s)=1/4 for all s.

The response random variable, R, is the number of spikes we measured in the neuron. We could assume that the response, given a particular bar, should follow a Poisson distribution: $p(R|S=s_i) \sim Poisson(\lambda_i)$. Then all we'd have to do is estimate λ_i for each bar to give us $p(R|S=s_i)$. From that, we could compute p(r,s) and p(r) (How?). Then we could compute the MI!

But what if we don't know the response follows a Poisson distribution? We could empirically determine the distribution. We could say:

$p(r|s_i)$ =(# of times we observe r when stimulus s_i is played) / total # of times stimulus s_i is played. (20)

Then we can proceed as before. Done! We will go over these methods in the homework sets you will do in the next couple of weeks.

But so what? What use is knowing this? Who cares if the information is 0.5 bits vs. 1.2 bits?

There are a few things that you can do to actually get useful scientific information out. For example, you could compare the number of bits you measure to the number of bits you would measure *assuming the spike count was linearly related to the stimulus!* This would tell you how much information you'd be throwing away be restricting yourself to linearity assumptions. The other thing you could do is compare the average amount of information carried by, say, one type of neuron versus another type of neuron. Maybe type I conveys more information about orientation but less about contrast, and type II the opposite. Then you'd be getting somewhere! In the slides today, I'll talk about another study where they used MI to say something interesting.

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?

Jensen's Inequality

For MI to correspond to our intuitive notions about information, it should be non-negative. It is. To prove it requires Jensen's Inequality, which is useful in other contexts to come, so it's worth really going over right now.

Definition of a convex function.

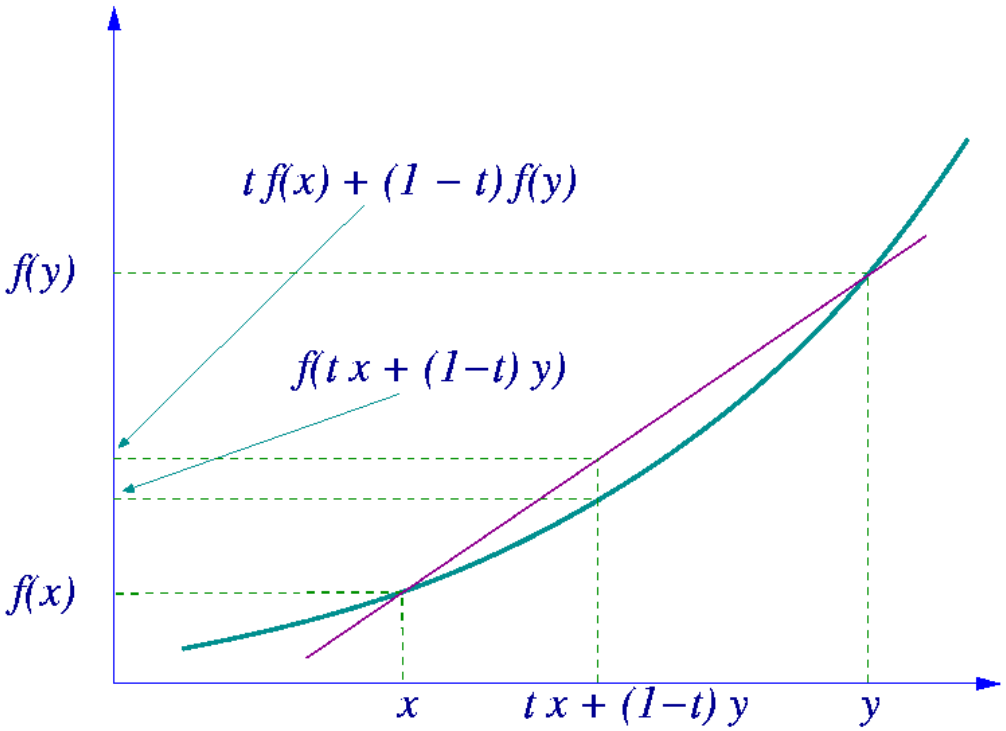
A function f(x) is said to be *convex* over an interval (a,b) if for every x_1, $x_2 \in (a,b)$ and $0 \le t \le 1$,

$$f\left(tx_1 + (1-t)x_2\right) \le tf\left(x_1\right) + (1-t)f\left(x_2\right)$$

It is *concave* if –f is convex.

Note: To be convex, it must always lie below any chord. To be concave, it must always lie above any chord.

CONVEX:



NOTE: The log function is concave.

Theorem: Jensen's inequality If f is a convex function and X is a random variable, then $E[f(X)] \geq f(E[X])$.

Proof – by induction:

Let's take a distribution that only has two mass points, at x_1 and x_2 (like the Bernoulli distribution). Let $p(x_1) = p_1$ and $p(x_2) = p_2$. Note that $p_2 = 1-p_1$.

Then

$E[f(x)] = p_1f(x_1)+p_2f(x_2) = p_1f(x_1)+(1-p_1)f(x_2) \geq f(p_1x_1+(1-p_1)x_2)$ by the definition of a convex function.

But $f(p_1x_1+(1-p_1)x_2) = f(E[X])$! So $E[f(x)] \geq f(E[X])$

Now, suppose the theorem is true for distributions with n-1 mass points. Now, let $p_i'=p_i/(1-p_n)$ for i=1,2,…,n-1. Then we have:

$$E[X]=\sum_{i=1}^{n} p_i f(x_i)= p_n f(x_n)+(1-p_n)\sum_{i=1}^{n-1} p_i' f(x_i)$$

But the RHS is a distribution with n-1 mass points. And since our theorem holds for these distributions (by the induction assumption), we have

$$E[X]=\sum_{i=1}^{n} p_i f(x_i)\geq p_n f(x_n)+(1-p_n) f\left(\sum_{i=1}^{n-1} p_i' x_i\right)$$

$$E[X]=\sum_{i=1}^{n} p_i f(x_i)\geq p_n f(x_n)+(1-p_n) f\left(\sum_{i=1}^{n-1} p_i' x_i\right)$$

And now the RHS follows the definition of a convex function exactly!

$$E[X]=\sum_{i=1}^{n} p_i f(x_i)\geq f\left(p_n x_n+(1-p_n)\left(\sum_{i=1}^{n-1} p_i' x_i\right)\right)$$

$$E[X]=\sum_{i=1}^{n} p_i f(x_i)\geq f(E[X])$$

Note that this proof can be extended to continuous distributions.

Some consequences of this theorem.

The information inequality theorem:

Let p(x), q(x), x∈X be two probability mass functions. Then D(p||q)≥0 with equality iff p(x)=q(x) for all x.

Proof: Let A={x:p(x)>0} be the support set of p(x). Then

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

$$-D(p\|q) \le \log\left(\sum_{x\in X} p(x)\left(\frac{q(x)}{p(x)}\right)\right) = \log\left(\sum_{x\in X} q(x)\right) = \log(1) = 0$$

$$D(p\|q) \ge 0$$

where the inequality step follows from Jensen's. And since log(x) is a strictly concave function of x, we have equality iff q(x)/p(x) = 1 everywhere, ie, p(x)=q(x).

Corollary:

MI(X;Y)≥0, with equality iff X and Y are independent.

Proof: MI(X;Y) = D(p(x,y)||p(x)p(y)).

Corollary:

D(p(y|x)||q(y|x))≥0, with equality iff p(y|x)=q(y|x) for all y and x with p(x)>0.

Corollary:

MI(X;Y|Z) ≥0, with equality iff X and Y are conditionally independent given Z.

Proof: MI(X;Y|Z) = MI(X|Z;Y|Z) = D(p(x|z,y|z)||p(x|z)p(y|z))≥0

The uniform distribution has maximum entropy theorem:

Not covered in class. Now as a homework problem.

Theorem: Conditioning reduces entropy

$H(X|Y) \leq H(X)$ with equality iff X and Y are independent.

Proof: $0 \leq MI(X;Y) = H(X) - H(X|Y)$

In other words, knowing the value of one random variable can *only* reduce the uncertainty in X. If the variable is unrelated to X, it does nothing to the uncertainty, otherwise it reduces it.

NOTE: This is not true for individual values of y! $H(X|Y=y)$ might be greater than, less than, or equal to $H(X)$. However, on average, $H(X|Y) = \text{sum\_y}\{p(y)H(X|Y=y) \leq H(X)$.

Y\X	1	2
1	0	1/2
2	1/4	1/4

Example: P(X) = [¼,¾] P(Y)=[½,½]. H(X) = 0.8113 bits. H(Y)=1 bit. H(X,Y) = 1.5 bits.

$H(X|Y=1)=0$ bits $< H(X)$. $H(X|Y=2) = 1$ bit $> H(X)$. $H(X|Y) = ½*0+½*1 = 0.5$ bits $< H(X)$.

The data processing inequality.

Definition of a Markov chain:

Random variables X,Y, and Z are said to form a Markov Chain (X->Y->Z) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, they form a Markov Chain if the joint probability mass function can be written as

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

(In general, $p(x,y,z)=p(z,y|x)p(x) = p(z|y,x)p(y|x)p(x)$. The difference is that $p(z|y,x)=p(z|y)$.)

Note: If Z=f(Y), then X->Y->Z. Even if Y depends on X, if we know Y, we have everything we need to know Z without referring back to X.

Data processing inequality theorem:

If X->Y->Z, then $I(X;Y)\geq I(X;Z)$.

Now, the joint information between X and Y,Z is:

$MI(X;Y,Z) = MI(X;Y) + MI(X;Z|Y)$. Also,

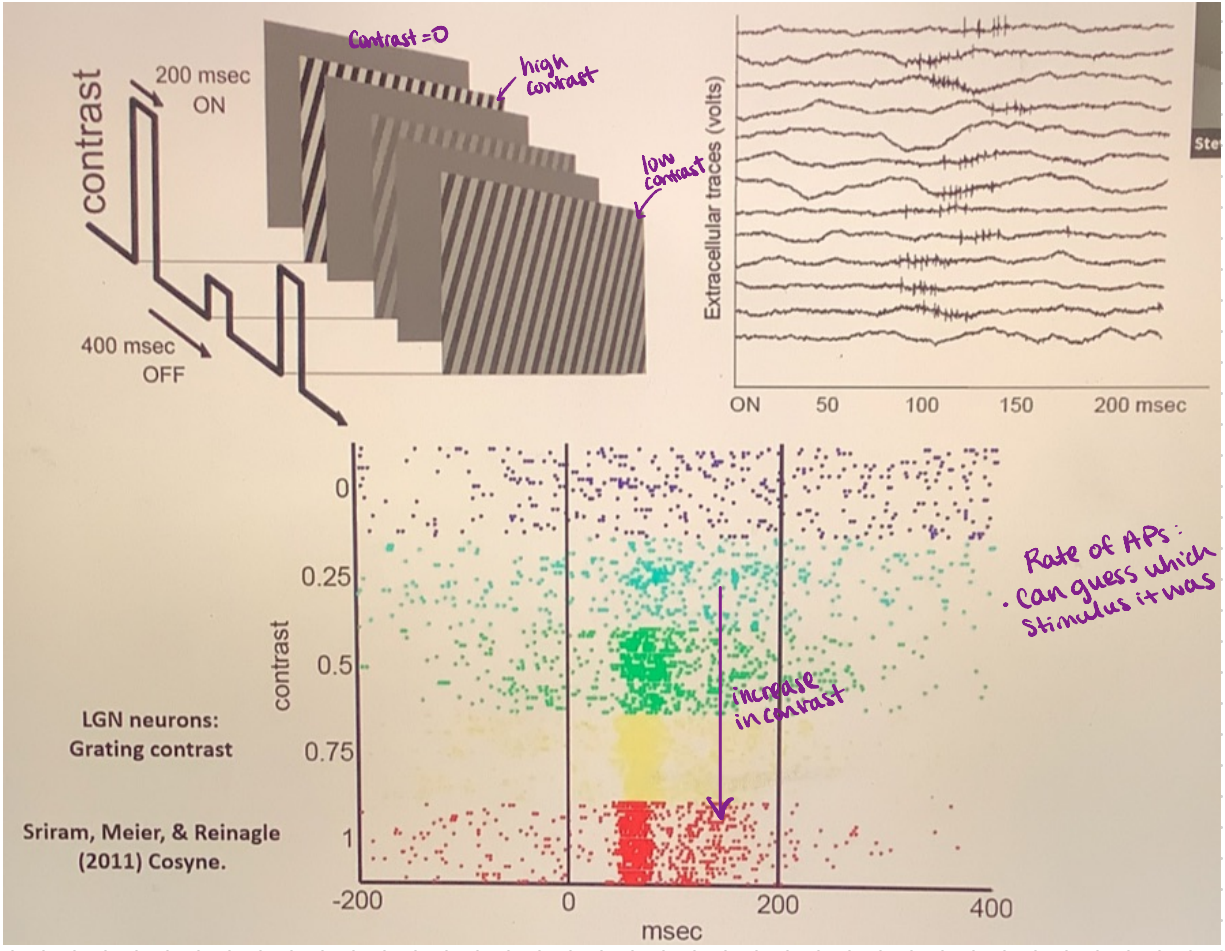
$MI(X;Y,Z) = MI(X;Z) + MI(X;Y|Z)$.

However, if Z|Y is independent of X, then $MI(X;Z|Y) = 0$! So, $MI(X;Y,Z) = MI(X;Y) = MI(X;Z) + MI(X;Y|Z)$.

Also, since we know $MI>0$, we must have $MI(X;Y)\geq MI(X;Z)$

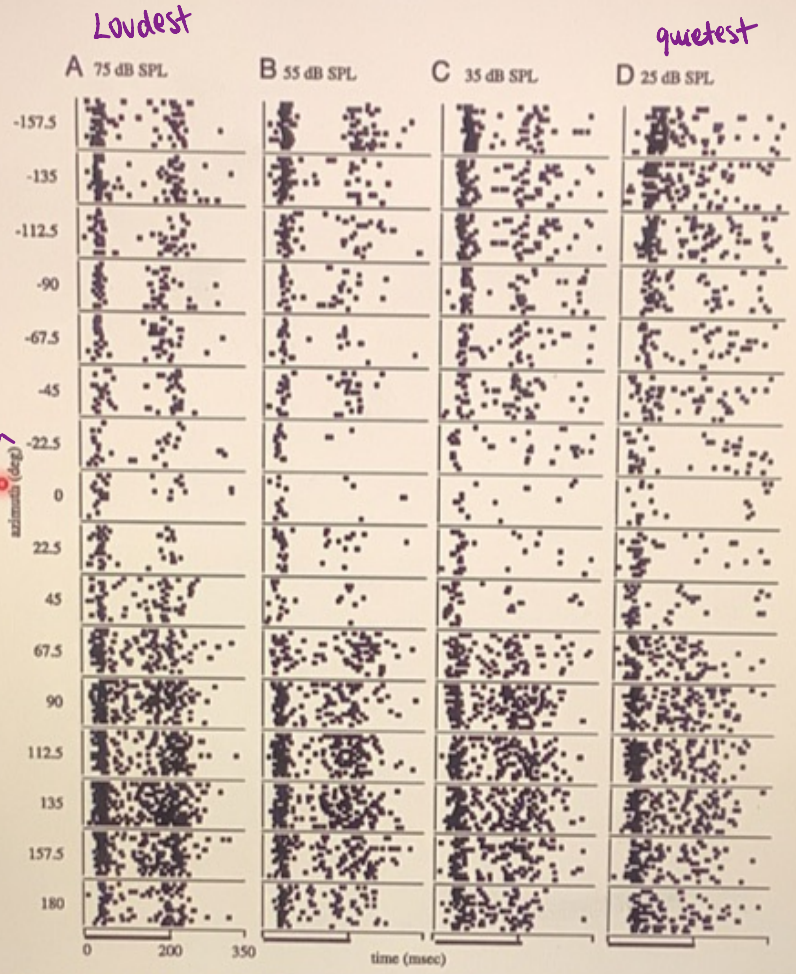
Visual System LGN/thalamus

Bandpass filtering



Rate of APs:
- Can guess which stimulus it was

Orientation matters more than volume of tone



(orientation in space)

A1 neurons:
Sound location and loudness

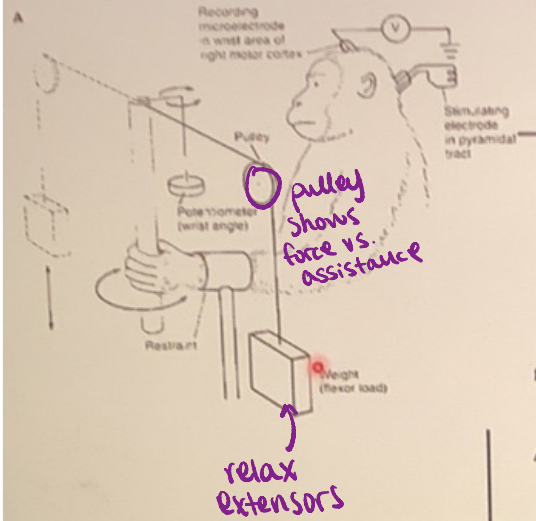
Woods T M et al. J Neurophysiol 2006;96:3323-3337

FR can give info about azimuth in primary auditory cortex

Direction of arm movements

M1 neurons:
Arm movement direction

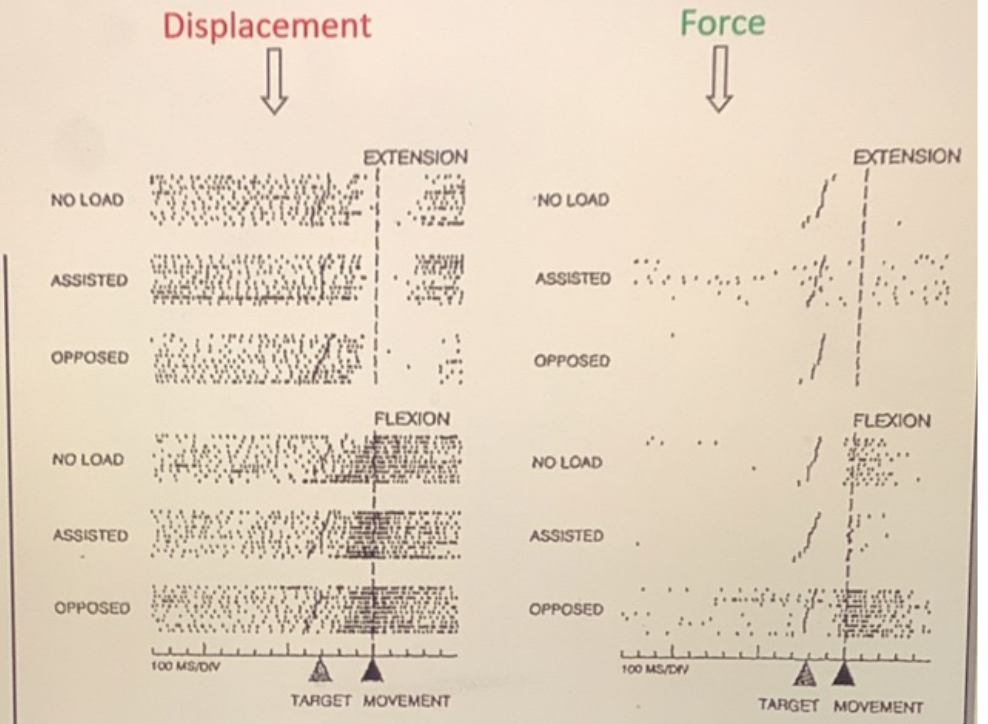
Moran & Schwartz (1999) J Neurophysiol



pulley shows force vs. assistance

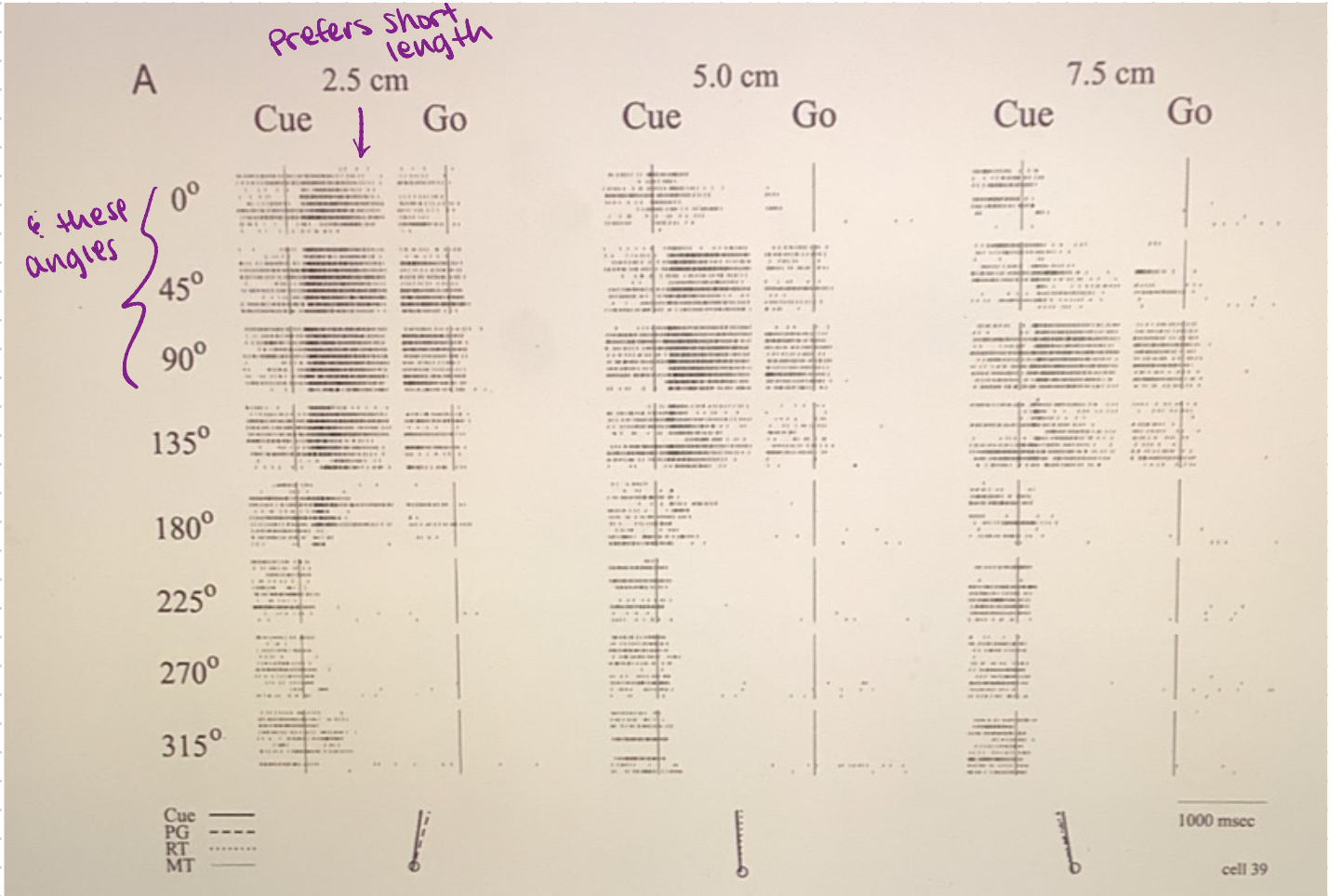
relax extensors

Neurons in motor
Subdivide for preference



M1 neurons:
Force and displacement

Crutcher & Alexander 1990



Rate in motor system

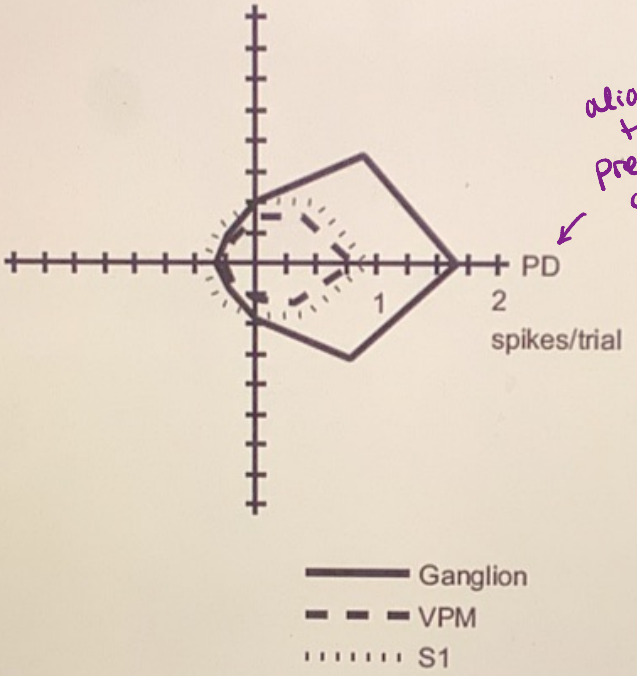
PMd neurons: Movement direction and distance

Messier J , Kalaska J F J Neurophysiol 2000;84:152-165

Journal of Neurophysiology

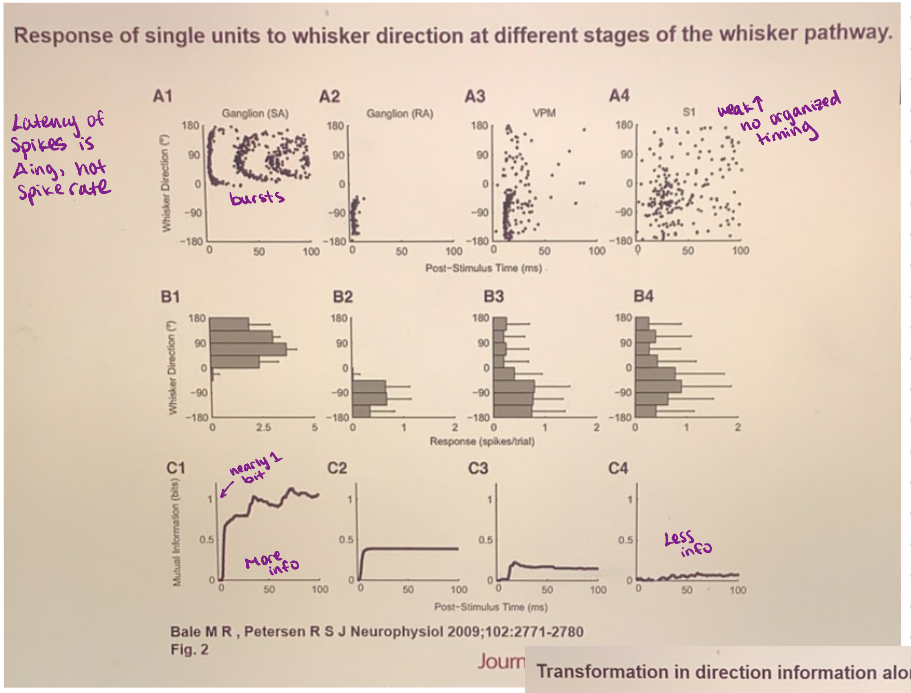
Rat whiskers

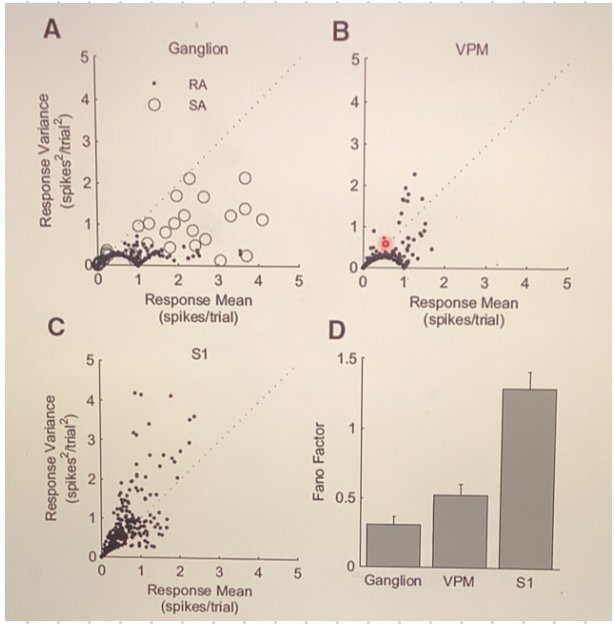
- Whisker deflection

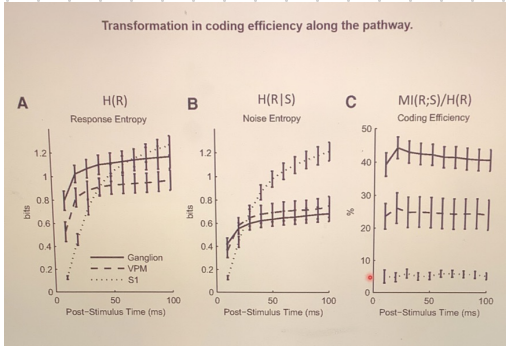


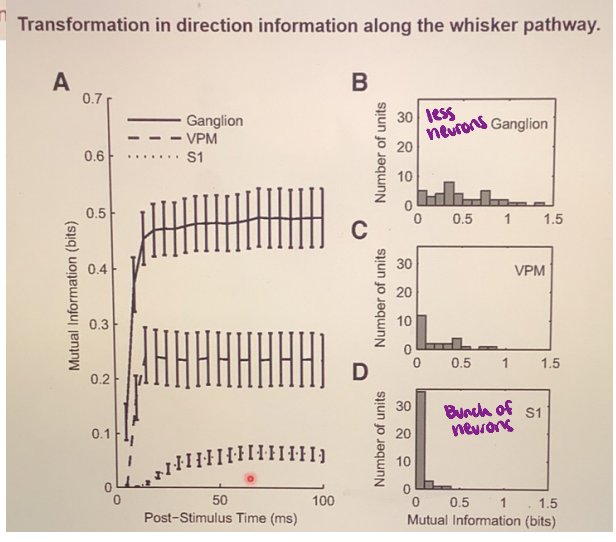
Bale M R , Petersen R S J Neurophysiol 2009;102:2771-2780
Fig. 1a

Journal of Neurophysiology









Conclusion

- Information for each neuron changed (↓) (ganglion → vpn → S1)
- Number of neurons representing stimulus increases on same pathway