

“Complexity Through Nonextensivity”

Bialek, Nemenman, & Tishby

and

“Predictability, Complexity and Learning”

Bialek, Nemenman, & Tishby

Summary:

- **Complexity.**

- Motivations.
- Definitions.
- Ties with predictions and learning
- An example...

- **Predictive information**

- Definitions.
- Extensivity, (and non!)

- **Learning**

- A test case.
- Parametric and non-parametric models.

- **Discussion.**

Complexity.

• Motivations.

- Make precise notions that some systems in the physical world evolve toward more complex states: i.e. classify those states.
- Quantifications of complexity can be used as Occam factors, i.e. penalties for more complex models over simple ones.
- General question of how hard it is to compute or describe the state of a complex system.

• Definitions

- Complexity as a measure of the complexity of the underlying dynamics of the system.
- Probabilistic or not?
- Distinguishing complexity from randomness...
 - * Algorithmic complexity.

• Our intuitive notion of complexity ties in with prediction and learning...

- Low complexity means ease of prediction...
- Learning is finding a model that explains or describes a set of observations...more complex systems mean more to learn.

An illustrative example from statistical mechanics.

- One dimensional chain of Ising spins.

$$H = - \sum_{i,j} J_{ij} \sigma_j \sigma_i \quad (1)$$

- Long range interactions allowed.
- Examine three cases:
 - * Nearest neighbor interactions with $J_{i,i+1} = 1$ nonzero.
 - * Nearest neighbor interactions with the strength of the coupling $J_{i,i+1}$ changed randomly every p spins.
 - * Strength of $J_{i,j}$ changed with same frequency as above but long range interactions allowed.
- Sequence consists of 2^N different overlapping words of length N .
- By counting up and binning these words we can calculate the entropy of these words for different wordlengths.

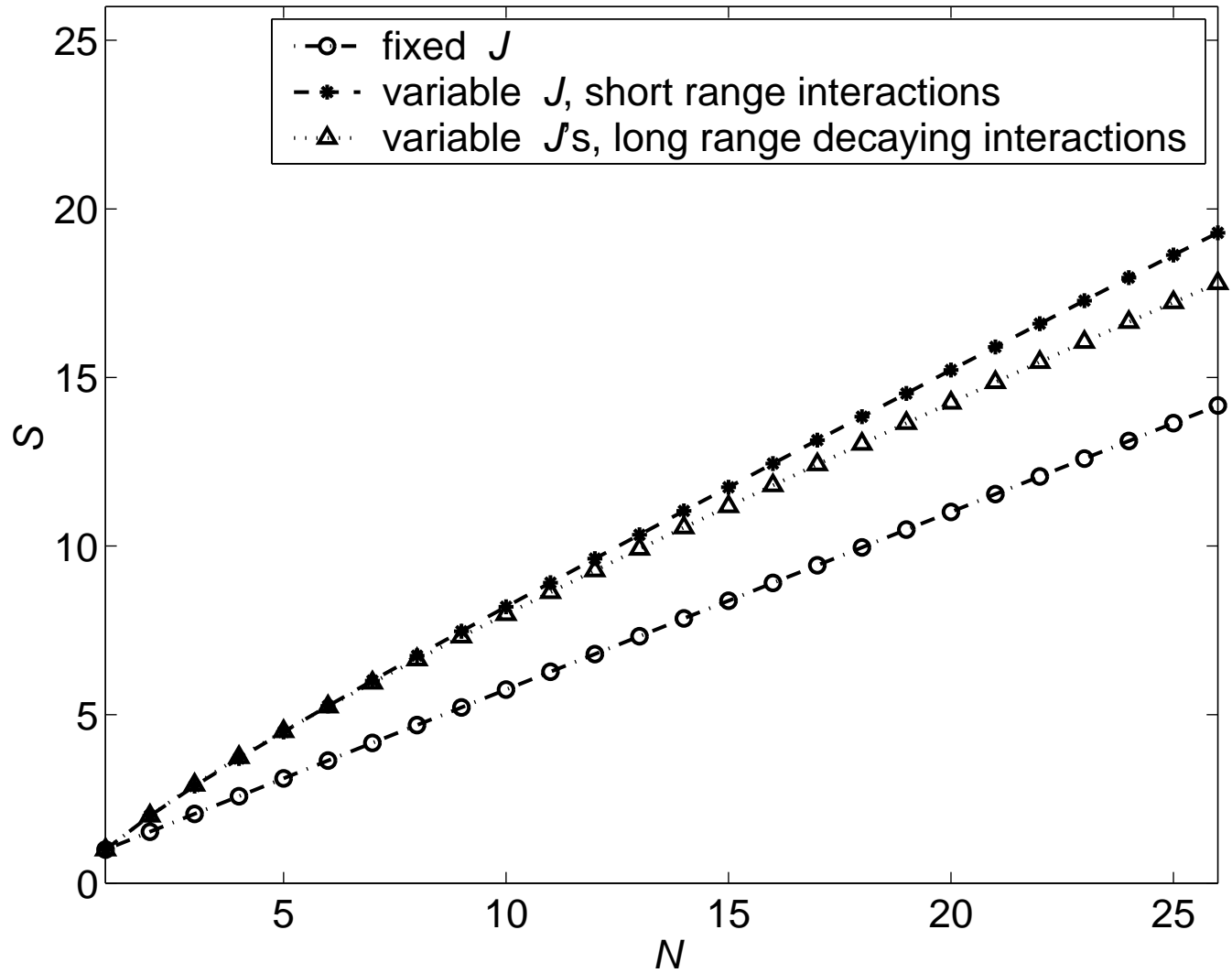


Figure 1: From “Predictability, complexity and learning”, Bialek, Nemenman & Tishby.

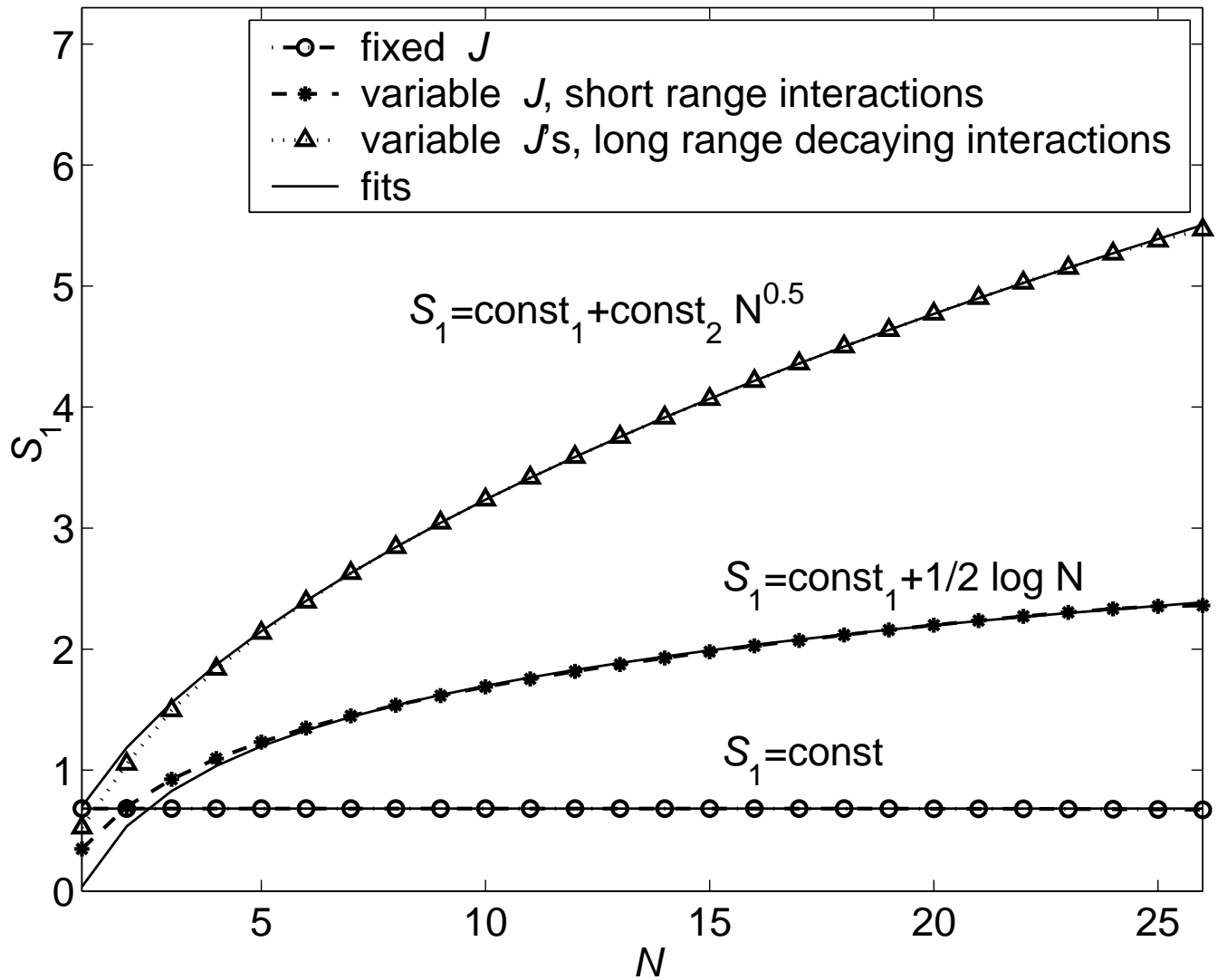


Figure 2: From “Predictability, complexity and learning”, Bialek, Nemenman & Tishby.

Predictive Information.

- Prediction an important part of neural computation.
 - Predictive info the only useful info?
- Generalization, in the learning sense is prediction.
- All predictions are probabilistic with priors.
- Our observations of the past lead to a tightening of the prior distribution of futures.
- Define the data stream $x(t)$ observed over an interval $-T < t < 0$ as x_{past} . We want to say something about a future observation, say in the time interval $0 < t < T'$ (x_{future}).
 - Futures are drawn from prior future distribution, $P(x_{\text{future}})$.
 - Observations of x_{past} tell us that futures will be drawn from conditional distribution, $P(x_{\text{future}}|x_{\text{past}})$.
 - The reduction in entropy of the prior distribution is quantified in Shannon's *predictive information*:

$$\mathcal{I}_{\text{pred}}(T, T') = \left\langle \log_2 \left[\frac{P(x_{\text{future}}|x_{\text{past}})}{P(x_{\text{future}})} \right] \right\rangle \quad (2)$$

$$= -\langle \log_2 P(x_{\text{future}}) \rangle - \langle \log_2 P(x_{\text{past}}) \rangle \\ - [-\langle \log_2 P(x_{\text{future}}, x_{\text{past}}) \rangle] \quad (3)$$

Predictive Information Continued...

- Interested in prediction, i.e. elements of the data stream that persist forever \rightarrow assume time translation invariance.
 - Entropy of past only depends on duration of observation... $S(T)$, and by the same argument $-\langle \log_2 P(x_{\text{future}}) \rangle = S(T')$.
 - Similarly, $-\langle \log_2 P(x_{\text{future}}, x_{\text{past}}) \rangle = S(T + T')$
 $\mathcal{I}_{\text{pred}}(T, T') = S(T) + S(T') - S(T + T')$. (4)
- Predictability is a deviation from extensivity.
 - Recall that entropy is extensive: $\lim_{T \rightarrow \infty} S(T)/T = \mathcal{S}_0$
 - These extensive parts cancel out of the equation for the predictive information.
 - If we write $S(T) = \mathcal{S}_0 T + S_1(T)$ then the predictive information is only related to $S_1(T)$.
 - What do we know about $S_1(T)$ '?

$$\lim_{T \rightarrow \infty} \frac{S_1(T)}{T} = 0. \quad (5)$$

* I.e. $S_1(T)$ must grow with T less rapidly than linear.

* Also, if we let the future extend forward for a very long time, $T' \rightarrow \infty$, then

$$I_{\text{pred}}(T) = \lim_{T' \rightarrow \infty} \mathcal{I}_{\text{pred}}(T, T') = S_1(T). \quad (6)$$

– Three cases:

* $\lim_{T \rightarrow \infty} S_1(T) = \text{constant}$

* $\lim_{T \rightarrow \infty} S_1(T) \propto \log T$

* $\lim_{T \rightarrow \infty} S_1(T) \propto T^n \quad 0 < n < 1$

– Note that most of what we observe is not predictive!

$$\lim_{T \rightarrow \infty} \frac{\text{Predictive Information}}{\text{Total Information}} = \frac{I_{\text{pred}}(T)}{S(T)} \rightarrow 0. \quad (7)$$

Learning a Simple Parametric model.

- Consider two streams of data, x and y , or equivalently a stream of pairs $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$. Assume that we know in advance that the x 's are drawn independently and at random from a distribution $P(x)$, while the y 's are noisy versions of some function acting on x ,

$$y_n = f(x_n; \alpha) + \eta_n, \quad (8)$$

- Start with very simple case in which function class is a linear combination of basis functions...

$$f(x; \alpha) = \sum_{\mu=1}^K \alpha_{\mu} \phi_{\mu}(x). \quad (9)$$

- After a bit of calculation we get the dominant subextensive piece of the entropy,

$$S_1(N) \rightarrow \frac{K}{2} \log_2 N \quad (\text{bits}). \quad (10)$$

- Note that the coefficient of the log is the number of parameters in the model over 2. Also note that $S_1(N)$ is independent of the coordinate system and of the basis functions we chose!
- Learning Curves

- Consider case where time is measured in discrete steps: N time steps x_1, x_2, \dots, x_N .
- The more we know the better we can predict the next data point, x_{N+1} , and hence the fewer bits we will need to describe the deviation of this datapoint from our observations. Let's make this quantitative...
- Average length of code word necessary to describe the point x_{N+1} given observation of the previous N points is

$$\ell(N) = -\langle \log_2 P(x_{N+1} | x_1, x_2, \dots, x_N) \rangle \text{ bits}, \quad (11)$$

or

$$\ell(N) = S(N+1) - S(N) \approx \frac{\partial S(N)}{\partial N}. \quad (12)$$

- If we define an cost as the difference of the above and an ideal length, $\ell_{\text{ideal}} = \lim_{N \rightarrow \infty} \ell(N)$ then we get,

$$\Lambda(N) \equiv \ell(N) - \ell_{\text{ideal}} \quad (13)$$

$$\begin{aligned} &= S(N+1) - S(N) - \mathcal{S}_0 \\ &= S_1(N+1) - S_1(N) \\ &\approx \frac{\partial S_1(N)}{\partial N} = \frac{\partial I_{\text{pred}}(N)}{\partial N}, \end{aligned} \quad (14)$$

- Learning curve is the derivative of the predictive information.

- How is this related to something more familiar?
- * Authors claim that in the large N limit, this universal learning curve is related to the “ χ^2 for generalization” which they write down as,

$$\begin{aligned}\langle \chi^2(N) \rangle &= \frac{1}{\sigma^2} \langle [y - f(x; \alpha)]^2 \rangle \\ &\rightarrow (2 \ln 2) \Lambda(N) + 1,\end{aligned}\quad (15)$$

Subject this to a simple test: generate data pairs, $\{x_n, y_n\}$, by choosing the x_n from a gaussian distribution and generate the y_n from

$$y_n = f(x_n; \alpha) + \eta_y \quad (16)$$

where η_y is gaussian with standard deviation σ_y .

- * Now, for each value of N , do cross-validation:
- * divide the data into p parts.
- * train on $p - 1$ parts (i.e. estimate the α using these data.)
- * Calculate the $\chi^2(N, p_i)$ for the compliment of the training data, (validation set), using the estimated α from the training data (note that we also estimate σ_y from the training data).
- * Do this p times by using each slice as the validation set.
- * Compute the average χ^2 for generalization as

$$\langle \chi^2(N) \rangle = \frac{p}{N} \frac{1}{p} \sum_{i=1}^p \chi^2(N, p_i)$$

$$= \frac{1}{N} \sum_{i=1}^p \chi^2(N, p_i) \quad (17)$$

- * This average chi squared minus 1 should go to $2 \ln 2 \Lambda(N)$ in the limit of large N .
- * In this example I chose a three parameter (linear) model ¹ and thus the subextensive part of the entropy goes as,

$$S_1(N) = \frac{K}{2} \log_2(N) \quad (18)$$

and the universal learning curve as defined above is

$$\Lambda(N) = \frac{\partial S_1(N)}{\partial N} = \frac{K}{2 \ln 2} \frac{1}{N} \quad (19)$$

Thus

$$\begin{aligned} \langle \chi^2(N) \rangle - 1 &\rightarrow 2 \ln 2 \Lambda(N) \\ &= \frac{3}{N} \end{aligned} \quad (20)$$

¹The three parameters are the y intercept and slope of the linear fit and the standard deviation of the distribution of y's. All of these were "learned" from the training data.

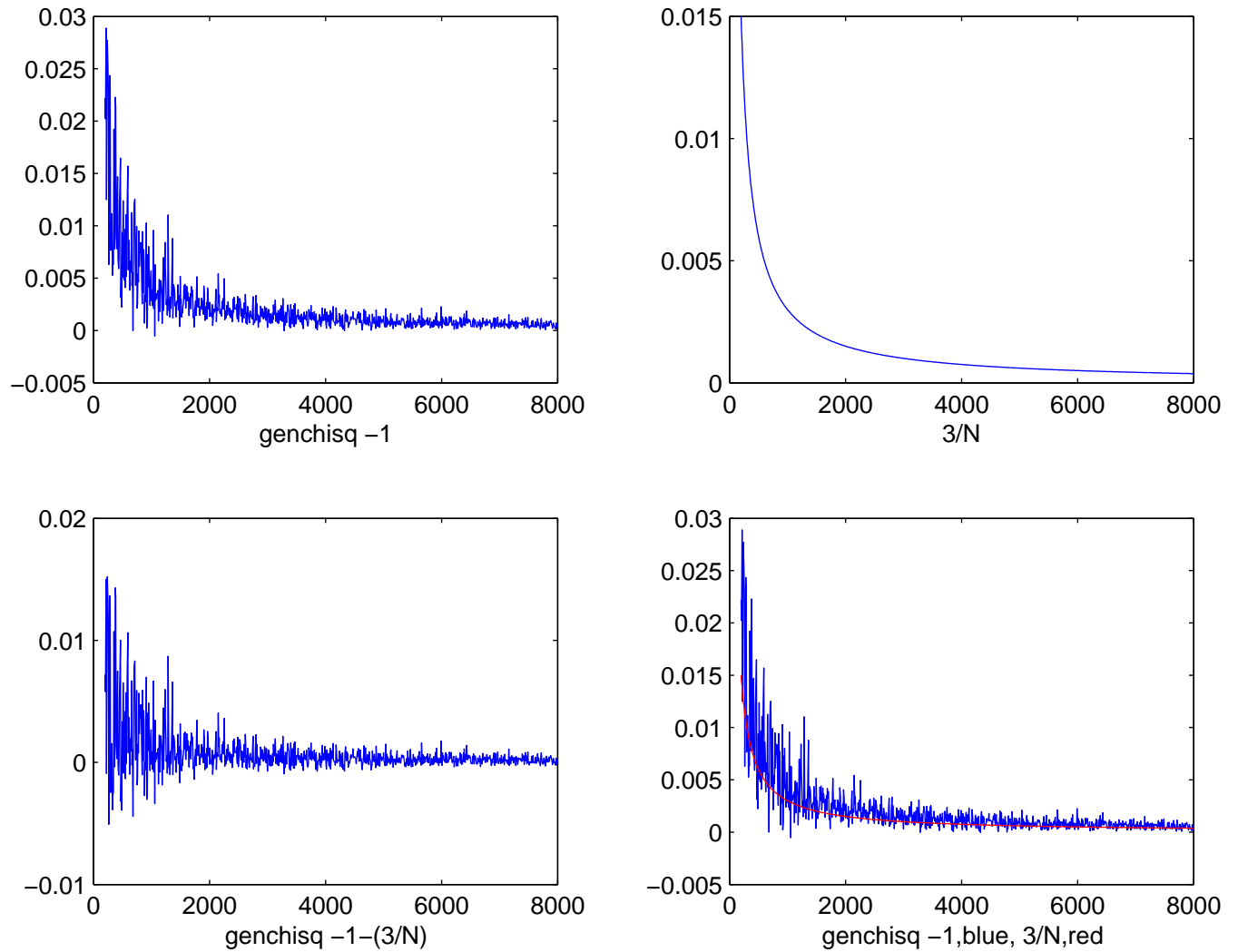


Figure 3: Simulations of learning a linear model. Top left: the chi squared for generalization -1. Top right: the theoretical asymptotic limit of the chi squared for generalization from equation 20. Bottom left: The error between $\langle \chi^2 \rangle - 1$ and the theoretical limit. Bottom right: The two curves plotted on top of each other. $\langle \chi^2 \rangle - 1$ in blue and $(2 \ln 2) \Lambda(N)$ in red. (Simulations by A. K. Schenk)

Learning more general models.

- Authors have shown that one gets the same result for more general K -parameter models, i.e. ones in which one has to learn some distribution $Q(\vec{x}|\alpha)$ of vectors \vec{x} that depends on a K long vector of parameters, α , as one gets for the simplest of parametric models. I.e.

$$S_1(N) \rightarrow \frac{K}{2} \log_2(N) \quad (21)$$

- Nonparametric models and models with growing number of parameters.
 - Example of nonparametric model: Learning a distribution $Q(x)$ for a continuous variable x , but rather than writing a parametric form of $Q(x)$, assume only that this function itself is chosen from some distribution that enforces a degree of smoothness.
 - * This example gives a subextensive part of the entropy that goes as \sqrt{N}
 - Example of model where number of parameters grow with more and more observation is English. As one reads for longer and longer times, one adds more and more parameters to the model of how English works. Some calcs have shown that this

system also has a subextensive entropy that diverges as \sqrt{N} .

Discussion

- If predictive information is the only kind of info that is useful to an organism then maybe the goal of the nervous system is to provide an efficient representation of the *predictive information*. This is very different from the notion that the nervous system is tuned to provide an efficient representation of the current state of the world. Can investigate this point by asking if the information which neural responses provide about the future of the inputs (say natural stimuli) is close to the limit set by the statistics of the input itself
- We know that the nervous system can learn probabilistic models, thus maybe these results could be used to analyze the dynamics of this learning... i.e. measure the universal learning curve. This observed learning curve will be limited by the predictive information in the time series of the stimulus trials. Comparing these two can give and measure of learning efficiency, which has as of yet not really been looked at in the brain.