Generating normative predictions with a variable-length rate code

S. Thomas Christie (tchristie@umn.edu)
Cognitive Science, 75 E River Rd
Minneapolis, MN, 55455, USA

Paul R. Schrater
Cognitive Science, 75 E River Rd
Minneapolis, MN, 55455, USA

Abstract

Cognitive science is an archipelago of concepts and models, with cross-pollination between topics of interest often prohibited by incompatible approaches. Despite this, behavioral performance universally depends on information transmission between brain regions and is limited by physical and biological constraints. These constraints can be formalized as information theoretic constraints on transmission, which provide normative predictions across a surprising range of cognitive domains. To illustrate this, we describe a simple variable-length rate coding model built with Poisson processes, Bayesian inference, and an entropy-based decision threshold. This model replicates features of human task performance and provides a principled connection between a high-level normative framework and neural rate codes. We thereby integrate several disjoint ideas in cognitive science by translating plausible constraints into information theoretic terms. Such efforts to translate concepts, paradigms and models into common theoretical languages are essential for synthesizing our rich but fragmented understanding of cognitive systems.

Keywords: information theory; bayesian inference; rate coding; response time; learning

Introduction and Background

Cognitive science is home to almost as many models as phenomena they purport to describe. While this *sui generis* approach to each problem allows rich and flexible descriptions, it stands in sharp contrast to the physical sciences, in which scientists strive for and expect simple unifying principles, like Newton's axiomatic laws, from which individual phenomena arise as particular circumstantial manifestations (Chater & Brown, 2008). In cognitive science, we would say that Newton's laws are *normative*. But where are our first principles, from which we can hope to derive a coherent set of expectations about how cognition *should* operate? In this paper, we consider information transmission from the environment, through the brain, to behavior. By constraining both the channel code and each transmitted signal to be optimally inferred under normative assumptions, we can construct a message-transmission system that replicates the Hick-Hyman law (Hick, 1952; Hyman, 1953) and the Power Law of Practice (Newell & Rosenbloom, 1981), illuminates the connection between transmission rate and energy use, and produces human-like response time distributions. Our information-theoretic approach affords a principled way to connect levels of analysis (Marr, 1982) by integrating energetic resource availability, message encoding and decoding schemes, and task performance characteristics into a single framework.

Applying information-theoretic concepts to the study of cognition is not new. The years following Claude Shannon's 'A Mathematical Theory of Communication' (1948) produced a wealth of information-theoretic analyses of cognitive function, perhaps the most famous of which resulted in the Hick-Hyman law (Hick, 1952; Hyman, 1953). This mathematical approach merged with optimal control theory to become Cybernetics (Wiener, 1965), which promised to understand cognition and behavior as just another system of information transmission, feedback, and control, and subject to the same constraints. Despite their successes, enthusiasm about both information theory and cybernetics has not persisted to the present day, partly because cybernetics was abstracted away from biological and neurological characterizations, and partly because the cognitive revolution led to a focus on the nature and calculus of representation.

The development of cognitive architectures has resulted in highly successful models of a broad array of tasks (Sun, 2008; Anderson et al., 1997; McClelland, 2009). In parallel, architecture-free computational principles like Bayesian inference, prediction, credit assignment, and generalization bounds on learning have provided a rich framework for normative thinking (Shiffrin, 2010; Griffiths et al., 2008). Computational architectures form a possible hybrid (Chater & Brown, 2008), using normative computational principles to structure a cognitive architecture. However, these principles are often expressed in mathematical language disconnected from cognitive and neural architectures, leading to a pervasive difficulty in translating between mathematical formulation and plausible neural implementaiton.

The inability to translate between cognitive models directly results in a lack of knowledge transfer between domains (cognitive processes, language, tasks, etc) and levels of analysis (high-level models to low-level mechanistic details). For example, consider cognitive control as a case-in-point illustration. 'Cognitive control' refers to the deployment of attention and memory resources in the service of competing tasks. Each of these (control, attention, and working memory) are famously limited in capacity and inextricably intertwined in their roles in executive function. It is well-known that task practice lessens the effort required to do tasks, lessens attentional load, reduces response times, and decreases the amount of cognitive control required (Logan, 1985; Moors, 2016; Pierce & McDowell, 2017). These effects mirror practice

effects in perceptuo-motor skill acquisition, suggesting there should be some common principles, but it is currently difficult to transfer insights gained in the study of one phenomenon to the study of others. This lack of transferability means that cognitive science has developed a series of 'knowledge islands,' making it almost impossible to share insights across boundaries.

Even within a single topic, with the same underlying concepts, there are often many theories that are difficult to relate to each other. For example, under the shared working assumption that mental effort is treated as a cost in a cost-benefit analysis, there is considerable disagreement about the nature of the cost. Depending on the theory employed, it may represent an opportunity cost from foregone tasks, a loss of the intrinsic reward of cognitive leisure, the tendency of mental effort to discourage use of limited-capacity resources like working memory, or simply the effort of cognitive control as a cost per se. Although Shenhav et al. (2017) show that these ideas share common computational principles, they also leave it as an open question how to compare them directly. The difficulty is that cognitive costs are exogenous to the computational architecture, which means there are many non-equivalent ways to import them. Without further normative constraints, there are many rational ways to import computational modeling ideas (like costs), which means each new model multiplies the translational difficulties for integrating and relating existing models, constraints and concepts.

Like Shenhav et al. (2017), we take as foundational that the brain is an information-processing organ, ultimately transferring information from the environment via sensation, through the brain, and back into the environment as behavior. Although high level theories of behavior are most easily expressed in decision- and control-theoretic terms, re-expressing these theories in information-theoretic terms affords the incorporation of constraints on information processing, as illustrated by work on bounded rationality (Ortega et al., 2015). Biological constraints involving energy availability and noise, when translated into information theory formalism, become *normative bounds* on the ability to transfer information. Similarly, limitations on information available to the organism provide bounds on task performance. In essence, information theory provides a well-known, well-understood and sophisticated language for translating models and theories that has largely untapped potential. We illustrate this potential by demonstrating its capacity to use common computational principles to reveal relationships between the seemingly unrelated phenomena of learning rates, response time distributions, and energetic resource utilization.

Framework

Whatever the task at hand, neurons performing task-related computations must infer, in a continuous-time and streaming manner, which 'messages' are being transmitted from other brain regions (Rieke et al., 1999). This inference process is noisy, imperfect, and time-dependent. We model this process by performing continuous-time inference about the configuration of stochastic processes, with a stopping criterion based on a posterior entropy threshold. This approach produces normative predictions that match the behavioral characteristics so commonly observed in experimental paradigms, including the shape of response-time distributions and the decrease in response times and mental effort as a function of practice.

Characterizing the relationship between inferential constraints and transmission efficiency is the domain of information theory (Cover & Thomas, 2012). Information theory has been transformative in its applications to electronic communications, and has provided useful normative predictions for neural characteristics (Bialek, 2012). In particular, information theoretic constraints underlie the Efficient Coding Hypothesis (Barlow et al., 1961; Simoncelli & Olshausen, 2001), which suggests that neural connectivity is structured in such a way as to encode information from the natural environment with maximum efficiency. Despite widespread evidence for the general validity of this hypothesis in early sensory systems (e.g. Laughlin (1981); Vinje & Gallant (2000); Pitkow & Meister (2012)), there is still significant uncertainty as to whether information theoretic principles are relevant at the level of cognitive processing. Central to this reservation is a concern that Shannon's proofs of the existence of arbitrarily efficient binary codes rely on his use of 'block codes,' in which several messages are combined into a single string in a way that increases the likelihood of error-free transmission (Shannon, 1948; Cover & Thomas, 2012). For example, Luce (2003) writes "Shannon's way of defining the concept [of channel capacity] requires that not individual signals be transmitted but rather very long strings of them so as to be rid of redundancies. That is rarely possible within psychological experiments." Another recent paper raises similar concerns that Shannon's method of encoding "requires complex computation and long delays to encode and decode in ways that achieve optimality," and that it only "applies to settings of perfect signal recovery, which may not be possible or even desirable in biological settings" (Park & Pillow, 2017).

Concerns about the applicability of Shannon's proofs to information transmission in the brain confuse levels of analysis (Marr, 1982). It is true that Shannon's reliance on block-coding to achieve efficient information transmission is an implementation-level detail applicable to discrete-time codes and not to the communication of information between neurons. However, the core conceptual contribution of information theory lies not in coding techniques but in providing a method for quantifying uncertainty. More broadly, the theory serves to characterize the ways in which noise and redundancy affect the reliability, efficiency, and rate of inference. From this broader perspective, it is surely applicable to the study of cognitive function. That an understanding of these factors can lead to the design of optimal codes is important, but the specifics of code design in a discrete-time system do not invalidate the application of general principles to the study of cognition.

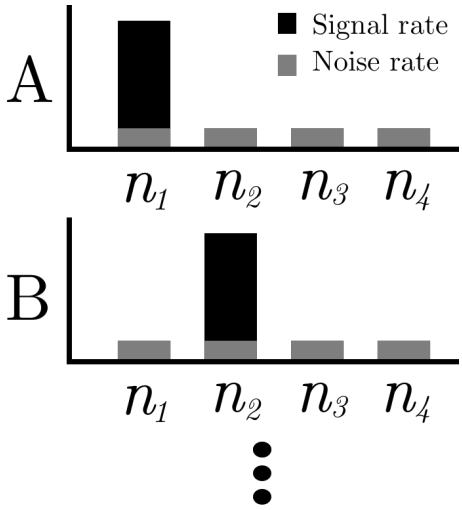

Figure 1: A codebook converts symbols A, B, etc. from a symbol alphabet into configurations of firing rates across Poisson processes $n_1, n_2, \ldots$. In this simple model, the codebook assigns a signal rate λ_S to a single Poisson process for a given symbol. Each Poisson process also emits spikes at a noise rate λ_N. As Poisson process rates are additive, this results in a total emission rate of $\lambda_N + \lambda_S$ for the 'activated' process.

In the remainder of this paper, we show an example of a continuous-time variable length coding mechanism, built using entropy and inference, that adheres to the principles of information theory while providing normative predictions of signal transmission time and accuracy. We emphasize that the continuous-time nature of the code means that signals are not discretized. Because of this, we are able to transmit messages such that transmission time is linearly related to message surprisal, replicating the Hick-Hyman law. By presenting such a code, we show that appropriate information-theoretic concepts can be applied to the study of neural information transmission.

Implementation

We model information transmission by having a sender encode a message into a configuration of Poisson process firing rates, and a receiver watch the generated spikes until they are confident about the configuration of underlying rates, and thus about the content of the encoded message (see Figure 2 for a schematic of the architecture). In more detail, the transmission mechanism consists of an encoder, a transmitter, a receiver, and a codebook. The transmitter is an array of Poisson processes, each continuously producing points or 'spikes' independently at a given noise rate λ_N. This can be viewed as a basic model of a neural rate code, as neural spikes trains are often modeled as Poisson processes (Rieke et al., 1999). The symbols to be communicated are taken from an alphabet of discrete symbols $\mathcal{A}$. The codebook describes a mapping between each symbol and a configuration of Poisson rates,

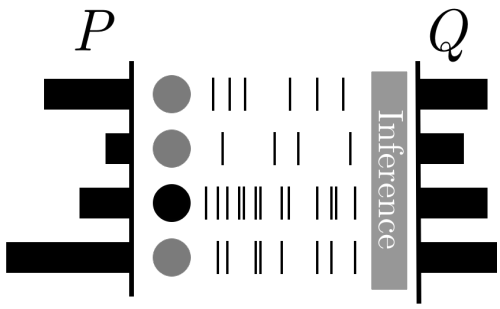

Figure 2: Messages are selected from a source distribution P. The codebook translates each message into a higher firing rate for a single process (a simplifying, but not restrictive, assumption). Poisson processes stochastically emit spikes, which are observed by the inference process. Bayesian inference combines the prior distribution Q with the likelihood of each message given the accumulated observations to produce a posterior distribution over possible messages.

and the mapping from a given symbol to rate configuration is carried out by the encoder. For the sake of expositional simplicity, we restrict the codebook to increasing the rate for a single Poisson process from the noise rate λ_N to a signal rate $\lambda_N + \lambda_S$, as shown in Figure 1. The neural analogue is that each Poisson process is 'tuned' to 'prefer' a particular symbol in a 1-hot manner, resulting in a sparse code.

The receiver observes the sequence of spikes emitting from each Poisson process and continuously attempts to infer which rate configuration is producing the spikes it observes, and thereby which symbol is being transmitted. We assume, again for simplicity and consistent with common information-theoretic analysis, that the receiver knows the values of both λ_N and λ_S. In standard binary or Gaussian channels, transmission is a discrete vector of amplitudes that takes a fixed time to transmit. Because of this, practitioners typically speak in terms of transmitting bits-per-signal, or bits-per-second (which are a constant multiple of each other). In our case, the receiver accumulates information about each transmission gradually, over time. In effect, observing for a longer period of time adds redundancy to the signal.

As observations continue, the receiver calculates and continuously updates a posterior probability distribution over possible messages, and stops decoding when the entropy of the posterior reaches a pre-specified stopping threshold. Let transmitted symbols be treated as realizations of a random variable X. The receiver begins each transmission at time $t = 0$ with an initial uncertainty $H_Q(X)$ regarding the symbol being transmitted, reflecting its prior distribution $Q(X)$ of the possible codewords. As time passes and observations $Y_t = \{y_1, \cdots, y_t\}$ are made, the receiver uses Bayesian inference to update the prior to obtain a posterior distribution $Q_t(X|Y_t)$ over messages according to Bayes rule, which yields an updated posterior entropy $H_{Q_t}(X|Y_t)$. The posterior entropy decreases non-linearly with time and reflects the de-

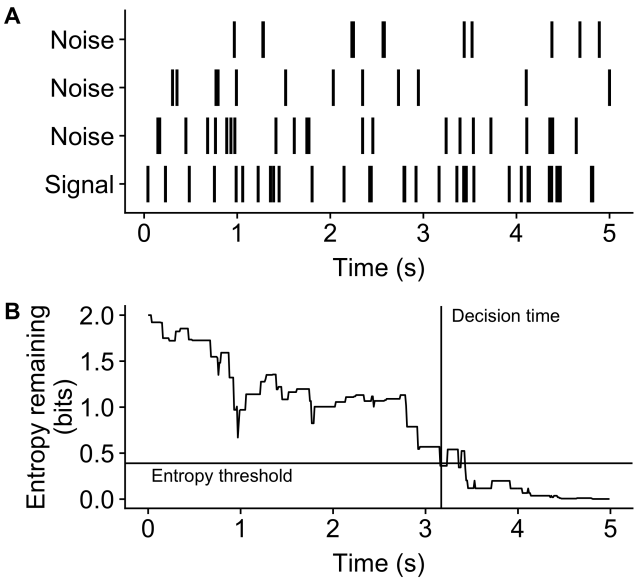

Figure 3: (A) Spikes are randomly emitted by each Poisson process as a function of time. The lower-most Poisson process is firing at a higher $\lambda_N + \lambda_S$ rate, while the others are firing at rate λ_N. (B) The receiver observes the spikes and infers which process is firing at rate $\lambda_N + \lambda_S$. The initial entropy is 2 bits, indicating a weak belief in equal probabilities for each of the 4 possible signals. The receiver's remaining entropy changes as the processes are observed and the posterior probability of each signal is calculated.

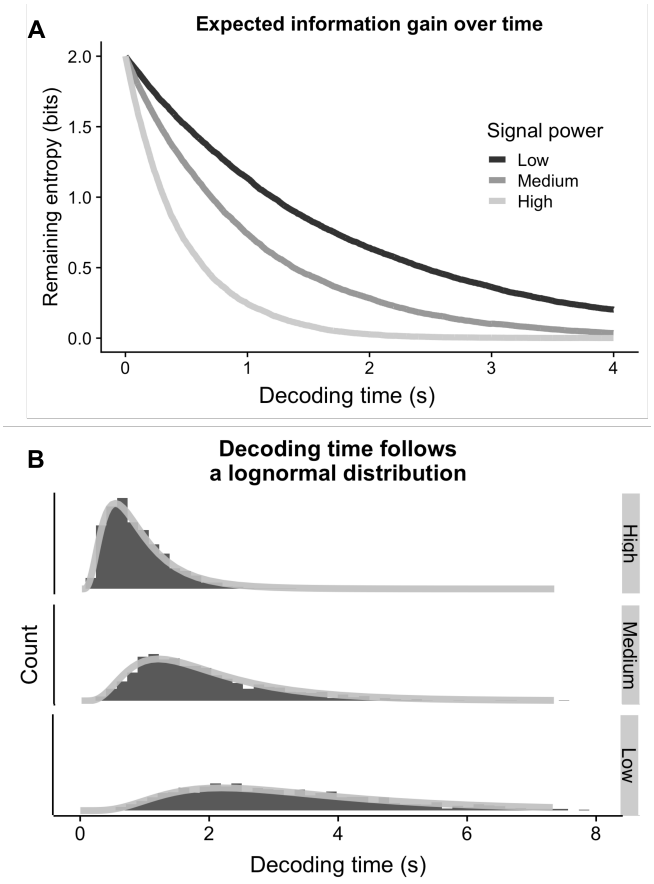

Figure 4: (A) The expected value of the receiver's entropy regarding four possible messages decreases as spikes are observed. Increasing the signal power λ_S changes the information transmission rate. (B) Response time distributions vary as a function of signal power λ_S, and in each case are well-fit by a log-normal distribution.

gree of confidence that a message has been correctly received. Transmission stops when $H_{Q_t}(X|Y_t)$ reaches a threshold. Figure 3 shows the change in posterior entropy over time for an example transmission.

Variable length transmissions

In the coding scheme introduced here, messages are *variable-length*: transmissions of messages with higher surprisal takes more time than messages with low surprisal, where surprisal is calculated using the prior probability distribution $Q(X)$ of the receiver. Recall that the surprisal $h(x)$ of a message x drawn from a distribution $P(X)$ is the logarithm of the inverse probability of the message, $h(x) = \log_2 \frac{1}{P(X=x)}$.

In 'entropy codes,' codeword length (and thus transmission time of each codeword) is roughly proportional to the surprisal of the encoded symbol in the absence of noise. When symbols are independently drawn according to a categorical probability distribution, this can manifest in two ways. In the first, increasing the number of possible symbols increases the surprisal of each individual symbol, and consequently the length of the code needed to encode its value. In the second, symbols drawn from a categorical distribution with unequal probabilities will have different surprisal values: more frequently transmitted messages will have lower surprisal and shorter codes than less frequent messages. We performed simulations to explore these scenarios in turn using our transmission model.

First, we varied codebook sizes and recorded transmission times using a fixed entropy threshold and a uniform source distribution. The nonzero entropy threshold occasionally results in transmission errors, as we see in human subjects. Information transmitted is thus less than the surprisal of each individual message, on average. We computed actual information transmitted by calculating the mutual information between transmitted symbols and received symbols, for each codebook size. The results are shown in Figure 5 and are a close qualitative match for the Hick-Hyman observations of human response times reported by Hick (1952) and Hyman (1953).

We next transmitted messages drawn from a non-uniform distribution $P(X)$ and measured transmission time for each message. For each transmission, we measured the information transmitted by comparing the receiver's prior probability distribution $Q(X)$ (which equals the source distribution

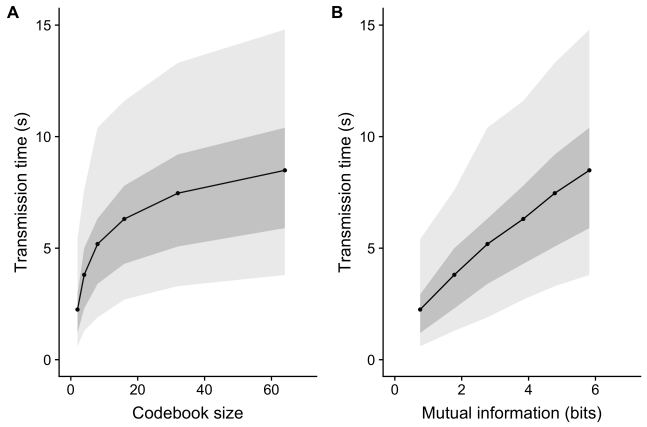

Figure 5: Mean transmission time increases logarithmically with codebook size and linearly with information transmitted, mirroring the Hick-Hyman law. Points represent mean transmission times and shaded regions represent the 50% and 90% high-density interval of the transmission time distribution. In each case, messages were transmitted according to a discrete uniform distribution $P(X)$ over messages, and the receiver maintained a uniform prior distribution $Q(X) = P(X)$ of the same dimensionality. For each transmission, an entropy threshold of 0.3 bits was used, with $\lambda_S = 4$ and $\lambda_N = 10$.

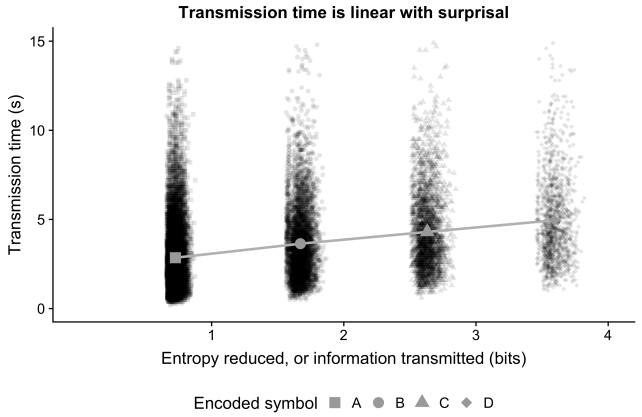

Figure 6: Mean transmission time increases linearly with actual information transmitted, echoing similar findings in humans by Hyman (1953). The quantity of information transmitted is calculated as the the KL-divergence between the prior distribution $Q(X)$ and the posterior distribution $P(X|Y)$ at decision time. Messages were drawn from a non-uniform source distribution $P(X)$. The receiver is assumed to know this source distribution and maintains a prior distribution $Q(X) = P(X)$. For each transmission, an entropy threshold of 0.3 bits was used, with $\lambda_S = 4$ and $\lambda_N = 10$.

$P(X)$, an assumption we relax below) with their posterior distribution $Q(X|Y)$ at decision time. We measured the difference in these distributions using the Kullback–Leibler divergence between the two distributions, $D_{KL}(Q(X|Y)||Q(X))$. The change between the receiver's prior and posterior distributions is equivalent to the decrease in the receiver's subjective uncertainty about which message is being transmitted. From the point of view of the receiver, this is equivalent to the amount of information transmitted, in bits. Figure 6 shows a linear relationship between message surprisal and transmission time, again qualitatively matching Hyman's reported results from human subjects.

Learning to efficiently transmit

As with source-coding systems, expected message transmission times are faster when more frequently transmitted messages are transmitted in less time than less frequently transmitted messages. In the our system, this is implemented by tailoring the receiver's prior distribution Q to match, as closely as possible, the source distribution P. This reveals an epistemic problem from the perspective of the receiver, which has no *a priori* knowledge of the source distribution: the prior must be learned and updated by observing message transmissions. The work of Hick and Hyman has been legitimately criticized for omitting this discussion (Laming, 2010).

Suppose we allow a receiver with an incorrect uniform prior message distribution Q_{init} to update its distribution to Q_{obs} in a Bayesian manner each time a message is received, so that the subsequent message transmission starts with the updated prior. As the receiver observes which messages are transmitted and at what relative frequency, Q_{obs} will become an ever-closer approximation to P, shrinking both $D_{KL}(P||Q_{obs})$ and the expected transmission times. Figure 7 shows message transmission times resulting from a uniform (naive) prior, a prior equal to the true source distribution, and an intermediate distribution, as might be expected to develop from a moderate level of experience with the task. In each case, response time is linearly related to message surprisal as calculated using Q. The slope depends on the amount of experience with the task: as experience accrues and Q_{obs} approaches P, response times more closely reflect the transmission frequencies of each message. The varying slopes are reminiscent of the subject-specific slope found by Hyman (1953).

As observations accumulate, the rate at which response times decrease as Q approaches P mirrors the Power Law of Learning (Newell & Rosenbloom, 1981). The Power Law of Learning is a ubiquitous finding that task response times have a power-law relationship with the number of practice episodes, when averaged across many subjects. We constructed a categorical source distribution P with $k = 16$ categories, but with most of the probability mass in two categories. We initialized Q_{init} to have a Dirichlet prior with concentration parameters 2, representing a weak prior belief that the source distribution is uniform. We simulated N message transmissions, for $N = 2$ to $N = 1024$, taken evenly in log space. For each value of N, we averaged the results across 1,000 simulated observers, resulting in an expected posterior

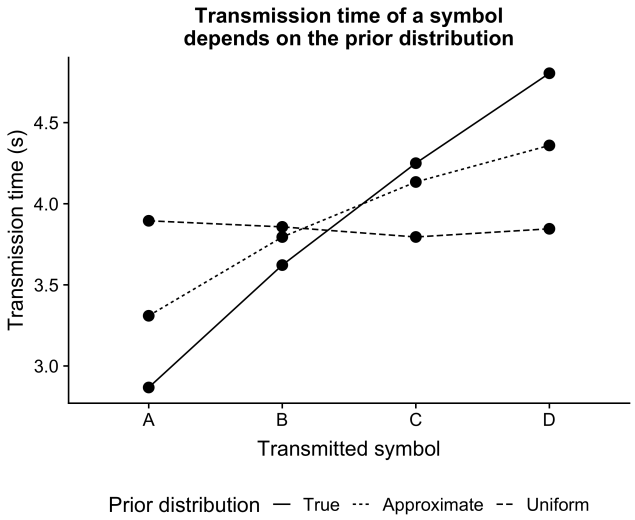

Figure 7: Mean transmission time is a function of the receiver's prior belief $Q(X)$ over messages, rather than the source distribution $P(X)$. In each case, messages were transmitted from the identical source distribution, where A was most frequent, followed by B, and so on. Each line connects response times arising from the same prior distribution. A uniform $Q(X)$ results in a flat line, while a $Q(X) = P(X)$ results in the steepest slope. In each case, the relationship between subjective surprisal and response time is approximately linear. For each transmission, an entropy threshold of 0.3 bits was used, with $\lambda_S = 4$ and $\lambda_N = 10$.

distribution Q_{obs} after N observations. For each Q_{obs} we then simulated more 2,000 message transmissions, with messages drawn with frequency defined by P, and calculated the transmission time for each. As illustrated in Figure 8, the relationship between observations N and transmission time is linear in log-log space, matching the Power Law of Learning.

The energy connection

Implicit in the above discussion is the notion that information transmission costs energy: transmission is initiated when an encoder assigns signal power λ_S to a Poisson process. If each spike costs energy, this implies a rate of energy expenditure. As shown in Figure 4, signal power has a direct effect on the rate of entropy decrease and the resulting transmission times. The framework introduced here allows us to explicitly describe the relationship between energy use (in terms of spikes), task novelty (in the form of naive Q estimates), task practice, and response times. If mental effort is a phenomenological correlate of signal transmission costs, it also provides a normative explanation for effort decrease as a function of practice, and provides weight to the currently tenuous relationship between mental effort and the utilization of metabolic resources.

Indeed, neural spikes are not free: an estimated 10% of an adult body's energy budget is allocated to neural information

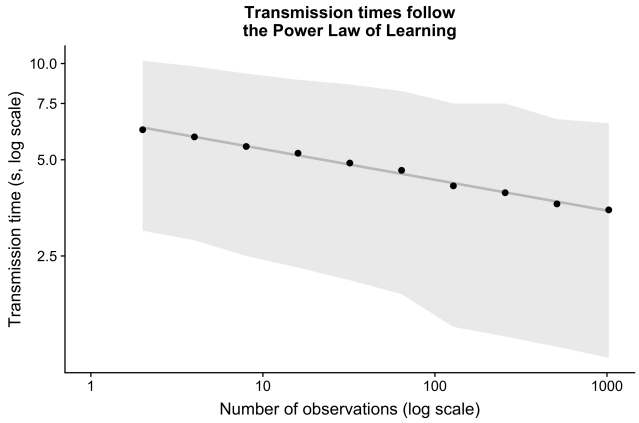

Figure 8: Simulated message transmission time decreases as a function of observations, as the prior Q approaches the source distribution P. Signals are transmitted with signal strength $\lambda_S = 4$, noise power $\lambda_N = 10$, and an entropy threshold of 0.3. Points represent mean transmission times, and the shaded region represents the 80% high-density interval of the response time distributions.

processing (Stone, 2018). In light of this, we might expect the brain to adopt a strategy of driving energetic efficiency by tailoring codes (represented by codebooks and Q distributions) to individual tasks. As stimulus distributions P are not equivalent between tasks, this would necessitate the creation and maintenance of a bank of task-specific codes, with a power-law response time trend repeated during the practice of each separate task (Newell & Rosenbloom, 1981). However, the power-law describes severely diminishing returns between task practice and transmission efficiency, and tasks in the world are not as discrete as in laboratory experiments. Because of this, in a naturalistic setting we might instead expect the brain to implement some 'universal' code (Cover & Thomas, 2012) that provides moderately efficient transmission across range of tasks (Vera et al., 2018). If this is the case, the brain would sacrifice efficiency to achieve flexibility, which is, after all, a chief characteristic of human cognition.

Conclusion

We have applied the principles of information theory to a simple rate-coding model of neural information transmission. We showed that placing normative bounds on the inference of both source distributions and the content of individual signals results in a coding mechanism that predicts the Hick-Hyman Law and the Power Law of Practice, describes a principled connection between information transmission and energy use, and produces realistic response-time distributions. By utilizing the information-theoretic principles relevant to a continuous-time system (in particular entropy and inference), and avoiding those that are not (block-coding), we have produced a simple and parsimonious explanation of a wide range of phenomena.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). Act-r: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439–462.

Barlow, H. B., et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*, 217–234.

Bialek, W. (2012). *Biophysics: searching for principles*. Princeton University Press.

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, *32*(1), 36-67.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11-26. doi: 10.1080/17470215208416600

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of experimental psychology*, *45*(3), 188.

Laming, D. (2010). Statistical information and uncertainty: A critique of applications in experimental psychology. *Entropy*, *12*(4), 72.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, *36*(9-10), 910–912.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*(2), 367.

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of general psychology*, *7*(2), 183.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information. w. h.* WH San Francisco: Freeman and Company.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263–287.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, *1*(1981), 1–55.

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E., & Tishby, N. (2015). Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*.

Park, I. M., & Pillow, J. W. (2017). Bayesian efficient coding. *bioRxiv*. doi: 10.1101/178418

Pierce, J. E., & McDowell, J. E. (2017). Reduced cognitive control demands after practice of saccade tasks in a trial type probability manipulation. *Journal of cognitive neuroscience*, *29*(2), 368–381.

Pitkow, X., & Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, *15*(4), 628.

Rieke, F., Warland, D., Steveninck, R. d. R. v., & Bialek, W. (1999). *Spikes: Exploring the neural code*. A Bradford Book.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*, 99–124.

Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in cognitive science*, *2*(4), 736–750.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193–1216.

Stone, J. (2018). *Principles of neural information theory: Computational neuroscience and metabolic efficiency*. Sebtel Press.

Sun, R. (2008). Introduction to computational cognitive modeling. *Cambridge handbook of computational psychology*, 3–19.

Vera, M., Vega, L. R., & Piantanida, P. (2018). Compression-based regularization with an application to multitask learning. *IEEE Journal of Selected Topics in Signal Processing*, *12*(5), 1063–1076.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273–1276.

Wiener, N. (1965). *Cybernetics, second edition: Or the control and communication in the animal and the machine*. The MIT Press.