Generating Functions in Neural Learning of Sequential Structures

Yanlong Sun (ysun@tamhsc.edu)

Hongbin Wang (hwang@tamhsc.edu)
Center for Biomedical Informatics, Texas A&M University Health Science Center
Houston, TX 77030 USA

Abstract

A cornerstone of human statistical learning is the ability to extract abstract regularities from sequential events. Here we present a unique method to derive the generating functions for the waiting time of sequential patterns, then compare these functions with the neural mechanisms for learning sequential structures. We show that the way the neocortex integrates information over time bears a striking resemblance to the way these normative functions operate. They both operate by organizing combinatorial objects into meaningful groups then compressing the representations by discarding irrelevant information. As a result, discrete-time signals are converted into frequency signals, and similarity-based structures are converted into abstract relational structures. Our analyses not only reveal surprisingly rich statistical structures embedded in the seemingly random sequences, but also offer an explanation for how higher-order cognitive biases may have emerged as a consequence of temporal integration.

Keywords: generating function; waiting time; statistical learning; temporal integration; compressed representation.

Introduction

The human mind has a unique capacity to find order in chaos (Gazzaniga, 2008). From betting cards in casinos to investing money in stocks, people constantly attempt to extract regularities from the seemingly random sequences. For theories dealing with human statistical learning, there are always two types of challenges: How is the implicit learning without instruction is connected with the explicitly structured rule learning? How can heuristics and biases deviate systematically from normative rules?

Consider the following situation. A fair coin is flipped repeatedly in independent Bernoulli trials. Which of the two patterns, two heads in a row (HH), or a head followed by a tail (HT), is more likely to happen? To measure the *frequency* of a pattern, let $E[T]$ denote the pattern's *mean time*, which is the expected number of coin flips between any two consecutive occurrences of the pattern,

$$E[T_{\mathtt{HH}}] = E[T_{\mathtt{HT}}] = (1/2)^{-2} = 4,$$

which means that on average, HH and HT are equally likely, each occurring once in every 4 flips.

This answer may sound simple. However, it appears to be at odds with a gambler's intuition. In a game of roulette at the Monte Carlo casino in 1913, black repeated a record 26 times, people began extreme betting on red after about 15 repetitions (Huff, 1959). The gambler's fallacy, which is often attributed to the representativeness bias, reflects the belief that chance is a self-correcting process such that it is more likely to produce alternating patterns than repeating ones (Tversky & Kahneman, 1974).

Now we take a different measure. Let $E[T^*]$ denote the pattern's *waiting time*, which is the expected number of flips since the beginning of the process until the *first* occurrence of the pattern, then,

$$E[T^*_{\mathtt{HH}}] = (1/2)^{-1} + (1/2)^{-2} = 6,$$
$$E[T^*_{\mathtt{HT}}] = (1/2)^{-1} + (1/2)^{-1} = 4.$$

That is, it actually takes longer to see the first HH (a repetition) than the first HT (an alternation).

The example of coin flipping demonstrates some intricate relations between our intuition about random sequences and the normative predictions of probability theory. In terms of statistical learning, implicit learning without instruction can be rapid and robust, but it does not always agree with explicit and structured rule learning (Aslin & Newport, 2012). Aiming at possible reconciliations, many theories propose that the evaluation of the biases in human randomness perception should consider other factors beyond a single normative measure, for example, the difficulty of encoding complexity (Falk & Konold, 1997), the limited short-term memory capacity (Hahn & Warren, 2009; Kareev, 2000), and inferences with competing generating processes (Nickerson, 2002; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Based on the waiting time statistics, we have argued that the alternation bias in the gambler's fallacy can be understood as a consequence of time in that repeating patterns are "delayed" than alternating patterns (Sun, Tweney, & Wang, 2010; Sun & Wang, 2010a, 2010b, 2012). However, these theories have been limited at the behavioral level where one could only align the overall human behavior with a certain normative measure in its abstract form. It remains to be answered where the alternation bias has originated, and more critically, how the abstract representations in the human mind have taken shape from the beginning.

In the present paper, we present a unique method to derive the generating functions for the waiting time statistics of sequential patterns. This method was first introduced by Graham, Knuth, and Patashnik (1994). Here we extend this method and elaborate on the procedures where combinatorial objects are perceptually organized then compressed into abstract and closed-form representations. We then discuss a neural network model that can actually capture the waiting time statistics with unsupervised learning (Sun et al., 2015). By comparing the generating functions with the neural learning mechanisms, we offer an explanation for how human randomness perception can take shape through mere exposure to the input stimuli without instruction, and how object representation can lead to probability induction.

Generating Functions for Waiting Times

Following the notations by Graham et al. (1994), a *generating function*, $A(z)$, is the sum of a power series that "organizes" an infinite sequence $\langle a_0, a_1, a_2, \ldots \rangle$ with an auxiliary variable z,

$$A(z) = a_0 + a_1 z + a_2 z^2 + \cdots = \sum_{k \geqslant 0} a_k z^k, \tag{1}$$

and, a *probability generating function*, $G_X(z)$, where X is a random variable that takes only nonnegative integer values, is the sum of the probability distribution,

$$G_X(z) = \sum_{k \geqslant 0} \Pr(X = k) z^k. \tag{2}$$

The coefficients of $G_X(z)$ sum to 1, which can be written as $G_X(1) = \sum_{k \geqslant 0} \Pr(X = k) = 1$. The function $G_X(z)$ contains the information of all cumulants in the distribution of X. For example, the first and the second cumulants, namely, the mean and variance, are given by

$$\begin{aligned} E[X] &= G'_X(1), \\ \mathrm{Var}(X) &= G''_X(1) + G'_X(1) - G'_X(1)^2. \end{aligned} \tag{3}$$

In the following, we use these definitions to derive the waiting time for patterns in binary sequences. We first introduce the solution by Graham et al. (1994) to the pattern HH's waiting time in independent Bernoulli trials. Then, we extend the method to the pattern HT and the waiting time in first-order dependent Markov trials.

Waiting Time in Bernoulli Trials

Assuming that a coin, with probability of heads, p, and probability of tails, $q = 1 - p$, is flipped repeatedly in independent Bernoulli trials. In waiting for the pattern HH, we consider the probability space consisting of all sequences that end with the first occurrence of HH:

$$\Omega = \{\mathtt{HH}, \mathtt{THH}, \mathtt{TTHH}, \mathtt{HTHH}, \mathtt{TTTHH}, \mathtt{THTHH}, \cdots\}.$$

Letting $S_{\mathtt{HH}}$ be the generating function that sums up all members of Ω:

$$S_{\mathtt{HH}} = \mathtt{HH} + \mathtt{THH} + \mathtt{TTHH} + \mathtt{HTHH} + \mathtt{TTTHH} + \mathtt{THTHH} + \cdots,$$

and by the expansion of the power series,

$$1 + z + z^2 + z^3 + \cdots = \sum_{n \geqslant 0} z^n = \frac{1}{1 - z},$$

we can write $S_{\mathtt{HH}}$ in a "closed-form":

$$S_{\mathtt{HH}} = \sum_{n \geqslant 0} (\mathtt{T} + \mathtt{HT})^n \mathtt{HH} = \frac{\mathtt{HH}}{1 - (\mathtt{T} + \mathtt{HT})}. \tag{4}$$

We can then obtain the probability generating function for the waiting time of HH by replacing each H with pz and each T with qz:

$$G_{\mathtt{HH}}(z) = \frac{p^2 z^2}{1 - qz - pqz^2}. \tag{5}$$

Letting $z = 1$, we have $G_{\mathtt{HH}}(1) = 1$, which means that the pattern HH eventually will happen with probability 1.

Then, from Equation 3, the pattern HH's waiting time is:

$$E[T^*_{\mathtt{HH}}] = \frac{1}{p} + \frac{1}{p^2}. \tag{6}$$

For example, at $p = 1/2$, we have $E[T^*_{\mathtt{HH}}] = 6$. (Hereafter we omit the calculation of the variance.)

Without losing any mathematical rigor, this method of deriving generating functions is remarkably simple. To recapture the critical steps, first of all, the generating function $S_{\mathtt{HH}}$ in Equation 4 partitions the probability space Ω with a "juxtaposition" (i.e., multiplication) of two terms: the binomial term $(\mathtt{T} + \mathtt{HT})^n$ and the pattern HH itself. The binomial term organizes all possible sequences where the pattern HH has *failed* to occur (such that the waiting has to start all over), into the power groups by the number of failures, n. For example, the sequence TTHT belongs to the group $(\mathtt{T} + \mathtt{HT})^3$, since it is obtained by stacking either T or HT 3 times. Then, the right-hand side of Equation 4 is simply the sum of a power series. Next, in Equation 5, the z-transformation from $S_{\mathtt{HH}}$ to $G_{\mathtt{HH}}(z)$ compresses the representation further by discarding the exact order of H's and T's. As a result, $G_{\mathtt{HH}}(z)$ only preserves the exact number of flips in each sequence, which is indexed by the power of z. Finally, averaging all sequence lengths with $G'_{\mathtt{HH}}(z = 1)$, which effectively removes the index z, we have the waiting time for the pattern HH.

What is even more remarkable about this method is that it may also shed light on how human randomness perception might have taken shape in a similar fashion. Particularly, this method directly operates on *object representations*. It illustrates how combinatorial objects, namely, sequences unfolding over time, can be organized into meaningful groups then compressed into an abstract and closed-form representation. We will elaborate further on this point in the next section.

Partitioning by Auxiliary Sum

The example above shows one way to partition the probability space. It should be noted that the way to organize the combinatorial objects can be rather flexible. One simple method is to use an *auxiliary sum*, which is the sum of all sequences that do not contain any occurrences of the expected pattern. In the following, we use this method to derive the generating functions for the pattern HT's waiting time in Bernoulli trials.

In waiting for the first occurrence of HT, we consider two sets of sequences: $S_{\mathtt{HT}}$ represents the sum of all sequences that end with the first HT, and M represents the auxiliary sum of all sequences that do not contain any HT.[1] We can then write two linear equations:

$$\begin{aligned} S_{\mathtt{HT}} + M &= M(\mathtt{H} + \mathtt{T}) + \mathtt{H} + \mathtt{T}, \\ S_{\mathtt{HT}} &= M\mathtt{HT} + \mathtt{HT}, \end{aligned}$$

[1] Note that different from the method by (Graham et al., 1994), our auxiliary sum here does not include the empty sequence. This is to emphasize the idea that all members in the sum are directly observable.

where the first equation partitions the entire probability space into either $S_{\mathtt{HT}}$ or M, and the second equation states that any member of $S_{\mathtt{HT}}$ is obtained either by extending a member of M with HT at the end or directly from the first two coin flips.

Solving for $S_{\mathtt{HT}}$, we have

$$S_{\mathtt{HT}} = \frac{\mathtt{HT}}{(\mathtt{H}-1)(\mathtt{T}-1)}.$$

Replace each H with pz and each T with qz, we have the probability generating function for the pattern HT's waiting time in Bernoulli trials:

$$G_{\mathtt{HT}}(z) = \frac{pqz^2}{(pz-1)(qz-1)}.$$

Then, from Equation 3, we have

$$E[T^*_{\mathtt{HT}}] = \frac{1}{p} + \frac{1}{1-p}, \tag{7}$$

For example, at $p = 1/2$, we have $E[T^*_{\mathtt{HT}}] = 4$, which is 2 flips short of the HH's waiting time (cf., Equation 6).

First-order Dependent Markov Trials

In studies on human randomness perception, another widely used model is the first-order dependent Markov trials, parameterized by the *probability of alternation* between consecutive trials (e.g., Budescu, 1987; Falk & Konold, 1997; Lopes & Oden, 1987; Nickerson, 2002; Oskarsson, Van Boven, McClelland, & Hastie, 2009; Sun & Wang, 2012). In the following, we derive the generating functions for both patterns HH and HT in such a process.

We first assume that the process is H-T symmetrical (i.e., exchangeable) with stationary probabilities,

$$\pi_{\mathtt{H}} = \pi_{\mathtt{T}} = 1/2,$$

which means that in the long run, heads and tails are equally likely. Then, we use the probability of alternation, p_A, to simplify the transition probabilities,

$$p_A = p_{\mathtt{H},\mathtt{T}} = p_{\mathtt{T},\mathtt{H}} = 1 - p_{\mathtt{H},\mathtt{H}} = 1 - p_{\mathtt{T},\mathtt{T}}.$$

In waiting for the pattern HH, we first consider the sum $S_{\mathtt{HH}}$ for all sequences that end with the first occurrence of HH. We then split the auxiliary sequences that do not contain the pattern into two parts, $M_{\mathtt{H}}$ for all sequences that end with H and $M_{\mathtt{T}}$ for all sequences that end with T. This partitioning is plotted as a Markov chain in Figure 1A, which shows that extending any member of $M_{\mathtt{H}}$ with a repetition (R) produces a member of $S_{\mathtt{HH}}$, extending any member of $M_{\mathtt{H}}$ with an alternation (A) produces a member of $M_{\mathtt{T}}$, and so on.

According to Figure 1A, we can write three equations,

$$\begin{aligned} M_{\mathtt{H}} &= \mathtt{H} + M_{\mathtt{T}}\mathtt{A}, \\ M_{\mathtt{T}} &= \mathtt{T} + M_{\mathtt{T}}\mathtt{R} + M_{\mathtt{H}}\mathtt{A}, \\ S_{\mathtt{HH}} &= M_{\mathtt{H}}\mathtt{R}. \end{aligned}$$

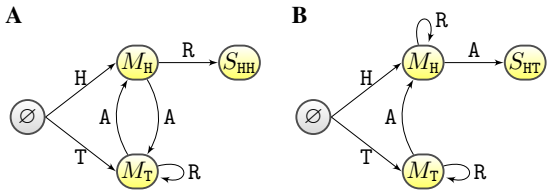

Figure 1: Markov chains for generating the first occurrence of the patterns HH (figure **A**) and HT (figure **B**). States $S_{\mathtt{HH}}$ and $S_{\mathtt{HT}}$ represent all sequences that end with the first occurrence of the expected pattern. States $M_{\mathtt{H}}$ and $M_{\mathtt{T}}$ represent all sequences that end with either an H or a T but do not contain the expected pattern. After the first transition out of the initial empty state (∅), later transitions are characterized by either a repetition (R) or an alternation (A).

Solving for $S_{\mathtt{HH}}$, we have the generating function

$$S_{\mathtt{HH}} = \mathtt{HR} + \frac{\mathtt{T}+\mathtt{HA}}{1-\mathtt{R}-\mathtt{AA}}\mathtt{AR}.$$

Replacing each H and each T with $z/2$ (since $\pi_{\mathtt{H}} = \pi_{\mathtt{T}} = 1/2$), each R with $(1-p_A)z$, and each A with $p_A z$, we have the probability generating function for the pattern HH's waiting time,

$$G_{\mathtt{HH}}(z) = \frac{(p_A-1)(2p_Az-z+1)z^2}{2(p_A^2z^2-p_Az+z-1)}.$$

Therefore, from Equation 3,

$$E[T^*_{\mathtt{HH}}] = 1 + \frac{1}{2p_A} + \frac{2}{1-p_A}. \tag{8}$$

For example, when $p_A = 1/2$, we have $E[T^*_{\mathtt{HH}}] = 6$, which is the same result from Equation 6. When $p_A = 1/3$, we have $E[T^*_{\mathtt{HH}}] = 5.5$.

Similarly, according to Figure 1B, the waiting time for the pattern HT can be solved from the following equations:

$$\begin{aligned} M_{\mathtt{H}} &= \mathtt{H} + M_{\mathtt{H}}\mathtt{R} + M_{\mathtt{T}}\mathtt{A}, \\ M_{\mathtt{T}} &= \mathtt{T} + M_{\mathtt{T}}\mathtt{R}, \\ S_{\mathtt{HT}} &= M_{\mathtt{H}}\mathtt{A}, \end{aligned}$$

resulting in the generating function,

$$S_{\mathtt{HT}} = \frac{\mathtt{HA}-\mathtt{HRA}+\mathtt{TAA}}{(1-\mathtt{R})^2},$$

and the probability generating function,

$$G_{\mathtt{HT}}(z) = \frac{p_A(2p_Az-z+1)z^2}{2(p_Az-z+1)^2}.$$

Therefore,

$$E[T^*_{\mathtt{HT}}] = 1 + \frac{1}{2p_A} + \frac{1}{p_A}. \tag{9}$$

For example, when $p_A = 1/2$, we have $E[T^*_{\mathtt{HT}}] = 4$, which is the same result from Equation 7. When $p_A = 1/3$, we have $E[T^*_{\mathtt{HT}}] = E[T^*_{\mathtt{HH}}] = 5.5$. That is, alternations have to be this much less frequent than repetitions to make the patterns HH and HT have the same waiting time.

Asymmetry in the Additional Time

The Markov chains in Figure 1 depict *a structural asymmetry* in the trajectories of different patterns. This asymmetry can be more obvious if we only look at a portion of the waiting time. Whereas the waiting time $E[T^*]$ is always counted from the beginning of the process (i.e., the initial state $\varnothing$ is empty), we can define the *additional time*, denoted by $E[T_{j|i}]$, as the expected number of transitions from any initial state i until the first arrival of the state j. For example, $E[T^*_{\mathtt{HH}}]$ and $E[T^*_{\mathtt{HT}}]$ in Equations 8 and 9 share the same component $E[T^*_{\mathtt{H}}] = E[T_{\mathtt{H}|\varnothing}] = 1 + {}^1\!/_{2p_A}$. Cancelling the common terms, we have

$$E[T_{\mathtt{HH}|\mathtt{H}}] = \frac{2}{1-p_A}, \quad E[T_{\mathtt{HT}|\mathtt{H}}] = \frac{1}{p_A}. \tag{10}$$

The difference between $E[T_{\mathtt{HH}|\mathtt{H}}]$ and $E[T_{\mathtt{HT}|\mathtt{H}}]$ is illustrated in Figure 1. Before reaching $S_{\mathtt{HH}}$, an alternation after state $M_{\mathtt{H}}$ leads the process to state $M_{\mathtt{T}}$ thus "delays" the transition to the destination. In contrast, before reaching $S_{\mathtt{HT}}$, a repetition after state $M_{\mathtt{H}}$ makes the process stay in the same state thus the distance to the destination is unchanged. Together, when repetitions and alternations are equally likely, $p_A = {}^1\!/_2$, the temporal distance from $M_{\mathtt{H}}$ to $S_{\mathtt{HH}}$ is greater than that from $M_{\mathtt{H}}$ to $S_{\mathtt{HT}}$: $E[T_{\mathtt{HH}|\mathtt{H}}] = 4$, and $E[T_{\mathtt{HT}|\mathtt{H}}] = 2$.

Similarly for independent Bernoulli trials, canceling the common terms in Equations 6 and 7, we have

$$E[T_{\mathtt{HH}|\mathtt{H}}] = \frac{1}{p^2}, \quad E[T_{\mathtt{HT}|\mathtt{H}}] = \frac{1}{1-p}.$$

By extending Figure 1 to longer sequences, we can show that the difference increases *exponentially* as the pattern length increases. When $p = {}^1\!/_2$, given an existing streak of k heads, despite that the next flip can be equally likely a head or a tail, the additional time until a streak of $(k+1)$ heads is much longer than that for the pattern of k heads followed by a tail:

$$E[T_{(k+1)\mathtt{H}|k\mathtt{H}}] = \frac{1}{p^{k+1}}, \quad E[T_{k\mathtt{HT}|k\mathtt{H}}] = \frac{1}{1-p}.$$

Neural Learning of Sequential Structures

The generating functions reveal a great deal about the rich statistical structures embedded in random sequences. A question that immediately follows then is whether these structures can be implicitly captured by the human mind. We have argued before that at the behavioral level, people's preference for alternating patterns (e.g., HT) over repeating ones (e.g., HH) appears to be driven by the patterns' waiting time statistics (Sun & Wang, 2010a, 2010b, 2012). Considering the way the generating functions for waiting times are derived, here we argue that the alternation bias might have actually emerged at the neural level.

In particular, the way these generating functions operate is to organize combinatorial objects into smaller groups then compress the representation into a closed form where irrelevant information is discarded. By doing so, discrete-time signals (e.g., sequences encountered over time) are transformed into frequency signals (e.g., as a power series), and similarity-based structures (e.g., sequences that end with the same elements) are transformed into abstract relational structures (e.g., given an H, the first HT arrives *earlier than* the first HH). Such transformations bear a striking resemblance to the currently proposed learning mechanisms in the human brain, for example, perceptual processing (Marr, 1982), temporal integration (Elman, 1990; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; O'Reilly, Wyatte, & Rohrlich, 2014), neural population encoding (Pouget, Beck, Ma, & Latham, 2013), and Bayesian abstraction (Tenenbaum et al., 2011). Then, it would be a plausible conjecture that processes similar to these generating functions may also take place in the human brain. That is, by merely observing random sequences unfolding over time, the mind should be able to naturally capture abstract structures summarized by the waiting time statistics.

Indeed, we have recently reported a biologically-motivated neural model that did just that (Sun et al., 2015). In the light of the generating functions developed above, here we recapture some of the major findings from the model.

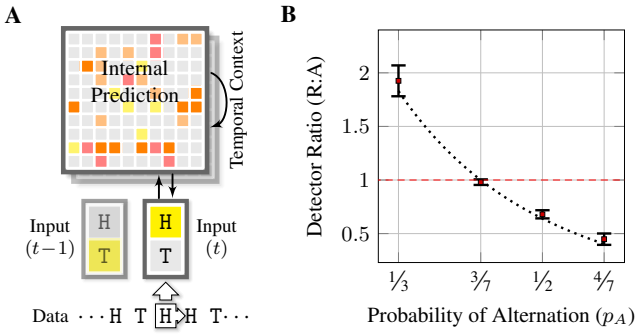

Figure 2: A neural network model of temporal integration (figures adopted from Sun et al., 2015). **A**. Architecture of the neural model. A single input layer scans a sequence of binary digits one digit at a time (input at time $t-1$ is for illustration only). An internal prediction layer, with its temporal context representation, attempts to predict the next input. **B**. Neural model behavior depicted by the ratio between the numbers of repetition and alternation detectors in response to the actual probability of alternation (p_A) in the input sequence. Error bars ($\pm$SEM) represent the variability of model predictions. The dotted line is the squared total time ratio between alternation and repetition patterns (Equation 11).

A Neural Model of Temporal Integration

Our neural model is extremely simple (Figure 2A). It employs a recently-developed neural algorithm for temporal integration (O'Reilly et al., 2014). At the sensory level, a 2-unit input layer scans non-overlapping signals of heads (H) versus tails (T) one digit at a time from sequences generated by the first-order dependent Markov trials. Then, a 100-unit internal prediction layer attempts to predict the next input, with the

benefit of a prior temporal context representation. The bidirectional activation dynamics between the input layer and the internal prediction layer allow us to use a single input layer for both providing inputs and receiving predictions.

By unsupervised learning, the model was trained with binary sequences generated at various levels of probability of alternation (p_A), each sequence consisting of 10,000 trials. After training, the model was tested with a sequence of 1,000 trials generated at the same p_A level. Through an activation-based receptive field analysis, we decoded the representations on the internal prediction layer and classified its units as either repetition detectors (sensitive to either HH or TT) or alternation detectors (sensitive to either HT or TH). We then counted the numbers of detectors and used the ratio (R/A, repetition over alternation) to measure the model's performance (Figure 2B).

We found that the model's behavior can be mostly replicated by a simple equation that averages the effects of the mean time and waiting time statistics (the dotted line in Figure 2B):

$$\frac{R}{A} \approx \left(\frac{E[T_{\mathtt{A}}] + E[T^*_{\mathtt{A}}]}{E[T_{\mathtt{R}}] + E[T^*_{\mathtt{R}}]} \right)^2, \tag{11}$$

where $E[T]$ is the mean time, $E[T^*]$ is the waiting time, and subscripts R and A represent repetition (either HH or TT) and alternation (either HT or TH), respectively. For example, at $p_A = 3/7$, $E[T_{\mathtt{HH}}] + E[T^*_{\mathtt{HH}}] = E[T_{\mathtt{HT}}] + E[T^*_{\mathtt{HT}}]$, the model showed about the same numbers of repetition and alternation detectors, $R/A \approx 1$.

Most interestingly, at $p_A = 1/2$ (i.e., flipping a fair coin independently), despite the same training frequency of the patterns (e.g., $E[T_{\mathtt{HH}}] = E[T_{\mathtt{HT}}]$ but $E[T^*_{\mathtt{HH}}] > E[T^*_{\mathtt{HT}}]$), the model consistently produced fewer repetition detectors than alternation detectors at a ratio of $R/A \approx .70$. We then used this R/A ratio to compute the *subjective probability of alternation*, p'_A, as the model's internal representation of its actually experienced p_A,

$$p'_A = \frac{A}{R+A} = \frac{1}{1+R/A} \approx 0.59.$$

This p'_A value was consistent with the value from empirical findings. From a comprehensive review of previous studies (Falk & Konold, 1997), a unanimous finding was that people perceived or generated random sequences with a p'_A value around $0.58 \sim 0.63$.

Generating Functions in the Human Brain

It should be noted that our neural model was not specifically tasked to capture the waiting time statistics. Rather, it was built on the well-established sensitivity in the neural learning of sequential structures (Elman, 1990) and implemented with biologically realistic algorithms that aim to explain the neural basis of cognition in a wide range of different domains (O'Reilly et al., 2012, 2014). Nevertheless, given that the neural model's behavior was systematically biased by sequential patterns' waiting time, here we offer an interpretation through the lens of the generating functions.

First of all, we argue that the alternation bias in human randomness perception is the consequence of *temporally distributed learning*. While neurons in the neocortex integrates information over time through different contributions of the deep versus superficial layers (e.g., layers 5*b* and 6, see, O'Reilly et al., 2014), they act in the same way as the generating functions by transforming discrete-time signals into representations of frequency. For example, Figure 1 and Equation 10 show that at $p_A = 1/2$, the additional time travelling from $M_{\mathtt{H}}$ to $S_{\mathtt{HT}}$ is shorter than that from $M_{\mathtt{H}}$ to $S_{\mathtt{HH}}$. This means that the neurons monitoring the $(\mathtt{H} \rightarrow \mathtt{HT})$ transitions are more likely to sustain their activations over time than those monitoring the $(\mathtt{H} \rightarrow \mathtt{HH})$ transitions. By the principles of self-organizing learning (Hebb, 1949), more neurons would be tuned to temporally associating an existing H with a future HT instead of a future HH. In a certain sense, in the process of maximizing the temporal correlation, these neurons have incidentally committed themselves to the gambler's fallacy.

Second, as we have seen in the generating functions, a critical step towards a compressed representation is to discard irrelevant information (e.g., in the z-transformation). Then, the structures based on perceptual similarity are transformed into the abstract and relational structures. In the same way, for neurons to maximize the correlation between temporally adjacent events, information such as the exact order of the past events has to be discarded. There are many reasons to believe that a primary function of the cortical processing is to actively discard massive amounts of information so that only the most relevant signals are retained and processed further (for a review, see, O'Reilly et al., 2014). For neurons that can only monitor sequential events unfolding over time, what is relevant in the past is determined by what happens in the future (i.e., following the arrow of time, see Figure 1).

Third, the generating functions we derived above are not meant to predict random sequences. Rather, they are built to capture the statistical structures embedded in time. Likewise, our neural model would generally fail to predict each coin flip. The predictions made by the model are rather *implicit* than *explicit*. This feature relieves the network from the burden of predicting every last detail of the input, and merely requires that the internal network state learn to be compatible with the new inputs. As a result, learning is distributed across populations of neurons and spanned over time, therefore allows the network to be more adaptive to the statistical structures of the learning environment.

Lastly, the generating functions can operate from both directions as either the summation of discrete-time objects or the expansion of a closed form (e.g., Equation 4). Mapped onto the bidirectional activation dynamics in our neural model, this corresponds to the integration of sensory inputs (bottom-up) and the prediction from more abstract internal representations (top-down). Different from a standard simple recurrent network that implements separate input and output layers (Elman, 1990), the bidirectional activation dynamics in our model allow flexible encoding and inferring of relational structures,

thus provide a more natural mechanism for predictive learning in the brain (George & Hawkins, 2009).

Conclusion

The generating functions presented in this paper provide a normative measure for the sequential structures embedded in time. By organizing combinatorial objects with a simple "juxtaposition" arithmetic, they break down the process of extracting statistical regularities into a set of summation and multiplication operations. In this aspect, these functions may help us understand the neural learning mechanisms in the process of extracting sequential relational structures, by revealing how object representations can build up to a compressed probabilistic representation, and vice versa, how a learned structure may bias predictions on discrete-time events. Overall, these functions can be a powerful tool to bridge the gap between implicit statistical learning without instruction and explicitly structured rule learning, and to reconcile the deviation of heuristics and biases from normative rules.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research (AFOSR) grant number FA9550-12-1-0457, the Office of Naval Research (ONR) grant number N00014-08-1-0042, and the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*(3), 170–176. doi: 10.1177/0963721412436806

Budescu, D. V. (1987). A Markov model for generation of random binary sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(1), 25–39. doi: 10.1037/0096-1523.13.1.25

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. doi: 10.1207/s15516709cog1402_1

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301–318. doi: 10.1037/0033-295x.104.2.301

Gazzaniga, M. S. (2008). *Human: The science behind what makes us unique*. New York: HarperCollins e-books.

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Computational Biology*, *5*(10), e1000532. doi: 10.1371/journal.pcbi.1000532

Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete mathematics*. Reading MA: Addison-Wesley.

Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review*, *116*(2), 454–461. doi: 10.1037/a0015241

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Huff, D. (1959). *How to take a chance*. New York: W. W. Norton.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, *107*(2), 397–402. doi: 10.1037/0033-295x.107.2.397

Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning Memory and Cognition*, *13*(3), 392–400. doi: 10.1037/0278-7393.13.3.392

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330–357. doi: 10.1037//0033-295X.109.2.330

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2012). *Computational cognitive neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu.

O'Reilly, R. C., Wyatte, D., & Rohrlich, J. (2014). Learning through time in the thalamocortical loops. *Preprint at: http://arxiv.org/abs/1407.3432*.

Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin*, *135*(2), 262–285. doi: 10.1037/a0014821

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178. doi: 10.1038/nn.3495

Sun, Y., OReilly, R. C., Bhattacharyya, R., Smith, J. W., Liu, X., & Wang, H. (2015). Latent structure in random sequences drives neural learning toward a rational bias. *Proceedings of the National Academy of Sciences*, *112*(12), 3788–3792. doi: 10.1073/pnas.1422036112

Sun, Y., Tweney, R. D., & Wang, H. (2010). Occurrence and nonoccurrence of random sequences: Comment on Hahn and Warren (2009). *Psychological Review*, *117*(2), 697–703. doi: 10.1037/a0018994

Sun, Y., & Wang, H. (2010a). Gambler's fallacy, hot hand belief, and time of patterns. *Judgment and Decision Making*, *5*(2), 124–132.

Sun, Y., & Wang, H. (2010b). Perception of randomness: On the time of streaks. *Cognitive Psychology*, *61*(4), 333–342. doi: 10.1016/j.cogpsych.2010.07.001

Sun, Y., & Wang, H. (2012). Perception of randomness: Subjective probability of alternation. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 1024–1029). Austin, TX: Cognitive Science Society.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. doi: 10.1126/science.1192788

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi: 10.1126/science.185.4157.1124