Manuscript

Draft 2014-0307

# Latent structure in random sequences drives neural learning toward a rational bias

Yanlong Sun

Randall C. O’Reilly

Hongbin Wang

# Abstract

TODO: need to rewrite: Recent developments in psychology and neuroscience suggests that the human mind develops structured probabilistic representations about the world and performs near-optimal Bayesian inferences. However, it remains elusive how the structured probability space has originated in the first place, and how cognitive biases can arise from normative probabilistic models. Here we show that without a pre-defined hypothesis space, structured representations of the learning environment, as well as “biases”, can naturally emerge through efficient neural encodings of the sensory inputs over time. We demonstrate the idea with a biologically realistic neural network model to simulate randomness perception. Our findings reveal that rich semantics can be developed in associative learning through temporal integration, providing the building blocks for the structured hypothesis space required by Bayesian inference. We further suggest that the statistical structures in the learning environment and the competition between implicit representations are the keys to bridging the gaps from object representation to probability induction and the gaps from low-level sensory processing to high-level cognition.

# [Introduction]

There is a surprising amount of systematic structure lurking within random sequences. For example, in the classic case of flipping a fair coin, where the probability of each outcome (Heads or Tails) is exactly .5 on every single trial, one would naturally assume that there is no possibility for some kind of interesting structure to emerge, given such a simple and extreme, desolate form of randomness. And yet, if one records the average amount of time it takes for a repetition (HH or TT) to first occur in a sequence (the *waiting time* statistic), it is significantly longer (6 tosses) than for an alternation (HT or HT, 4 tosses). This is despite the fact that, on average, repetitions and alternations are equally probable (as one would expect) – the *mean time* statistic. For both of these facts to be true, it must be that repetitions are more bunched together over time – they come in bursts, with greater spacing between, compared to alternations. Intuitively, this difference comes from the fact that repetitions can build upon each other (e.g., HHH contains 2 repetitions of HH), whereas alternations cannot. Another source of insight comes from the transition graph (Figure 1).

Is this difference just a strange mathematical curiosity, or could it possibly have deep implications for our cognitive-level perceptions of randomness? We show that a neural model based on a detailed biological understanding of the way the neocortex integrates information over time when processing sequences of events \cite{LeabraTI}, is naturally sensitive to both the mean time and alternation time statistics. Indeed, its behavior is explained by a simple averaging of the influences of both of these statistics, and this behavior emerges in the model over a wide range of parameters. Furthermore, this averaging dynamic directly produces the best-fitting bias-gain parameter for an existing Bayesian model of randomness judgments \cite{GT01} – this was previously an unexplained free parameter and obtained only through parameter fitting. We also show that we can extend this Bayesian model to better fit the full range of human data by including a higher-order pattern statistic, and the neurally-derived bias-gain parameter still provides the best fit to the human data in this extended model. Overall, this model provides a neural grounding for the pervasive *gambler’s fallacy* bias in human judgments, where people systematically discount repetitions, and emphasize alternations (e.g., in a famous example in the Monte Carlo casino in 1913, black repeated a record 26 times in a game of roulette – people’s began extreme betting on red after about 15 repetitions).

Our neural model (Figure 1a) is extremely simple, consisting of a sensory input layer with distinct nonoverlapping patterns for H vs. T, and an internal prediction layer that attempts to predict the next input, with the benefit of a prior temporal context information based on properties of the deep neocortical neurons (layers 5b and 6; \cite{LeabraTI}). Of course, this model will generally fail to accurately predict for a random sequence, but nevertheless the exercise (which is much more successful for all the systematic structure in the rest of the environment) leads to the development of representations encoding sequences of inputs, where the prior inputs in the sequence are encoded in the temporal context. We decoded these sequence representations through a reverse correlation technique, and classified them as Repetition (R) or Alternation (A). For a fair coin probability of ½, the model had a ratio of R/A = .70 – repetition detectors were significantly less likely than alternation detectors. This demonstrates the gambler’s fallacy bias emerging naturally in the model, as a consequence of its sensitivity to the waiting time advantage of alternations compared to repetitions. To further characterize the nature of this bias, we systematically varied the probability of alternation (PA) in generating the sequence (i.e., departures from a fair coin), and measured the effects on the R/A ratio. We found a smooth curve, with an equilibrium point (R/A = 1) at PA = 3/7 (Figure 1b). In other words, alternations have to be this much less frequent to cancel out the bias in favor of alternations. See the supplemental online material for extensive exploration of model parameters and alternative model inputs, which demonstrate the robustness of this result, and that it depends critically on temporal integration and the waiting time statistics.

The neural model’s behavior can be replicated by a simple equation that averages the effects of the mean time and waiting time statistics:

R/A = E[THT] + E[T\*HT] / E[THH] + E[T\*HH] , (1)

where E[T] is the expected mean time, and E[T\*] is the expected waiting time (Figure 1c). This establishes a clear higher-level explanation for the emergent behavior of the model, allowing us to summarize its behavior as simply averaging the effects of these two relevant statistics over the random sequences.

Next, we asked whether it was possible to relate this emergent behavior of the neural model to an existing Bayesian model of randomness judgments \cite{GT01}. This model was fit to a massive database of 20,099 participant’s judgments of how random each of the length 5 sequences of H, T are \cite{Goodfellow38}. To account for the gambler’s fallacy bias, this model included a bias gain factor **λ** that weights the contribution of a local representativeness factor Lk in determining the probability of the k’th response (Rk) in a sequence being heads:

P(Rk = H) = 1 / (1 + e- **λ** Lk) (2)

Lk = … (3)

This equation shows that the bias gain value modulates the strength of the alternation bias – a value of **λ**= 0 produces “rational” judgments in accord with the mean time statistic, and higher values produce an increasing alternation bias. \cite{GT01} found that a **λ**value of around .51 produced the optimal fit to the human data (Figure 2a) – a moderate alternation bias. But they had no independent basis for specifying this value, beyond parameter fitting. In contrast, we are able to show that this value can be derived directly from the behavior of our neural model.

Specifically, given the form of equation (2), we can show that **λ**can be derived directly from the R/A ratio:

**λ**= -log2(R/A) (4)

For a fair coin, R/A = .7, resulting in **λ**= .51 – precisely the value that optimizes the fit to the human data. Thus, by using the emergent behavior of the neural model, we can independently anchor this previously free parameter, and obtain the best fit to the human data. This represents a remarkable convergence across multiple levels of analysis, and further bolsters the validity of our understanding for the nature and origin of the systematic preference for alternating sequences, and against repeating ones.

Finally, the \cite{GT01} model produced a very bad fit to one of the sequence inputs: HTHTH, which was judged by people to not be a very good random sequence, but the model ranked it highly. It seems that people have a bias against the higher-order repetition of the HT sequence. We were able to add this bias into the model as an additional additive term in equation 2, and this augmented model now produces an excellent fit to the full set of sequence data points (Figure 2b). Again, the optimal value of the bias gain factor, which now applies to both sources of bias, is the .51 value predicted by our neural model.

In conclusion, we find that the latent structure in simple probabilistic sequences shapes the learning dynamics in a neural model, producing a novel “rational” explanation for what has generally been considered a curious failure of human probabilistic understanding. The remarkable fit of the parameters derived from this neural model with a Bayesian model derived from very different considerations reinforces the idea that the temporal integration mechanisms in our neural model provide a good account of human information integration over time. We have also recently shown how this same neural temporal integration and learning framework can account for human causal learning \cite{Blicket}, in a way that is also compatible with existing Bayesian models of causal reasoning \cite{GriffithsEtc}. This ability to bridge between levels of analysis across multiple domains represents a rare and important development, with the potential to both ground these abstract models in underlying neural mechanisms, and provide a simpler explanatory understanding of the emergent behavior of these neural models.

# References:

1. A. Pouget, J. M. Beck, W. J. Ma, P. E. Latham, Probabilistic brains: Knowns and unknowns. *Nature Neuroscience* **16**, 1170 (09//print, 2013).

2. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction. *Science* **331**, 1279 (March 11, 2011, 2011).

3. A. Gopnik, H. M. Wellman, Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin* **138**, 1085 (2012).

4. J. L. McClelland *et al.*, Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* **14**, 348 (2010).

5. A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases. *Science* **185**, 1124 (Sep. 27, 1974, 1974).

6. G. F. Marcus, Neither size fits all: Comment on McClelland et al. and Griffiths et al. *Trends in Cognitive Sciences* **14**, 346 (2010).

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

8. A. T. Oskarsson, L. Van Boven, G. H. McClelland, R. Hastie, What's next? Judging sequences of binary events. *Psychological Bulletin* **135**, 262 (2009).

9. R. S. Nickerson, The production and perception of randomness. *Psychological Review* **109**, 330 (2002).

10. D. Kahneman, A. Tversky, Subjective probability: A judgment of representativeness. *Cognitive Psychology* **3**, 430 (1972).

11. T. Gilovich, R. Vallone, A. Tversky, The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology* **17**, 295 (1985).

12. J. B. Tenenbaum, T. L. Griffiths, in *Proceedings of the 23rd annual conference of the cognitive science society,* J. D. Moore, K. Stenning, Eds. (Lawrence Erlbaum Associates, Mahwah, NJ, 2001), pp. 1036-1041.

13. T. L. Griffiths, J. B. Tenenbaum, in *Proceedings of the 23rd annual conference of the cognitive science society,* J. D. Moore, K. Stenning, Eds. (Lawrence Erlbaum Associates, Mahwah, NJ, 2001), pp. 398-403.

14. Y. Kareev, Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review* **107**, 397 (2000).

15. R. Falk, R. Falk, P. Ayton, Subjective patterns of randomness and choice: Some consequences of collective responses. *Journal of Experimental Psychology: Human Perception and Performance* **35**, 203 (2009).

16. R. Falk, C. Konold, Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review* **104**, 301 (1997).

17. S. A. Huettel, P. B. Mack, G. McCarthy, Perceiving patterns in random series: Dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience* **5**, 485 (2002).

18. Y. Sun, H. Wang, Perception of randomness: On the time of streaks. *Cognitive Psychology* **61**, 333 (2010).

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

20. D. M. Oppenheimer, B. Monin, The retrospective gambler’s fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making* **4**, 326 (2009).

21. B. Aisa, B. Mingus, R. C. O’Reilly, The Emergent neural modeling system. *Neural Networks* **21**, 1146 (2008).

22. J. L. Elman, Finding structure in time. *Cognitive Science* **14**, 179 (1990).

23. R. B. Ivry, R. T. Knight, Making order from chaos: The misguided frontal lobe. *Nature Neuroscience* **5**, 394 (2002).

24. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, Bayesian inference with probabilistic population codes. *Nature Neuroscience* **9**, 1432 (2006).

25. M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429 (01/24/print, 2002).

26. G. T. M. Altmann, Why emergentist accounts of cognition are more theoretically constraining than structured probability accounts: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences* **14**, 340 (2010).

27. S. Pinker, *How the mind works*. (Norton, New York, 1997).