Mental Sampling in Multimodal Representations

Jian-Qiao Zhu
Department of Psychology
University of Warwick
j.zhu@warwick.ac.uk

Adam N. Sanborn
Department of Psychology
University of Warwick
a.n.sanborn@warwick.ac.uk

Nick Chater
Behavioural Science Group
Warwick Business School
nick.chater@wbs.ac.uk

Abstract

Both resources in the natural environment and concepts in a semantic space are distributed “patchily”, with large gaps in between the patches. To describe people’s internal and external foraging behavior, various random walk models have been proposed. In particular, internal foraging has been modeled as sampling: in order to gather relevant information for making a decision, people draw samples from a mental representation using random-walk algorithms such as Markov chain Monte Carlo (MCMC). However, two common empirical observations argue against simple sampling algorithms such as MCMC. First, the spatial structure is often best described by a Lévy flight distribution: the probability of the distance between two successive locations follows a power-law on the distances. Second, the temporal structure of the sampling that humans and other animals produce have long-range, slowly decaying serial correlations characterized as $1/f$-like fluctuations. We propose that mental sampling is not done by simple MCMC, but is instead adapted to multimodal representations and is implemented by Metropolis-coupled Markov chain Monte Carlo (MC^3), one of the first algorithms developed for sampling from multimodal distributions. MC^3 involves running multiple Markov chains in parallel but with target distributions of different temperatures, and it swaps the states of the chains whenever a better location is found. Heated chains more readily traverse valleys in the probability landscape to propose moves to far-away peaks, while the colder chains make the local steps that explore the current peak or patch. We show that MC^3 generates distances between successive samples that follow a Lévy flight distribution and $1/f$-like serial correlations, providing a single mechanistic account of these two puzzling empirical phenomena.

1 Introduction

In many complex domains, such as vision, motor control, language, categorization or common-sense reasoning, human behavior is consistent with the predictions of Bayesian models (e.g., [3, 25, 5, 2, 14, 17, 37, 39]). Bayes’ theorem prescribes a simple normative method to combine prior beliefs with new information to make inferences about the world. However, the sheer number of hypotheses that must be considered in complex domains makes exact Bayesian inference intractable. Instead it must be that individuals are performing some kind of approximate inference [34, 24].

Sampling is a way to perform approximation for Bayesian models in complex problems that makes many difficult computations easy: instead of integrating over vast hypothesis spaces, samples of hypotheses can be drawn from the posterior distribution. The computational cost of sample-based approximations only scales with the number of samples rather than the dimensionality of the hypothesis space, though using small numbers of samples result in particular biases in inference.

Interestingly, the biases in inference that are introduced by using a small number of samples match some of the biases observed in human behavior. For example, probability matching [34], anchoring effects [18], and many reasoning fallacies [7, 24] can all be explained in this way. However, there is as of yet no consensus on the exact nature of the algorithm used to sample from human mental representations.

Previous work has posited that people either use direct sampling or Markov chain Monte Carlo (MCMC) to sample from their posterior distribution over hypotheses [34, 18, 7, 24]. In this paper, we demonstrate that these algorithms cannot explain two key empirical effects that have been found in a wide variety of tasks. In particular, these algorithms do not produce distances between samples that follow a Lévy flight distribution, and separately they do not produce autocorrelations that follow $1/f$ scaling. To find a sampling algorithm that does match these empirical effects, we note that mental representations have been shown to be "patchy" with high probability regions separated by large regions of low probability. We then compare one of the first algorithms developed for sampling from multimodal distributions, Metropolis-coupled MCMC (MC^3), and demonstrates that it produces both key empirical phenomena. Previously Lévy flight distributions and $1/f$ scaling have been explained separately as the result of efficient search and the signal of self-organizing behavior respectively [33, 31], and we provide the first account that can explain both phenomena as the result of the same purposeful mental activity.

1.1 Spatial structure of mental samples

In the real world, resources are rarely distributed uniformly in the environment. Food, water, and other critical nature resources often occur in spatially isolated patches with large gaps in between. Therefore, humans and other animals' foraging behaviors should adapt to such patchy environments. In fact, foraging behaviors have been observed to display a Lévy flight, which is a class of random walk whose step lengths follow a heavy-tailed power-law distribution [26]. In a Lévy flight distribution, the probability of executing a jump of length l is given by:

$$P(l) \sim l^{-\mu} \tag{1}$$

where $1 < \mu \leq 3$. The values $\mu \leq 1$ do not correspond to normalizable probability distributions. Examples of mobility patterns following the Lévy flight has been recorded in Albatrosses [32], marine predators [27], monkeys [21], and humans [13].

Lévy flights are advantageous in patchy environments where resources are sparsely and randomly distributed because the probability of returning to a previously visited target site is smaller than in a standard random walk. In the same patchy environment, Lévy flights can visit more new target sites than a random walk does [4]. Interestingly, it has been proven that in foraging the optimal exponent is $\mu = 2$ regardless the dimensionality of the space if (1) the target sites are sparse, (2) they can be visited any number of times, and (3) the forager can only detect and remember a target site in a close vicinity [33].

Remarkably, mental representations of concepts are also patchy and the distance between mental samples also follows a Lévy flight distribution. For example, in semantic fluency tasks (e.g., asking participants to "name as many animals as you can"), the retrieved animals tend to form clusters (e.g., pets, water animals, African animals) [30, 1]. This same task has also been found to produce Lévy flight distributions of inter-response intervals (IRI) [23], which can be considered a measure of distance between samples by making the reasonable assumption that there is a linear relationship between IRI and mental distance[1].

[1] There are various ways to make the link between IRI and distance between samples. One is to assume that it takes longer to transition to a sample that is further away in the mental space. A second is to assume that while generating any sample takes the same fixed amount of time, there are unreported samples that are generated between each reported sample, and that the sampler has travelled further the more unreported samples that are generated; unreported samples are plausible in this task because participants are only given credit for each new animal name they report.

1.2 Temporal structure of mental samples

Besides the spatial structure of the distance between two successive locations following a power-law distribution, a number of studies has reported that the temporal structure of many cognitive activities contains long-range, slowly decaying serial correlations. These correlations tend to follow a $1/f$ scaling law [16]:

$$C(k) \sim k^{-\alpha} \tag{2}$$

where $C(k)$ is the autocorrelation function of temporal lag k. The same phenomenon is often expressed in the frequency domain:

$$S(f) \sim f^{-\alpha} \tag{3}$$

where f is frequency, $S(f)$ is spectral power and $\alpha \in [0.5, 1.5]$ is considered $1/f$ scaling. The power spectra can be derived from submitting the time series to Fourier analysis. $1/f$ noise is also known as pink or flicker noise, which varies in predictability intermediately between white noise (no serial correlation, $S(f) \sim 1/f^0$) and brown noise (no correlation between increments, $S(f) \sim 1/f^2$). Note that Lévy flights are random walks so they do not produce $1/f$ noise, but $1/f^2$ noise instead.

$1/f$-like temporal fluctuations in human cognition were first reported in time estimation and spatial interval estimation tasks in which participants were asked to repeatedly estimate a pre-determined time interval of 1 second or spatial interval of 1 inch [12]. Subsequent studies have shown $1/f$ scaling laws in the response times of mental rotation, lexical decision, serial visual search, and parallel visual search [11], as well as the time to switch between different percepts when looking at a bistable stimulus (i.e., a Necker cube [8]).

Given that sampling can be described as a Lévy flight spatially and has $1/f$ autocorrelations (see Table 1 for summary), we now investigate which sampling algorithms can capture both the spatial and temporal structure of human cognition.

Table 1: Key findings for $1/f$ noise and Lévy flight in human mental sampling

Papers	Experiments	Main findings
[12]	Time interval estimation	Power spectra slopes of $[-1.2, -0.90]$
	Spatial interval estimation	Power spectra slope of -1
[11]	Mental rotation	RT power spectra slope of -0.7
	Lexical decision	RT power spectra slope of -0.9
	Serial search	RT power spectra slope of -0.7
	Parallel search	RT power spectra slope of -0.7
[23]	Memory retrieval task	IRI power-law exponents $\hat{\mu} \in [1.37, 1.98]$
[22]	Natural scene perception	Eye movement trajectories follow both $1/f$ noise and Lévy flight

2 Sampling algorithms

We consider three possible sampling algorithms that might be employed in human cognition: Direct Sampling (DS), Random walk Metropolis (RwM), and Metropolis-coupled MCMC (MC^3). We define DS as independently drawing samples in accord with the posterior probability distribution. DS is the most efficient algorithm for sampling of the three, but it may not be possible to implement in human cognition as it often requires calculating intractable normalizing constants that scale exponentially with the dimensionality of the hypothesis space [19, 6]. DS has been used to explain biases in human cognition such as probability matching [34].

MCMC algorithms can bypass the problem of the normalizing constant by simulating a Markov chain that transitions between states according only to the ratio of the probability of hypotheses [19]. We define RwM as a classical Metropolis-Hastings MCMC algorithm, which can be thought of as a random walker exploring the probability landscape of hypotheses, preferentially climbing the peaks of the posterior probability distribution [20, 15]. However, with limited number of samples, RwM is very unlikely to reach modes in the probability distribution that are separated by large regions of low probability. This leads to biased approximations of the posterior distribution [28, 24]. Random walks have been used to model clustered responses in memory retrieval [1], and RwM in particular has been used to model multistable perception [9], the anchoring effect [18], and various reasoning biases [7, 24].

Our third algorithm is MC^3, also known as parallel tempering or replica-exchange MCMC, was one of the first algorithms to successfully tackle the problem of multimodality [10]. MC^3 involves running M Markov chains in parallel, each at a different temperature: $T_1, T_2, ..., T_M$. In general, $1 = T_1 < T_2 < ... < T_M$, and T_1 is the temperature of the interest where the target distribution is unchanged. The purpose of the heated chains is to traverse valleys in the probability landscape to propose moves to far-away peaks (by sampling from heated target distributions: $\pi^{1/T}$), while the colder chains make the local steps that explore the current probability peak or patch. MC^3 decides whether to swap the states between two randomly chosen chains in every iteration [10]. In particular, swapping of chain i and j is accepted or rejected according to a Metropolis rule; hence, the name Metropolis-coupled MCMC

$$A^{swap} = \min\{1, \frac{\pi(x_j)^{1/T_i}\pi(x_i)^{1/T_j}}{\pi(x_i)^{1/T_i}\pi(x_j)^{1/T_j}}\} \tag{4}$$

Coupling induces dependence among the chains, so each chain is no longer Markovian. The stationary distribution of the entire set of chains is thus $\prod_{i=1}^{M} \pi^{1/T_i}$ but we only use samples from the cold chain ($T = 1$) to approximate the posterior distribution [10]. Pseudocode for MC^3 is presented below. Note that MC^3 reduces to RwM when the number of parallel chains $M = 1$.

Algorithm Metropolis-coupled Markov chain Monte Carlo

```
Choose a starting point $x_1$.
for $t = 2$ to $L$ do
    for $m = 1$ to $M$ do                                                  ▷ update all $M$ chains
        Draw a candidate sample $x' \sim \mathcal{N}(x_{t-1}^m, \sigma)$   ▷ Gaussian proposal distribution
        Sample $u \sim U[0, 1]$
        $A^m = \min\{1, [\frac{\pi(x')}{\pi(x_{t-1}^m)}]^{1/T_m}\}$
        if $u < A^m$ then $x_t^m = x'$ else $x_t^m = x_{t-1}^m$ end if     ▷ Metropolis acceptance rule
    end for
    repeat floor($M/2$) times                                              ▷ swapping scheme for Markov chains
        Randomly select two chain $i, j$ without repetition
        Sample $u \sim U[0, 1]$
        $A^{swap} = \min\{1, \frac{\pi(x_t^j)^{1/T_i}\pi(x_t^i)^{1/T_j}}{\pi(x_t^i)^{1/T_i}\pi(x_t^j)^{1/T_j}}\}$
        if $u < A^{swap}$ then swap($x_t^i, x_t^j$) end if                 ▷ Metropolis-coupled swapping rule
    end repeat
end for
```

3 Results

In this section, we evaluate whether the two key empirical effects of Lévy flights and $1/f$ autocorrelations can be produced by the Direct Sampling (DS), Random walk Metropolis (RwM), and Metropolis-coupled MCMC (MC^3) algorithms.

3.1 Lévy flight

We simulated a 2D patchy environment with $N_{mode} = 15$ Gaussian mixtures where the means are uniformly generated from $[-r, r]$ for both dimensions, where $r = 9$ and the covariance matrix is fixed as the identity matrix for all mixtures. This method will produce a patchy environment (for example the top panel of Figure 1). We ran DS, RwM, and MC^3 on this multimodal probability landscape, and the first 100 positions for each algorithm can be found in the top panel of Figure 1. The empirical flight distances were obtained by calculating the Euclidean distance between two consecutive positions of the sampler. For MC^3, only the positions of the cold chain ($T = 1$) were used.

Power-law distributions should produce straight lines in a log-log plot. Therefore, the power-law exponents were fitted by linear regression on the window-averaged log-binned flight distance data [23]. We used 10 non-overlapping windows that evenly split the x-axis, and cell means are represented in the yellow filled dots in the bottom panel of Figure 1. Fitting the cell means provides a lower-variance method for estimating the slope than fitting the log-binned data directly. Figure 1 (bottom panel) shows that only MC^3 can reproduce the distributional property of flight distance as a Lévy flight with estimated power-law exponent $\hat{\mu} = 1.27$. Both DS ($\hat{\mu} = -0.59$) and RwM ($\hat{\mu} = 0.91$) [2] produced values outside the range of a Lévy flight.

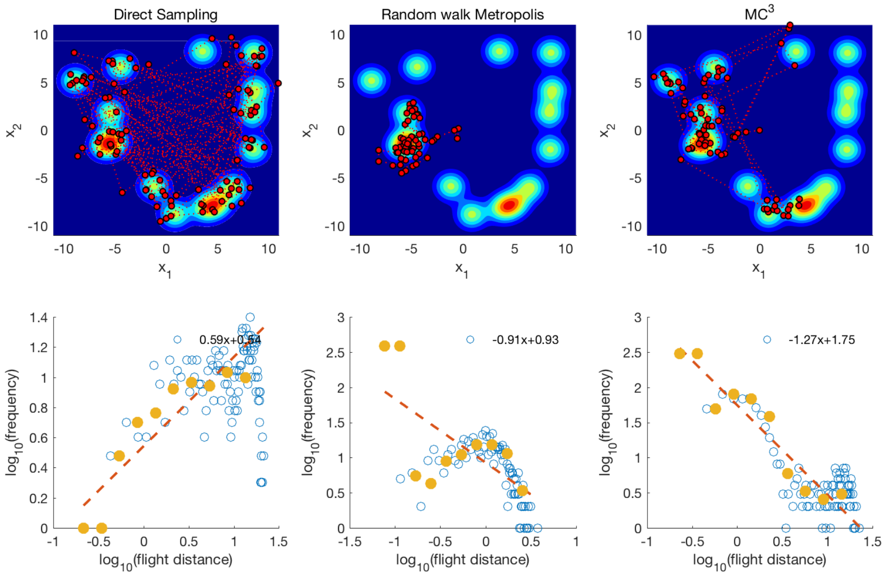

Figure 1: Searching behavior in a simulated 2D patchy environment of a 15 component Gaussian mixture. (*Left Panel*) the trajectory of first 100 positions (red dots in top panel) and the log-log plot of flight distance (bottom panel) for DS. The best-fitted lines used to estimate the Levy flight exponent (bottom panel red dashed lines) were based on cell means using non-overlapping windows (yellow filled dots) of the log-binned data (blue dots). (*Middle Panel*) the same plots for RwM algorithm. The Gaussian proposal distribution was an identity covariance matrix. (*Right Panel*) the same plots for MC^3 algorithm with 8 parallel chains with only the cold chain shown here. The Gaussian proposal distributions for all 8 chains had the same identity covariance matrix. For all the algorithms only the first 1024 samples were used.

We then investigated the impact of spatial sparsity on the estimated power-law exponents. In this simulation, the same number of Gaussian mixture were used but the range r was varied. The spatial sparsity was computed as the mean distance between Gaussian modes. With small or moderate spatial sparsity we found a positive relationship between spatial sparsity and the estimated power-law exponents for both DS and MC^3 (see Figure 2 left). In this range, only MC^3 produced power-law exponents in the range reported in human mental foraging studies (see Table 1), while both DS and RwM failed to do so. For all three algorithms, once spatial sparsity was too great only a single mode was explored and no large jumps were made.

We also checked whether MC^3 really is more suitable to explore patchy mental representations than RwM. In our simulated patchy environment, which used 15 identical Gaussian mixtures with identity covariance matrix, an optimal sampling algorithm should visit each mode equally often, hence will produce a uniform distribution of visit frequencies over all the modes. To this end, the effectiveness of exploring the representation was examined by computing a Kullback-Leibler divergence (KL) [19] between a uniform distribution over all modes and a the relative frequency of how often an algorithm visited each mode:

[2] The log-binned data for RwM shows that the algorithm is certainly not producing a power-law as the empirical flight distance distribution is not a straight line in the log-log plot

$$D_{KL}(\mathcal{H}_{1:t}||U) = \sum_{i=1}^{N_{mode}} \mathcal{H}_{1:t} \log \frac{\mathcal{H}_{1:t}}{1/N_{mode}} \tag{5}$$

where U is a discrete uniform distribution, N_{mode} is the number of identical Gaussian mixtures, and $\mathcal{H}$ is the empirical frequency of visited modes up to time t. Samples were assigned to the closest mode when determining these empirical frequencies. The faster the KL divergence for an algorithm reaches zero, the more effective the algorithm is at exploring the underlying environment and the DS algorithm serves as a benchmark for the other two algorithms. As shown in Figure 2 (middle), MC^3 quickly catches up to DS, while RwM lags far behind in exploring this patchy environment.

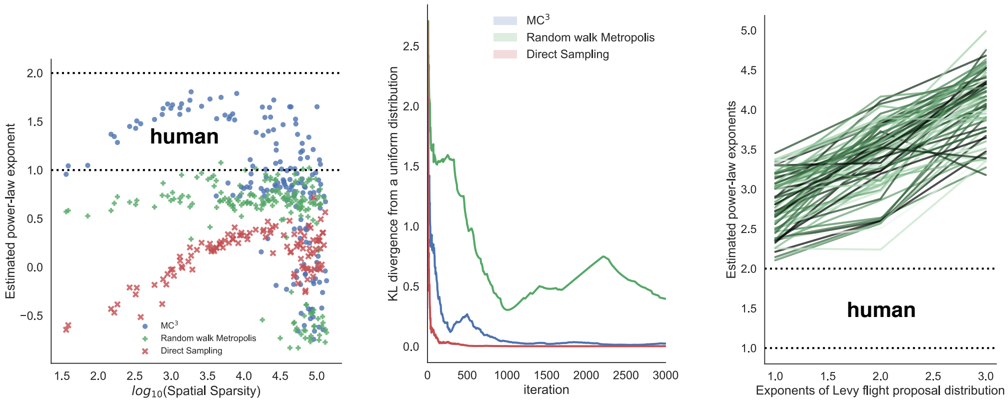

Figure 2: (*Left*) Estimated power-law exponents for flight distance distributions for the three sampling algorithms, manipulating the spatial sparsity of the Gaussian mixture environment. Spatial sparsity measurement was defined as the mean distance between modes. All three algorithms used the same settings as in Figure 1. The dashed lines show the range of human data. (*Middle*) KL divergence of mode visiting from the true distribution for the three sampling algorithms. The underlying patchy environments are the same for all three algorithms. (*Right*) Simulated RwM with a Lévy flight proposal distribution. Darker colors represent higher spatial sparsities. The dashed lines show the range of human data.

Of course it may seem that we were simply using the wrong proposal distribution for RwM. Instead of using a Gaussian proposal distribution we can use a Lévy flight proposal distribution, which will straightforwardly produce Lévy flights if the posterior distribution is uniform over the entire space (i.e., every proposed flight will be accepted). However, in a patchy environment a Lévy flight proposal distribution will not typically produce a Lévy flight distribution of distances between samples that has estimated power-law exponents in the range of human data, as also can be seen in Figure 2 (right) with different spatial sparsities. The reason for this is that the long jumps in the proposal distribution are unlikely to be successful: these long jumps often propose new states that lie in regions of nearly zero posterior probability.

3.2 $1/f$ noise

A typical interval estimation task requests participants to repeatedly produce an estimation for the same target interval over many repeated trials [12, 11]. For instance, participants were first given an example of a target interval (e.g., 1 second time interval or 1-inch spatial interval) and then repeated the judgments again and again without feedback for up to 1000 trials. These time series produced by human subjects showed $1/f$ noise, with an exponent close to 1. However, the log-log plot in human data is typically observed flatten out for the highest frequencies [12]. This effect has been explained as the result of two processes: fractional Brownian motion combined with white noise at the highest frequencies [12].

Figure 3 shows an example of time series for the first 1024 samples generated by DS (left), RwM (middle), and MC^3 (right). We used a simple Gaussian target distribution in this simulation because the distribution of responses produced by participants was indistinguishable from a Gaussian [12]. Note that RwM and MC^3 were initiated at the mode of the Gaussian distribution, and there was no burn-in period in our simulation. This results show that only MC^3 produces $1/f$ noise ($\hat{\alpha} = 1.01$), whereas DS produces white noise ($\hat{\alpha} = -0.01$) and RwM is closest to brown noise ($\hat{\alpha} = 1.64$). RwM tends to generate brown noise because, if every proposed sample is accepted, then the algorithm reduces to first-order autoregressive process (i.e., AR(1)) [38]. This is shown through simulation in Figure 4: when the Gaussian width (σ_{target}) becomes much greater width of the Gaussian proposal distribution ($\sigma_{proposal}$), RwM produces brown noise.

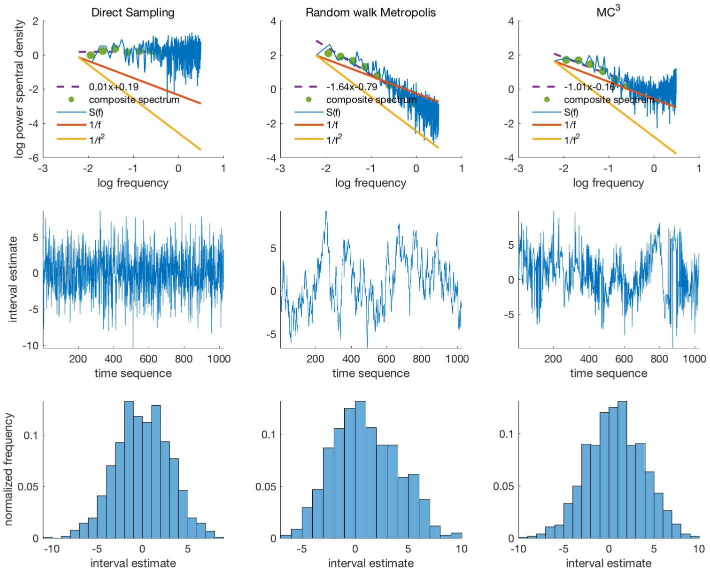

Figure 3: Sampling from a unimodal Gaussian $\mathcal{N}(0, 3)$. (*Left Panel*) the power spectra, traceplot, and sample distribution of DS from top to bottom. The best-fitted lines in power spectra (top panel purple dashed lines) are estimated based on block-averaged periodograms (green filled dots; [29]). (*Middle Panel*) the same plots for RwM algorithm. (*Right Panel*) the same plots for MC^3 algorithm with 2 parallel chains and only the cold chain is shown here. The result will be similar to using 8 parallel chains if we restrict swapping between neighboring chains only. For all the algorithms here, the first 1024 samples were used.

In contrast, MC^3 has a tendency to produce $1/f$ noise when the acceptance rate is high. It has been shown that the sum of as few as three AR(1) processes with widely distributed autoregressive coefficients produces an approximation to $1/f$ noise [36]. As the higher-temperature chains can be thought of as very roughly similar to AR(1) processes with lower autoregressive coefficients, this may explain why the asymptotic behavior of the MC^3 is $1/f$ noise.

What is also interesting about MC^3 is that it is a single process that is able to produce both the $1/f$ slope at lower frequencies as well as the flattening of the slope at higher frequencies, which was ascribed to two different processes by [12]. The reason MC^3 produces this result appears to be because when two chains with similar temperatures find states with similar posterior probability they will repeatedly swap back and forth, which can produce high frequency oscillations in the coldest chain.

4 Discussion

Lévy flights are advantageous in a patchy world, and have been observed in many foraging task with humans and other animals. A random walk with Gaussian steps does not produce the occasional long-distance jump as a Lévy flight does. However, the swapping scheme between parallel chains of MC^3 enables it to produce Lévy-like scaling in the flight distance distribution. Additionally MC^3

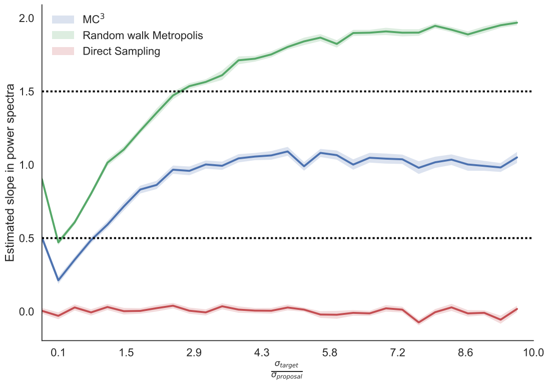

Figure 4: Estimated slopes in the power spectra are related to the ratio between Gaussian width and proposal step size. When the ratio is low, the acceptance rate of proposed sample should be low; it is the opposite case for the high ratio. The asymptotic behaviors of MC^3 are $1/f$ noise, of RwM are brown noise, and of DS are white noise.

produces the long-range slowly-decaying temporal correlations of $1/f$ scaling. This long-range dependence rules out any sampling algorithm that draws independent samples from the posterior distribution, such as DS, since the sample sequence would have no serial correlation (i.e., white noise). It also rules out RwM because the current sample solely depends on the previous sample. Both of these results suggest that the algorithms people use to sample mental representations are more complex than DS or RwM, and, like MC^3, are instead adapted to sampling from multimodal distributions.

However, if people are adapted to multimodal distributions, their behavior appears to have the same temporal pattern even when they are actually sampling from a unimodal distribution. In Gilden's experiments, the overall distribution of estimated intervals (i.e., ignoring serial order) was not multimodal, instead it was indistinguishable from a Gaussian distribution [12]. Assuming that the posterior distribution in the hypothesis space is also unimodal then it is somewhat inefficient to use MC^3 rather than simple MCMC. Potentially the brain is hardwired to use particular algorithms, or it is slow to adapt to unimodal representations because it is very difficult to know that a distribution is unimodal rather than just a single mode in a patchy space.

Previous explanations of scale-free phenomenon in human cognitions such as self-organized criticality argue that $1/f$ noise is generated from the interactions of many simple processes that produce such hallmarks of complexity [31]. Other explanations assume that it is due to a mixture of scaled processes like noise in attention or noise in our ability to perform cognitive tasks [35]. These approaches argue that $1/f$ noise is a general property of cognition, and do not tie it to other empirical effects. Our explanation of this scale-free process is more mechanistic, assuming that they reflect the cognitive need to gather vital resources in a multimodal world. While autocorrelations make samplers less effective when sampling from simple distributions, they may need to be tolerated in a multimodal world in order to sample other isolated modes.

Of course, we do not claim that MC^3 is the only sampling algorithm that is able to produce both $1/f$ noise and Lévy flights. It is possible that other algorithms that deal better with multimodality than MCMC, such as running multiple non-random walk Markov chains in parallel or Hamiltonian Monte Carlo, could produce similar results. Future work will explore which algorithms can match these key human data.

References

[1] J. T. Abbott, J. L. Austerweil, and T. L. Griffiths. Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems*, pages 3050–3058, 2012.

[2] J. R. Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3):409, 1991.

[3] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

[4] G. Berkolaiko, S. Havlin, H. Larralde, and G. Weiss. Expected number of distinct sites visited by n lévy flights on a one-dimensional lattice. *Physical Review E*, 53(6):5774, 1996.

[5] N. Chater and C. D. Manning. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344, 2006.

[6] N. Chater, J. B. Tenenbaum, and A. Yuille. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291, 2006.

[7] I. Dasgupta, E. Schulz, and S. J. Gershman. Where do hypotheses come from? Technical report, Center for Brains, Minds and Machines (CBMM), 2016.

[8] J. Gao, V. A. Billock, I. Merk, W. Tung, K. D. White, J. Harris, and V. P. Roychowdhury. Inertia and memory in ambiguous visual perception. *Cognitive Processing*, 7(2):105–112, 2006.

[9] S. J. Gershman, E. Vul, and J. B. Tenenbaum. Multistability and perceptual inference. *Neural Computation*, 24(1):1–24, 2012.

[10] C. J. Geyer. Markov chain monte carlo maximum likelihood. 1991.

[11] D. L. Gilden. Fluctuations in the time required for elementary decisions. *Psychological Science*, 8(4):296–301, 1997.

[12] D. L. Gilden, T. Thornton, and M. W. Mallon. $1/f$ noise in human cognition. *Science*, 267(5205):1837, 1995.

[13] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[14] T. L. Griffiths and J. B. Tenenbaum. Predicting the future as bayesian inference: people combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, 140(4):725, 2011.

[15] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[16] C. T. Kello, G. D. Brown, R. Ferrer-i Cancho, J. G. Holden, K. Linkenkaer-Hansen, T. Rhodes, and G. C. Van Orden. Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14(5):223–232, 2010.

[17] C. Kemp and J. B. Tenenbaum. Structured statistical models of inductive reasoning. *Psychological Review*, 116(1):20, 2009.

[18] F. Lieder, T. Griffiths, and N. Goodman. Burn-in, bias, and the rationality of anchoring. In *Advances in Neural Information Processing Systems*, pages 2690–2798, 2012.

[19] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[20] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[21] G. Ramos-Fernández, J. L. Mateos, O. Miramontes, G. Cocho, H. Larralde, and B. Ayala-Orozco. Lévy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi). *Behavioral Ecology and Sociobiology*, 55(3):223–230, 2004.

[22] T. Rhodes, C. T. Kello, and B. Kerster. Distributional and temporal properties of eye movement trajectories in scene perception. In *33th Annual Meeting of the Cognitive Science Society*, 2011.

[23] T. Rhodes and M. T. Turvey. Human memory retrieval as lévy foraging. *Physica A: Statistical Mechanics and its Applications*, 385(1):255–260, 2007.

[24] A. N. Sanborn and N. Chater. Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12):883–893, 2016.

[25] A. N. Sanborn, V. K. Mansinghka, and T. L. Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, 120(2):411, 2013.

[26] M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch. Lévy flights and related topics in physics. *Lecture Notes in Physics*, 450:52, 1995.

[27] D. W. Sims, E. J. Southall, N. E. Humphries, G. C. Hays, C. J. Bradshaw, J. W. Pitchford, A. James, M. Z. Ahmed, A. S. Brierley, M. A. Hindell, et al. Scaling laws of marine predator search behaviour. *Nature*, 451(7182):1098–1102, 2008.

[28] R. H. Swendsen and J.-S. Wang. Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.

[29] T. L. Thornton and D. L. Gilden. Provenance of correlations in psychological data. *Psychonomic Bulletin & Review*, 12(3):409–441, 2005.

[30] A. K. Troyer, M. Moscovitch, and G. Winocur. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, 11(1):138, 1997.

[31] G. C. Van Orden, J. G. Holden, and M. T. Turvey. Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132(3):331, 2003.

[32] G. M. Viswanathan, V. Afanasyev, S. Buldyrev, E. Murphy, et al. Lévy flight search patterns of wandering albatrosses. *Nature*, 381(6581):413, 1996.

[33] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. Da Luz, E. Raposo, and H. E. Stanley. Optimizing the success of random searches. *Nature*, 401(6756):911–914, 1999.

[34] E. Vul, N. Goodman, T. L. Griffiths, and J. B. Tenenbaum. One and done? optimal decisions from very few samples. *Cognitive science*, 38(4):599–637, 2014.

[35] E.-J. Wagenmakers, S. Farrell, and R. Ratcliff. Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin & Review*, 11(4):579–615, 2004.

[36] L. M. Ward. *Dynamical Cognitive Science*. MIT press, 2002.

[37] D. M. Wolpert. Probabilistic models in human sensorimotor control. *Human Movement Science*, 26(4):511–524, 2007.

[38] J. Xu and T. L. Griffiths. How memory biases affect information transmission: A rational analysis of serial reproduction. In *Advances in Neural Information Processing Systems*, pages 1809–1816, 2009.

[39] A. Yuille and D. Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.