LETTER **Communicated by Robert Kass**

A Set Probability Technique for Detecting Relative Time Order Across Multiple Neurons

Anne C. Smith
annesmith@ucdavis.edu
Department of Anesthesiology and Pain Medicine, University of California at Davis, Davis, CA 95616, U.S.A.

Peter Smith
p.smith@maths.keele.ac.uk
Department of Mathematics, University of Keele, Keele, Staffordshire, ST5 5BG, U.K.

With the development of multielectrode recording techniques, it is possible to measure the cell firing patterns of multiple neurons simultaneously, generating a large quantity of data. Identification of the firing patterns within these large groups of cells is an important and a challenging problem in data analysis. Here, we consider the problem of measuring the significance of a repeat in the cell firing sequence across arbitrary numbers of cells. In particular, we consider the question, given a ranked order of cells numbered 1 to N, what is the probability that another sequence of length n contains j consecutive increasing elements? Assuming each element of the sequence is drawn with replacement from the numbers 1 through N, we derive a recursive formula for the probability of the sequence of length j or more. For $n < 2j$, a closed-form solution is derived. For $n \geq 2j$, we obtain upper and lower bounds for these probabilities for various combinations of parameter values. These can be computed very quickly. For a typical case with small N (<10) and large n (<3000), sequences of 7 and 8 are statistically very unlikely. A potential application of this technique is in the detection of repeats in hippocampal place cell order during sleep. Unlike most previous articles on increasing runs in random lists, we use a probability approach based on sets of overlapping sequences.

1 Introduction

Development of analysis techniques to find temporal patterns in large sets of neural spike train data is an important current research problem (Buzsáki, 2004; Brown, Kass, & Mitra, 2004). This is increasingly critical because of technology advances allowing large numbers of individual neurons (from 10 up to over 100) to be recorded simultaneously. One common target of this multielectrode recording technology is the hippocampal place cell. Place

cells are neurons that fire in a specific temporal order when rats navigate through space (O'Keefe & Dostrovsky, 1971). There is growing experimental evidence from place cells in rats (Pavlides & Winson, 1989; Wilson & McNaughton, 1994; Skaggs & McNaughton, 1996; Nádasdy, Hirase, Czurkó, Csicsvari, & Buzsáki, 1999; Kudrimoti, Barnes, & McNaughton, 1999; Louie & Wilson, 2001; Lee & Wilson, 2002, 2004) and other animals (Dave & Margoliash, 2000; Hoffman & McNaughton, 2002) that temporal ordering of place cell firing from behavior persists during both rapid eye movement (REM) and slow-wave sleep (SWS) stages of subsequent sleep. However, the detection of these patterns is particularly difficult due to shortcomings in statistical analysis techniques for multiple neurons and due to the lack of an observed behavioral correlate often present during awake recording.

Techniques used to analyze temporal order among multiple neurons range from analysis of pairwise correlations (Wilson & McNaughton, 1994; Kudrimoti et al., 1999), to the joint peristimulus time histogram (JPSTH) for triplets (Aertsen, Gerstein, Habib, & Palm, 1989), to template matching techniques for multiple neurons (Abeles & Gerstein, 1988; Nádasdy et al., 1999; Louie & Wilson, 2001). Other techniques include the unitary events analysis (Grün, Diesmann, & Aertsen, 2001) and gravity methods (Gerstein & Aertsen, 1985). Recently, Lee and Wilson (2002, 2004) have employed a combinatorial method. This method is different from previous approaches as it relies on an initial parsing of the data and then focuses on whether longer sequences within the data are likely to have occurred by chance. Here, using a similar overall approach but with different assumptions, we introduce a complementary set probability technique.

Our approach can be described succinctly using elementary probability as follows. Assume an urn contains balls numbered 1 to N. Pick a ball at random, record its number, and replace it in the urn. Repeat this procedure n times, resulting in a list with n numbers. This list is called a *word*. The goal is to compute the probability that the word has a strictly increasing sequence (or *run*) of numbers of length j or more. For example, with $j = 5$, $n = 9$, and $N = 8$, one possible word is $\{5, 3, \mathbf{2, 3, 5, 7, 8}, 2, 2\}$ where the run of 5 is shown in bold. In the example of place cell firing, the parameters correspond as follows. The place cells observed during behavior are each assigned a number (1 through N) according to their position in the firing order. After smoothing the sleep data, we next write a list of length n, composed of numbers from 1 to N, corresponding to the observed order of cell firing during a subsequent sleep epoch. Parameter n can be larger or smaller than N. The parameter j corresponds to the length of the longest sequence of strictly increasing numbers observed within the list of n numbers. If the computed probability of j or more occurring by chance is low enough, then we can conclude that the temporal order of cell firing from behavior has been preserved during sleep. This article focuses on the development of an efficient technique to compute recursively upper and lower bounds for this probability of a strictly increasing subsequence of length j or more. Because

the calculations can be carried out very quickly, it is possible to study the properties of the probability function across a wide range of parameter values.

Although longest increasing subsequence (LIS) problems have been studied in probability and combinatorics (Wolfowitz, 1944; Schensted, 1960; Knuth, 1970; Chryssaphinou & Vaggelatou, 2001), as well as in computer science (Albert et al., 2004; Bespamyatnikh & Segal, 2000), our approach is different from these methods and from that of Lee and Wilson (2004) in two respects. First, our approach based on set probability yields explicit formulas for the required probabilities. Second, we assume each element in the word is chosen with replacement from the reference sequence of length N. Thus, the resulting hypothesis is not restrictive, and the probability bounds are computationally feasible over a wide range of parameter values.

2 Methods

We wish to compute the probability of j or more strictly increasing consecutive elements occurring by chance in a word of length n, where each element is chosen independently with replacement from a reference sequence $\{1,\ldots,N\}$. We do this in two steps. First, we compute the probability that j numbers picked independently with replacement from N numbers are strictly increasing. Second, we construct disjoint events based on the starting position of the strictly increasing sequence within the word. Since the events are now disjoint, it is possible to compute the required probability by summing them.

2.1 Step 1. Define the event $F_r(j, n, N)$ as j consecutive increasing elements starting at element r. It follows that $1 \le r \le n - j + 1$. Other numbers in the word can take any values. We start by asking how many ways one can pick j different numbers from an ordered list $\{1,\ldots,N\}$. This is $N!/(j!(N-j)!)$. Since these numbers are distinct, they can always be ordered in such a way as to make them strictly increasing. We then divide this number by the total possible combinations of j numbers from N, making

$$\Pr[F_r(j, n, N)] = (N^{-j})\frac{N!}{j!(N-j)!}, \tag{2.1}$$

which is in fact independent of r and n. We call this probability $p_{j,N}$.

An alternative method to compute this probability is by construction. If we count all the possible ways that we can get j strictly increasing combinations from N numbers, we get

$$\Pr[F_r(j, n, N)] = (N^{-j}) \sum_{i_{j-1}=1}^{N-j+1} \sum_{i_{j-2}=1}^{i_{j-1}} \cdots \sum_{i_1=1}^{i_2} i_1. \tag{2.2}$$

Using induction, it can be shown that the multiple sum is the binomial coefficient, that is,

$$\frac{N!}{j!(N-j)!} = \sum_{i_{j-1}=1}^{N-j+1} \sum_{i_{j-2}=1}^{i_{j-1}} \cdots \sum_{i_1=1}^{i_2} i_1. \tag{2.3}$$

We calculate the probability $p_{j,N}$ using equation 2.1, as it is computationally less intensive.

2.2 Step 2. To avoid overlapping events, we now define $F_r^*(j, n, N)$ to be the event that a strictly increasing word of length j starts at position r with no words of length j occurring before or including part of $F_r^*(j, n, N)$. It follows that the number at position $r-1 (r \geq 2)$ must not be less than the number at position r and that there can be any numbers from positions $r+j$ to n. Since there can be any numbers in positions $r+j$ to n, our calculations yield the probability of a strictly increasing sequence of j or more. The event $F_r^*(j, n, N)$ can be expressed in the form

$$F_r^*(j, n, N) = \begin{cases} F_1 & r = 1 \\ F_r \backslash F_{r-1} & r < j+1 \\ (F_r \backslash F_{r-1}) \backslash (F_1 \cup F_2 \cup \cdots \cup F_{r-j}) & r \geq j+1 \end{cases} \tag{2.4}$$

where the notation $A \backslash B$ means the event A but not the event B, and the terms in parentheses (j, n, N) have been dropped, for simplicity, on the right-hand side.

Since these events are disjoint, the desired total probability of a strictly increasing run of length j or more in n, denoted by H_j, is given by

$$H_j = \sum_{r=1}^{n-j+1} \Pr[F_r^*(j, n, N)]. \tag{2.5}$$

Note that the probability of exactly j strictly increasing numbers can be computed from $H_j - H_{j+1}$. It remains to compute the terms in equation 2.4. For the top equation where $r = 1$, F_1 is computed directly from equation 2.1 and is the probability of a sequence of j or more strictly increasing elements starting at the first position. For $1 < r < j+1$, we note that

$$\Pr(F_r \backslash F_{r-1}) = \Pr(F_r) - \Pr(F_r \cap F_{r-1}), \tag{2.6}$$

where

$$\Pr(F_r \cap F_{r-1}) = \Pr[F_{r-1}(j+1, n, N)]. \tag{2.7}$$

That is, the probability of an increasing sequence of length j or more, starting at position r and with an element greater than or equal to the element at $r-1$, is equal to the probability of a sequence length j or more, starting at r, minus the probability of a sequence of length $j+1$ or more starting at position $r-1$. Note that $\Pr(F_r \backslash F_{r-1})$ does not depend on its starting position in the word, r, or the word length, n. For $n < 2j$, the computations are complete, and H_j can be computed exactly from

$$\begin{aligned} H_j &= p_{j,N} + (n-j)(p_{j,N} - p_{j+1,N}) \\ &= \frac{N![(N+1)j(n-j) + N(j+1)]}{N^{j+1}(j+1)!(N-j)!}. \end{aligned} \tag{2.8}$$

To define disjoint events at larger values of r ($r \geq j+1$) the computation is more complicated. The formula in equation 2.4 can be expanded as follows:

$$\begin{aligned} \Pr(F_r^*) &= \Pr[(F_r \backslash F_{r-1}) \backslash (F_1 \cup F_2 \cup \cdots \cup F_{r-j})] \\ &= \Pr[(F_r \backslash F_{r-1})] - \Pr(F_r \backslash F_{r-1}) \cup (F_1 \cup F_2 \cup \cdots \cup F_{r-j}) \\ &= \Pr(F_r) - \Pr(F_r \cap F_{r-1}) - \Pr\left(F_r \cap \bigcup_{s=1}^{r-j} F_s\right) \\ &\quad + \Pr\left(F_r \cap F_{r-1} \cap \bigcup_{s=1}^{r-j} F_s\right). \end{aligned} \tag{2.9}$$

In this expression,

$$\Pr\left(F_r \cap \bigcup_{s=1}^{r-j} F_s\right) = \Pr(F_r)\Pr\left(\bigcup_{s=1}^{r-j} F_s\right) = \Pr(F_r)\sum_{s=1}^{r-j} \Pr(F_s^*), \tag{2.10}$$

since F_r does not overlap any $F_{s,}$ and

$$\begin{aligned} \Pr\left(F_r \cap F_{r-1} \cap \bigcup_{s=1}^{r-j} F_s\right) &= \Pr\left(F_r \cup F_{r-1} \cup \bigcup_{s=1}^{r-j} F_s\right) - \Pr(F_r) - \Pr(F_{r-1}) \\ &\quad - \Pr\left(\bigcup_{s=1}^{r-j} F_s\right) + \Pr(F_r \cap F_{r-1}) + \Pr\left(F_r \cap \bigcup_{s=1}^{r-j} F_s\right) \\ &\quad + \Pr\left(F_{r-1} \cap \bigcup_{s=1}^{r-j} F_s\right). \end{aligned} \tag{2.11}$$

Therefore,

$$\begin{aligned}\Pr(F^*_{r,j,N}) &= \Pr\left(F_r \cup F_{r-1} \cup \bigcup_{s=1}^{r-j} F_s\right) \\ &\quad - \Pr(F_{r-1}) - \Pr\left(\bigcup_{s=1}^{r-j} F_s\right) + \Pr\left(F_{r-1} \cup \bigcup_{s=1}^{r-j} F_s\right) \\ &= \Pr\left(F_r \cup F_{r-1} \cup \bigcup_{s=1}^{r-j} F_s\right) - \Pr\left(F_{r-1} \cup \bigcup_{s=1}^{r-j} F_s\right). \end{aligned} \tag{2.12}$$

The multiple union terms can be evaluated using the identity (Grimmett & Stirzaker, 1982)

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_i \Pr(A_i) - \sum_{i_1<i_2} \Pr\left(A_{i_1} \cap A_{i_2}\right) \\ &\quad + \sum_{i_1<i_2<i_3} \Pr\left(A_{i_1} \cap A_{i_2} \cap A_{i_3}\right) - \ldots \\ &\quad + (-1)^n \Pr\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}\right). \end{aligned} \tag{2.13}$$

Thus,

$$\begin{aligned}\Pr(F^*_{r,j,N}) &= \Pr(F_r) - \left(\Pr(F_r \cap F_{r-1}) + \Pr(F_r) \sum_{s=1}^{r-j} \Pr(F_s)\right) \\ &\quad + \Pr\left(F_r \cap F_{r-1} \cap \bigcup_{s=1}^{r-j} F_s\right). \end{aligned} \tag{2.14}$$

For relatively small values of n and N (<20), equation 2.11 can be evaluated exactly making use of the construction of disjoint events. For example, for the pairs-intersection terms,

$$\Pr(F_{r_1} \cap F_{r_2}) = \begin{cases} p^2_{j,N} & r_1 - r_2 \geq j \\ p_{r_1-r_2+j,N} & r_1 - r_2 < j \end{cases}.$$

That is, if there is a large gap (greater than or equal to j) between the runs' starting points, then the probability of a sequence of j or more is just squared. Alternatively, if the gap is small (less than j), then we compute the probability of a longer sequence of increasing elements. For larger values of n and N, the number of loops to be evaluated when expanding the union

terms becomes prohibitive, and it is more efficient to work with upper and lower bounds.

2.3 Upper Bound. An upper bound can be computed using

$$\begin{aligned}\Pr(F_r^*) &= \Pr[(F_r \setminus F_{r-1}) \setminus (F_1 \cup F_2 \cup \ldots \cup F_{r-j})] \\ &\leq \Pr[(F_r \setminus F_{r-1}) \setminus (F_1 \cup F_2 \cup \ldots \cup F_{r-j-1})] \\ &= \Pr(F_r \setminus F_{r-1}) - \Pr\left(F_r \cap \bigcup_{s=1}^{r-j-1} F_s\right) + \Pr\left(F_r \cap F_{r-1} \cap \bigcup_{s=1}^{r-j-1} F_s\right) \\ &\leq (p_{j,N} - p_{j+1,N})\left(1 - \sum_{s=1}^{r-j-1} \Pr(F_r^*)\right). \end{aligned} \tag{2.15}$$

Intuitively, this upper bound should be close to the true probability because only one event has been neglected. This is the event that an increasing sequence of length j or more starting at $r - j$ is part of a sequence of length j or more that also increases from r onwards.

2.4 Lower Bound. For the lower bound, we use Boole's inequality ($\Pr(\bigcup_i A_i) \leq \sum_i \Pr(A_i)$) and thus

$$\begin{aligned}\Pr(F_r^*) &= \Pr[(F_r \setminus F_{r-1}) \setminus (F_1 \cup F_2 \cup \ldots \cup F_{r-j})] \\ &\geq \Pr(F_r \setminus F_{r-1}) - \Pr(F_r) \sum_{s=1}^{r-j} \Pr(F_r) \\ &= \begin{cases} p_{j,N} - p_{j+1,N} - (r-j)p_{j,N}^2 & \text{if positive} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.16}$$

Because of the recursive structure of equations 2.15 and 2.16, the upper and lower bounds can be computed in seconds using Matlab (Mathworks, Natick, MA). (The software to perform these calculations is available online at www.ucdmc.ucdavis.edu/anesthesiology/staff/asmith.html.) In the next section we illustrate typical values for various parameter combinations.

2.5 Combinatorial/Shuffled Data Method. Our technique is most similar to the combinatorial technique developed by Lee and Wilson (2002, 2004) and used in the analysis of hippocampal place cell firing during non-REM sleep. Using the same definitions of reference sequence and word, their technique computes the probability that shuffled versions of the observed words contain a match or better to the reference sequence. As with the above

approach, if this probability is low, the experimenter might conclude that the order of the original reference sequence has been preserved. Their definition of a match or better is more complicated than used in our set probability technique and allows interruptions in the increasing sequence. In particular, they define a word as containing an (x, y) match if there are $x + y$ consecutive letters in the word of which at least x are strictly increasing. This means there are possibly as many as y interruptions in the increasing sequence. In addition, they define the parameter k as the number of distinct letters in the observed word. They suggest computing the required match probability by either a sequence shuffling technique, which can be computationally intensive, or by an approximate technique. To use the approximate technique, it is necessary to rank the matches based on a (subjective) decision about the acceptable balance between the number of interruptions and the length of the increasing sequence. In the approximate match ranking case, it is possible to derive algorithms to compute upper and lower bounds for the probability of the (x, y) matches. Across many trials, the final output of their method is a match-trial ratio with corresponding Z-score relative to the match-trial ratio expected by chance ($1/j!$ in our notation).

The main differences between our technique and the Lee and Wilson technique are that we do not allow sequence interruptions and we assume the observed word is chosen independently and with replacement from the reference sequence rather than basing the probabilities on the shuffled observed sequence. For comparison purposes with our method, we have computed the exact formulas for two of the match probabilities we would get using the Lee and Wilson technique in the special case where each letter is observed only once within the word and the reference sequence equal to the word length ($n = N = k$). We call this probability $P_{\text{shuffled}}(n, j)$. The probability of exactly shuffling a sequence of length j and getting j strictly increasing is

$$P_{\text{shuffled}}(j, j) = 1/j! \tag{2.17}$$

and for $j = n - 1$

$$P_{\text{shuffled}}(j, j-1) = (2j-1)/j! \tag{2.18}$$

We compare our technique with this technique in the next section.

3 Results

We illustrate our approach by considering place cells in rat hippocampal area CA1 and how to assess if their temporal order is preserved during sleep. First, we compare our method with some examples from the more elaborate method of Lee and Wilson (2004). Second, we discuss two theoretical

scenarios from non-REM and REM sleep using parameter values consistent with experimental measurements (Lee & Wilson, 2002; Nádasdy et al., 1999; Kudrimoti et al., 1999). By ordering the place cells in behavior from 1 to N, we compute either the exact probability or bounds for the probability of j or more consecutive increasing elements in any word of length n using equations 2.8, 2.15, and 2.16.

3.1 Comparison with Combinatorial/Shuffled Data Approach. At first inspection, one might assume that by making the assumption of replacement in our technique, our probabilities might always be lower than those computed using the shuffle technique of Lee and Wilson (2002, 2004). This is true for some parameter combinations. It is true, for example, for the cases derived in equations 2.17 and 2.18, as can be shown by comparing them with equation 2.8.

However, for other situations, this is not the case. Consider the example in Lee and Wilson (2004) where the reference sequence is (1,2,3,4,5,6,7,8,9), and the observed word is (5,1,4,6,9,7,8,4). They report the best match using their (x, y) matching procedure to be (5, 1) (based on the run of numbers **1 4 6** 9 **7 8)** and report the probability of this match or better to be 0.0580 based on exact computations, or bounded by 0.0195 and 0.1038. Using equations 2.15 and 2.16, our technique asks what the likelihood is of finding a run of four strictly increasing elements (here **1 4 6 9**) in eight numbers from a reference sequence of length 9. We estimate this event to be more likely, lying between 0.0871 and 0.0875. With such a short sequence (nine observations) and a reasonable spread of values (4 appears twice, and 2 and 3 do not appear), it appears impossible to determine statistically whether these numbers have been chosen with or without replacement from the reference sequence. Given this uncertainty and the simplicity of the current technique's hypothesis, this larger probability estimated by the set probability technique will be less likely to indicate a significant sequence replay.

It is also instructive to consider what happens to the computed match probabilities in cases where the reference sequence is longer. Consider, for example, that the same word (5,1,4,6,9,7,8,4) is observed, but now the reference sequence is composed of numbers 1 through 20. In the Lee and Wilson formulation, the computed match or better probability will be unaltered as the calculations rely on shuffling the observed word. Using our set probability method with $N = 20$, we would estimate the probability of four or more strictly increasing numbers as even larger and lying in the interval [0.1311, 0.1320].

3.2 Example: Non-REM (SWS) Sleep. In slow-wave (non-REM) sleep, it is hypothesized that compressed encoding of behavioral sequences may occur during sharp wave/ripple events (Buzsáki, 1989; Skaggs & McNaughton, 1996; August & Levy, 1999; Nádasdy et al., 1999; Lee & Wilson, 2002). These sharp/wave ripple events occur approximately once

per second and last for approximately 100 msec. During the event, the firing rate increases about seven-fold (Csicsvari, Hirase, Czurkó, Mamiya, & Buzsáki, 1999). We assume for our theoretical examples that the data have been preprocessed and parsed into words using the techniques outlined by previous authors (Nádasdy et al., 1999; Lee & Wilson, 2002). For example, one could assume that complex spike bursts (i.e., spike bursts with interspike intervals of less than, say, 6 msec) could be represented by a single spike occurring at the time of the first spike of the burst (Ranck, 1973; Nádasdy et al., 1999).

As a first application, we consider estimating the probability of sequence repeats within short SWS/ripple events. We call each SWS/ripple event a *trial*. To be consistent with experiments, we choose parameters as follows: j is 4 or 5, N is 8 or 9, and n ranges from 5 to 10 cell firings. Our computed probabilities, either exact or the upper or lower bounds, are shown in Table 1. Note that in the low-probability regime and with these choices of parameters, the upper and lower bounds computed using equations 2.14 and 2.15 are very close to one another. For fixed N, as the word length increases, we find the probability of finding a j or more length strictly

Table 1: Tabulation of Probability of j or More Consecutive Increasing Numbers Chosen from N in a Word of Length n.

j	N	n	$H_j = \Pr(j$ or more strictly increasing numbers)	E (number of matches in 300 trials)
4	8	5	0.0325	9
		6	0.0479	14
		7	0.0632	18
		8	[0.0783 0.0786]	[23 23]
		9	[0.0931 0.0937]	[27 28]
		10	[0.1076 0.1086]	[32 32]
4	9	5	0.0363	10
		6	0.0533	16
		7	0.0704	21
		8	[0.0871 0.0875]	[26 26]
		9	[0.1035 0.1042]	[31 31]
		10	[0.1194 0.1207]	[35 36]
5	9	5	0.0021	0
		6	0.0041	1
		7	0.0061	1
		8	0.0081	2
		9	0.0100	3
		10	[0.0120 0.0120]	[3 3]

Notes: When a single number is shown, the value is exact (to four decimal places) and is computed using equation 2.8. When two numbers are shown, the values are lower and upper bounds computed using equations 2.16 and 2.15. The last column shows the approximate expected number of times one might find j or more strictly increasing numbers in 300 Bernoulli trials ($300 \times H_j$).

increasing sequence moves from quite unlikely ($p < 0.05$) to quite likely. This is particularly noticeable when the $j = 4$ case is considered. For example, if $N = 9$ and $n > 5$, then the chance of observing a run of length $j = 4$ or more is greater than 0.05.

On average approximately one SWS/ripple event occurs every second. We consider the case of 300 such events, corresponding to observations over approximately 300 sec (5 min) of recording and consistent with the number of longer sequences ($n \geq 4$) observed by Lee and Wilson (2002). Assuming each event is a Bernoulli trial, we compute in Table 1 the expected number of trials with j or more strictly increasing numbers expected to occur by chance in 300 trials. For $j = 4$ and as n increases, we find the expected number of trials increases from as low as 9 in 300 to as high as 36 in 300. Note that for fixed n, there is not a big difference between the expected numbers of trials when $N = 8$ and $N = 9$ for the two $j = 4$ cases. However by increasing j from 4 to 5, the expected number of trials drops rapidly, and even in a word of length $n = 10$, we might expect to observe only 3 words in 300.

A second approach would be to concatenate all words together and look again for sequences of length j or more within the (much longer) word. By doing this, we treat the entire sleep episode as a time series. Over 50 minutes of sleep, one may expect to observe as many as 50×60 trials, each of which may contain a sequence with significant temporal ordering. In this case, we consider parameter values of $N = 10$ and $N = 20$ and vary n from 5 to 300 for various values of j. If the computed probability of j or more strictly increasing is still lower than p-values of 0.05 or 0.01, then this provides evidence that a statistically significant event is occurring. As expected, as sequence length j increases, the probability of observing a sequence of that length decreases (see Figure 1). As the length of the word (n) increases, this probability increases. In the case of $N = 10$ and $n = 3000$ (see Figure 1A), an observation of a single sequence of length 7 or more strictly increasing numbers is sufficient to conclude that the temporal order of the place cells has been preserved. Observation of a single sequence of six or more strictly increasing numbers would not be sufficient to indicate replay. For the larger $N = 20$ case (see Figure 1B), we require eight or more strictly increasing numbers to determine statistical significance.

3.3 Example: REM Sleep. In REM sleep, there is evidence of replay similar to the speed of the place cell firing during behavior (Louie & Wilson, 2002). In this case, for our theoretical study, we take $5 < n < 70$. This is consistent with up to seven words with $N = 10$ cells and is about the maximum number of replays that might fill a 5 minute episode of REM. In this case we show probability bounds for two values of N (10 and 20) for various choices of j (see Figure 2).

Interestingly, for both $N = 10$ (see Figure 2A) and $N = 20$ (see Figure 2B), an observation of a strictly increasing sequence of six or more is statistically

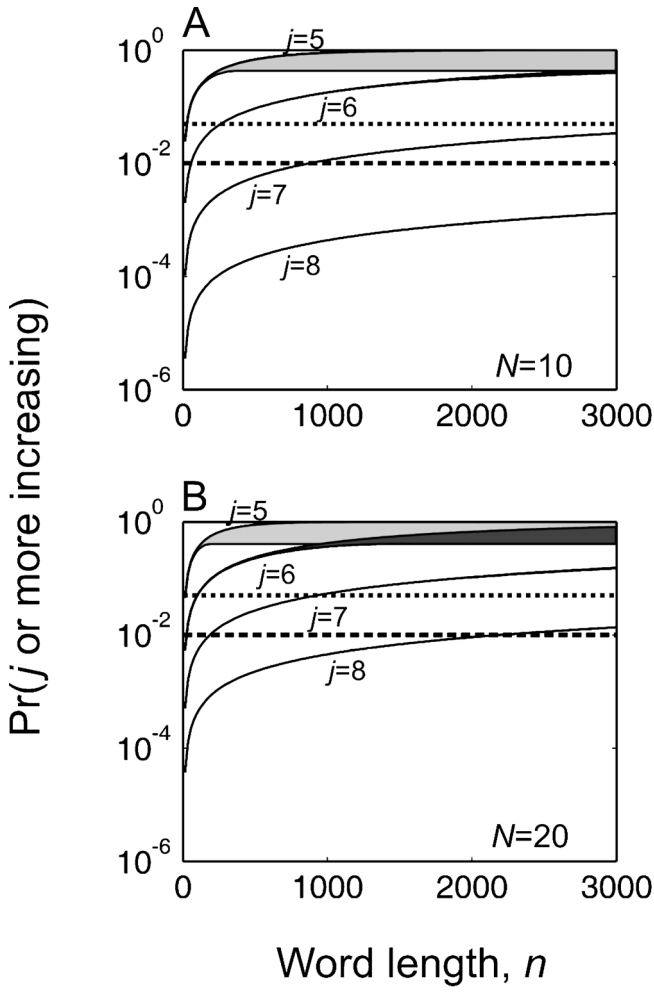

Figure 1: Probability bounds relevant for the analysis of sequences in long experiments. We show the upper and lower bounds of probability of j or more strictly increasing numbers as n increases for four different values of j and two reference sequence sizes (panel A, $N = 10$; panel B, $N = 20$). The region between the upper and lower bounds is shaded gray. As j increases, the difference between the upper and lower bounds becomes indistinguishable. The horizontal dashed lines indicate the standard probability cutoffs of 0.01 and 0.05. For the largest j value of 8, the probability curves lie below 0.05 for both reference sequence lengths and all values of n considered here.

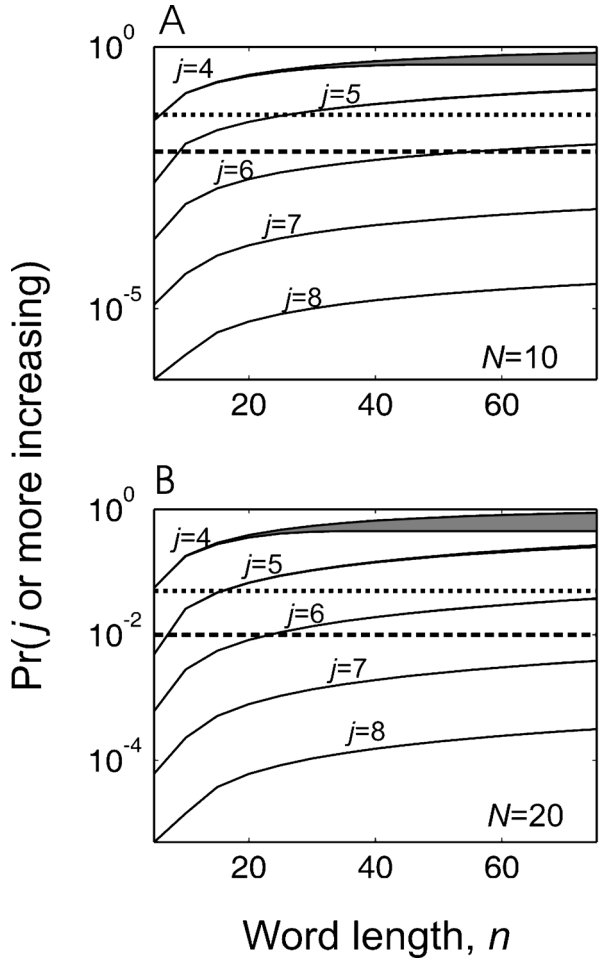

Figure 2: Probability bounds relevant for the analysis of sequences in short experiments. Shown are upper and lower bounds of probability of j or more strictly increasing numbers as n increases for five different values of j and two reference sequence sizes (panel A, $N = 10$; panel B, $N = 20$). The region between the upper and lower bounds is shaded gray. As j increases, the difference between the upper and lower bounds becomes indistinguishable. The black horizontal dashed lines indicate the standard probability cutoffs of 0.01 and 0.05. For the largest shown j value of 8, the probability curves lie below 0.01 for both reference sequence lengths and all values of n considered here.

significant enough to indicate replay at the $p < 0.05$ level. As j increases, this level of significance increases considerably. In contrast, if there are 20 place cells ($N = 20$), an observation of only 5 strictly increasing numbers in a word of length 20 might easily occur by chance (see Figure 2A).

4 Discussion

In this letter, we outline a set probability technique for analyzing spatiotemporal order across neurons. We have derived formulas for the computation of bounds for the probability of a sequence of j strictly increasing elements in a word of length n, chosen from a reference sequence of length N. Our derivation makes use of techniques from elementary probability theory, and calculations can be performed in the order of seconds. While similar problems have been studied in the context of reliability (Wolfowitz, 1944) and in combinatorics (Schensted, 1960), as far as we are aware, the results presented here have not been reported previously.

Some of the inferences about repeating spatiotemporal patterns in neural data have been questioned on statistical grounds (see the points made by Baker & Lemon, 2000; Oram, Wiener, Lestienne, & Richmond, 1999; Moore, Rosenberg, Hary, & Breeze, 1996; Vertes, 2004). Our technique has not been applied to experimental data and therefore does not either prove or disprove any existing experimental conclusions. However, it is interesting to note, for example, that previous analyses of sequences of place cells' order within SWS/ripple events have tended to focus on short words (e.g., j values less than 6). Depending on the parameter space being explored, our analysis indicates that observations of runs of this size are possible by chance alone. In contrast, a single strictly increasing sequence of length 8 is highly unlikely even within a word of length 3000 (see Figure 1) and should be enough for the experimenter to conclude that replay has occurred based on one single observation.

4.1 Comment on Choice of Null Hypothesis. Because we make the assumption that elements from the reference sequence are picked with replacement, our probability estimates will be different from those of a permutation technique. Choosing between with replacement and without replacement techniques is analogous to deciding between the bootstrap technique and a randomization technique, respectively. Our assumption that the word is composed of a number chosen with replacement was made because it is general and allows analytic tractability in solving the problem.

Both combinatorial (Lee & Wilson, 2002, 2004) and template-matching techniques (Abeles & Gerstein, 1988) make use of permutations, or shuffling, to compute the probability of a match. Shuffling techniques are useful for surrogate spike data as they allow the experimenter to preserve firing rate or theta phase relations within the data set (Nádasdy et al., 1999; Gerstein, 2004). This is not so critical in this sequence problem as we assume

the spike data have already been parsed; that is, we assume the data have been smoothed in order that the place cells' firing order during behavior can be represented by numbers 1 though N. A similar smoothing process is carried out on the spikes during sleep, though potentially on a different timescale.

4.2 Advantages of This Approach. An advantage of this approach over shuffling and combinatorial techniques (Nádasdy et al., 1999; Baker & Lemon, 2000; Lee & Wilson, 2004) is that it is straightforward to run through various parameter combinations very quickly. In addition the technique yields analytic forms for upper and lower bounds, and therefore its accuracy does not depend on the number of shuffles or simulations carried out.

Our approach is most similar to the recent combinatorial approach of Lee and Wilson (2002, 2004). They take an observed word, identify the best match within that word, and compute how likely permutations of the sequence would contain that match or better within the word. The final output of their method is a match-trial ratio with corresponding Z-score relative to the chance match-trial ratio ($1/j!$ in our notation). Most importance was assigned in their approach to the low-probability trials involving words of length 4 or more. An advantage of our approach is that both the hypothesis and the calculation of probabilities are considerably simpler.

One of the recommendations from our current analysis would be that even if a combinatorial method is employed later, a more practical approach to identification of statistically significant repeats would be to search for sequences of length 5 or more, rather than 4 or more, since sequences of 4 or more appear frequently by chance alone (see Table 1).

4.3 Limitations of This Approach. Currently this technique computes the probability of at least one sequence of length j or more within a longer word. If we were to observe two sequences of length j or more within the word, corresponding in some cases to the gaps that are used in Lee and Wilson (2004), our current technique would yield a larger estimate of the probability of the event's occurring by chance than necessary. The probability derivation in this case becomes much more difficult (and we leave this for a later publication).

4.4 Future Approaches. As with analysis of behavioral data (Smith et al., 2004), methods from time-series analysis and signal processing may yield more helpful results than techniques based entirely on probability. This is because neurons are inherently noisy and because our current model when applied to rat hippocampal data ignores an important source of data, namely the rat's position, which is available during the awake portion of rat hippocampal experiments. A model that specifically takes into account the

stochastic nature of spike firing and positional information data should be better able to determine if replay of sequences is significant. For example, an area where our method fails is that it does not take into account the extent of place field overlap between cells. If two cells have a large overlap, then it makes sense that it is harder to say which cell fires before the other. This ambiguity should also be taken into account when looking for replay of sequences. Statistical models have been used to describe rat behavior during awake exploration. While they have been applied very successfully in this area (Zhang, Ginzburg, McNaughton, & Sejnowski, 1998; Brown, Frank, Tang, Quirk, & Wilson, 1998), they have not yet been applied to the noisier problem of decoding sleep.

Acknowledgments

This work was supported by the Department of Anesthesiology and Pain Medicine, UC Davis, and NIH grant MH071847.

References

Abeles, M., & Gerstein G. L. (1988). Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. *J. Neurophysiol.*, *60*, 909–924.

Aertsen, A. M. H. J., Gerstein, G. L., Habib, M. K., & Palm G. (1989). Dynamics of neuronal firing correlation: Modulation of "effective connectivity." *J. Neurophysiol.*, *61*, 900–917.

Albert, M. H., Golynski, A., Hamel, A. M., Lopez-Ortiz, A., Rao, S. S., & Safari, M. A. (2004). Longest increasing subsequences in sliding windows. *Theoret. Computer Science*, *321*, 405–414.

August, D. A., & Levy, W. B. (1999). Temporal sequence compression by an integrate and fire model of hippocampal area CA3. *J. Comput. Neurosci.*, *6*, 71–90.

Baker, S. N., & Lemon, R. N. (2000). Precise spatiotemporal repeating patterns in monkey primary and supplementary motor areas occur at chance levels. *J. Neurophysiol.*, *84*, 1770–1780.

Bespamyatnikh, S., & Segal, M. (2000). Enumerating longest increasing subsequences and patience sorting. *Information Proc. Letters*, *76*, 7–11.

Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., & Wilson, M. A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *J. Neurosci.*, *18*, 7411–7425.

Brown, E. N., Kass, R. E., & Mitra, P. P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neurosci.*, *7*(5), 456–461.

Buzsáki, G. (1989). Two-stage model of memory trace formation: A role for noisy brain states. *Neuroscience*, *31*, 551–570.

Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. *Nature Neurosci.*, *7*(5), 446–451.

Chryssaphinou, O., & Vaggelatou, E. (2001). Compound Poisson approximation for long increasing sequences. *J. Appl. Prob.*, *38*, 449–463.

Csicsvari, J., Hirase, H., Czurkó, A., Mamiya, A., & Buzsáki, G. (1999). Oscillatory coupling of hippocampal pyramidal cells and interneurons in the behaving rat. *J. Neurosci., 19*, 274–287.

Dave, A. S., & Margoliash, D. (2000). Song replay during sleep and computational rules for sensorimotor vocal learning. *Science*, *290*, 812–816.

Gerstein, G. L. (2004). Searching for significance in spatio-temporal firing patterns. *Acta Neurobiol. Exp.*, *64*, 203–207.

Gerstein, G. L., & Aersten, A. M. H. J. (1985). Representation of cooperative firing activity among simultaneously recorded neurons. *J. Neurophysiol.*, *54*, 1513–1528.

Grimmett, G. R., & Stirzaker, D. R. (1982). Probability and random processes. Oxford: Clarendon Press.

Grün, S., Diesmann, M., & Aersten, A. (2001). Unitary events in multiple single-neuron spiking activity. I. Detection and significance. *Neural Comp., 14*, 43–80.

Hoffman, K. L., & McNaughton, B. L. (2002). Coordinated reactivation of distributed memory traces in primate neocortex. *Science, 297*(5589), 2070–2073.

Knuth, D. E. (1970). Permutations, matrices and generalized Young tableaux. *Pacific. J. Math., 34*(3), 709–727.

Kudrimoti, H. S., Barnes, C. A., & McNaughton, B. L. (1999). Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and EEG dynamics. *J. Neurosci., 19*, 4090–4101.

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*, 1183–1194.

Lee, A. K., & Wilson, M. A. (2004). A combinatorial method for analyzing sequential firing patterns involving an arbitrary number of neurons based on relative time order. *J. Neurophys., 92*, 2555–2573.

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron, 29*, 145–156.

Moore, G. P., Rosenberg, J. R., Hary, D., & Breeze, P. (1996). "Replay" of hippocampal "memories." *Science, 274*(5290), 1216–1216.

Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., & Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *J. Neurosci., 19*, 9497–9507.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Res., 34*, 171–175.

Oram, M. W., Wiene, M. C., Lestienne, R., & Richmond, B. J. (1999). Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *J. Neurophysiol., 81*, 3021–3033.

Pavlides, C., & Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *J. Neurosci., 9*, 2907–2918.

Ranck, J. B. Jr. (1973). Studies on single neurons in dorsal hippocampal formation and septum in unrestrained rats. I. Behavioral correlates and firing repertories. *Exp. Neurol., 41*, 461–531.

Schensted, C. (1960). Longest increasing and decreasing sequences. *Canad. J. Math., 13*, 179–191.

Skaggs, W. E., & McNaughton, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science, 271*, 1870–1873.

Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A. M., Suzuki, W., & Brown, E. N. (2004). Dynamic analysis of learning in behavioral experiments. *J. Neurosci., 24*(2), 447–461.

Vertes, R. P. (2004). Memory consolidation in sleep: Dream or reality? *Neuron, 44*(1), 135–148.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 265*, 676–679.

Wolfowitz, J. (1944). Asymptotic distribution of runs up and down. *Ann. Math. Statist., 15*, 163–172.

Zhang, K. C., Ginzburg, I., McNaughton, B. L., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *J. Neurophys., 79*(2), 1017–1044.

Received May 20, 2005; accepted September 9, 2005.