Expected Number of Distinct Non-Consecutive Patterns in Random Permutations

Anant Godbole

East Tennessee State University

October 5, 2024

International Conference on Combinatorial Methods and Probability Models. A Conference in Memory of Professor Charalambos Charalambides

Finding the Expected Value of Random Quantities; LCS

- Finding the expected value of random quantities is often non-trivial!

Finding the Expected Value of Random Quantities; LCS

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;

Finding the Expected Value of Random Quantities; LCS

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;
- For example, if we have two binary strings of length n, then it is natural to ask what can be said about the length L_n of their longest common subsequence (LCS).

Finding the Expected Value of Random Quantities; LCS

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;
- For example, if we have two binary strings of length n, then it is natural to ask what can be said about the length L_n of their longest common subsequence (LCS).
- This could be of biological relevance in the case of two DNA strings.

Finding the Expected Value of Random Quantities; LCS

- Finding the expected value of random quantities is often non-trivial!
- This occurs when the quantity in question is unexpectedly nuanced;
- For example, if we have two binary strings of length n, then it is natural to ask what can be said about the length L_n of their longest common subsequence (LCS).
- This could be of biological relevance in the case of two DNA strings.
- Subadditivity arguments are easy to apply to prove that $L = \lim_{n\to\infty} \frac{E(L_n)}{n}$ exists.

More on the LCS Problem

- The value of the limit L, however, is still not known!

More on the LCS Problem

- The value of the limit L, however, is still not known!
- The best known bounds are, roughly, $0.78 \leq L \leq 0.82$.

More on the LCS Problem

- The value of the limit L, however, is still not known!
- The best known bounds are, roughly, $0.78 \leq L \leq 0.82$.
- The variance is of order n and in 2014, Houdré proved a CLT.

LIS

- More is known about the length of the longest increasing subsequence (LIS) of a random permutation, the study of which culminated in the celebrated paper of Baik, Deift, and Johansson (1998).

LIS

- More is known about the length of the longest increasing subsequence (LIS) of a random permutation, the study of which culminated in the celebrated paper of Baik, Deift, and Johansson (1998).
- But even here, calculation of the expected value was non-trivial.

LIS

- More is known about the length of the longest increasing subsequence (LIS) of a random permutation, the study of which culminated in the celebrated paper of Baik, Deift, and Johansson (1998).
- But even here, calculation of the expected value was non-trivial.
- The combined results of Vershik and Kerov; and Logan and Shepp from the 1970's gave

$$\lim \frac{EL_n}{\sqrt{n}} = 2.$$

Tracy Widom Distribution

This was followed by concentration results—due to Bollobas and Janson; Kim; and Frieze among others—that revealed that the standard deviation of the size of the longest monotone subsequence (LMS) is of order $\Theta(n^{1/6})$,

Tracy Widom Distribution

This was followed by concentration results—due to Bollobas and Janson; Kim; and Frieze among others—that revealed that the standard deviation of the size of the longest monotone subsequence (LMS) is of order $\Theta(n^{1/6})$,

This culminated with the work of Baik, Deift and Johansson (cited earlier) that exhibited the limiting law of a normalized version of the LMS (Tracy Widom distribution).

Tracy Widom Distribution

This was followed by concentration results—due to Bollobas and Janson; Kim; and Frieze among others—that revealed that the standard deviation of the size of the longest monotone subsequence (LMS) is of order $\Theta(n^{1/6})$,

This culminated with the work of Baik, Deift and Johansson (cited earlier) that exhibited the limiting law of a normalized version of the LMS (Tracy Widom distribution).

This is often cited as one of the crowning achievements of Probability/Analysis of the 20th Century. An *AMS Notices* article of Aldous and Diaconis gives a great summary.

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.
- First, we consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.
- First, we consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.
- For example the string 11111 has 5 subsequences, namely 1, 11, 111, 1111, and 11111, whereas

The Two Examples and our Problem

- The above two examples are of two problems about which a lot is known after a slow start.
- First, we consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.
- For example the string 11111 has 5 subsequences, namely 1, 11, 111, 1111, and 11111, whereas
- The string 10110 contains the subsequences 0, 1, 01, 10, 11, 00, 100, 101, 110, 111, 011, 010, 1011, 1010, 1110, 0110, and 10110.

- The above two examples are of two problems about which a lot is known after a slow start.
- First, we consider a random binary string and ask how many subsequences are embedded in it. We will make the slow start.
- For example the string 11111 has 5 subsequences, namely 1, 11, 111, 1111, and 11111, whereas
- The string 10110 contains the subsequences 0, 1, 01, 10, 11, 00, 100, 101, 110, 111, 011, 010, 1011, 1010, 1110, 0110, and 10110.
- What is the average case behavior?

Collaborators

This talk about joint work with

- Biers-Ariel and Kelley (*DMTCS*);

Collaborators

This talk about joint work with

- Biers-Ariel and Kelley (*DMTCS*);
- Allen, Cruz-Fonseca, Dobbs, Downs, Fokuoh, Papanikolaou, Soto, and Yoshikawa (the so-called9PP) (*PuMA*);

Collaborators

This talk about joint work with

- Biers-Ariel and Kelley (*DMTCS*);
- Allen, Cruz-Fonseca, Dobbs, Downs, Fokuoh, Papanikolaou, Soto, and Yoshikawa (the so-called9PP) (*PuMA*);
- Swickheimer (*DMTCS 2024*); and

Collaborators

This talk about joint work with

- Biers-Ariel and Kelley (*DMTCS*);
- Allen, Cruz-Fonseca, Dobbs, Downs, Fokuoh, Papanikolaou, Soto, and Yoshikawa (the so-called9PP) (*PuMA*);
- Swickheimer (*DMTCS 2024*); and
- Borras-Serrano, Byrne, Jackson, LeBlanc and Veimau (preprint).

Existence

In the DMTCS paper, we proved

Theorem

Let $s_1, s_2, \ldots$ be a sequence of independent and identically distributed random variables with $Pr(s_1 = j) = \alpha_j, j = 1, 2, \ldots, d, \sum_j \alpha_j = 1$. Set $\alpha = (\alpha_1, \ldots, \alpha_d)$. Let $\phi(S_n)$ be the number of distinct subsequences in $S_n = (s_1, \ldots, s_n)$. Let $\psi(n) = E(\phi(S_n))$. Then there exists $c = c_{d,\alpha} \geq 1$ such that

$$\psi(n)^{1/n} \to c; n \to \infty,$$

where $c = 1$ iff $d \geq 1$ and $\max_j \alpha_j = 1$.

Discussion

- The above theorem is hardly surprising, but raises other questions, namely whether the "true" numbers contain, additionally, polynomial factors as do several Stanley-Wilf limits in the theory of pattern avoidance (note that there are no polynomial factors in our next result with $d = 2$) Also, in general the existence of limits is not automatic, as seen by the following example:

Discussion

- The above theorem is hardly surprising, but raises other questions, namely whether the "true" numbers contain, additionally, polynomial factors as do several Stanley-Wilf limits in the theory of pattern avoidance (note that there are no polynomial factors in our next result with $d = 2$) Also, in general the existence of limits is not automatic, as seen by the following example:
- Assume that n balls are independently thrown into an infinite array of boxes so that box j is hit with probability $1/2^j$ for $j = 1, 2, \ldots$. Let π_n be the probability that the largest occupied box has a single ball in it. Then, as proved by several people in the 1990's, $\lim_{n\to\infty} \pi_n$ does not exist, and $\limsup_{n\to\infty} \pi_n$ and $\liminf_{n\to\infty} \pi_n$ differ in the fourth decimal place! Such behavior does not however occur in our context, as the theorem states.

The case of $d = 2$

Theorem

Suppose $Pr[s_i = 1] = \alpha \in [0, 1]$ *for all* $1 \leq i \leq n$, *and* $Pr[s_i = 0] = 1 - \alpha$, $\alpha \neq 0, 1$. *Then we have*

$$\psi(S_n) = \frac{A + B}{2\sqrt{\alpha(1-\alpha)}},$$

where

$$A = \left(1 - 2\sqrt{\alpha(1-\alpha)}\right)\left(1 - \left(1 - \sqrt{\alpha(1-\alpha)}\right)^n\right)$$

and

$$B = \left(1 + 2\sqrt{\alpha(1-\alpha)}\right)\left(\left(1 + \sqrt{\alpha(1-\alpha)}\right)^n - 1\right)$$

Result was Previously Known for $\alpha = 0.5$

It was shown in a 2004 EJC paper of Flaxman et al.that when $\Pr[s_i = 1] = .5$ then $E[\phi(S_n)] \sim k(\frac{3}{2})^n$ for a constant k. Later, Collins improved this result by finding that $E[\phi(S_n)] = 2(\frac{3}{2})^n - 1$. We generalized this in the previous theorem to non-uniform letter generation. Two state Markov chains were also considered.

Permutations

How do these questions translate to *permutations*? What is the number of distinct *patterns* contained (*consecutively or non-consecutively*) in a random permutation on $[n]$?

Permutations

How do these questions translate to *permutations*? What is the number of distinct *patterns* contained (*consecutively or non-consecutively*) in a random permutation on $[n]$?

•The permutation 34152 consecutively contains the patterns 1, 12, 21, 231, 213, 132, 2314, 3142, and 34152. This is the maximum possible.

Permutations

How do these questions translate to *permutations*? What is the number of distinct *patterns* contained (*consecutively or non-consecutively*) in a random permutation on $[n]$?

•The permutation 34152 consecutively contains the patterns 1, 12, 21, 231, 213, 132, 2314, 3142, and 34152. This is the maximum possible.

•The permutation 34152 non-consecutively contains the patterns 1,12, 21, 231, 213, 132, 123, 312, 2314, 3142, 2341, 3412, and 34152. This is NOT the maximum possible.

Permutations

How do these questions translate to *permutations*? What is the number of distinct *patterns* contained (*consecutively or non-consecutively*) in a random permutation on $[n]$?

•The permutation 34152 consecutively contains the patterns 1, 12, 21, 231, 213, 132, 2314, 3142, and 34152. This is the maximum possible.

•The permutation 34152 non-consecutively contains the patterns 1,12, 21, 231, 213, 132, 123, 312, 2314, 3142, 2341, 3412, and 34152. This is NOT the maximum possible.

•The minimum number of consecutive or non-consecutive patterns is n, as given by the permutation $123\dots n$.

Permutations

How do these questions translate to *permutations*? What is the number of distinct *patterns* contained (*consecutively or non-consecutively*) in a random permutation on $[n]$?

•The permutation 34152 consecutively contains the patterns 1, 12, 21, 231, 213, 132, 2314, 3142, and 34152. This is the maximum possible.

•The permutation 34152 non-consecutively contains the patterns 1,12, 21, 231, 213, 132, 123, 312, 2314, 3142, 2341, 3412, and 34152. This is NOT the maximum possible.

•The minimum number of consecutive or non-consecutive patterns is n, as given by the permutation $123\ldots n$.

•Alison Miller (2009) proved, answering a question by Wilf from 2003, and improving previous results due to e.g., Coleman (2004), Albert et al (2007), that

$$2^n - O(n^2 2^{n-\sqrt{2n}}) \le \max_{\pi_n \in S_n} \phi(\pi_n) \le 2^n - \Theta(n 2^{n-\sqrt{2n}}).$$

Consecutive Case

- In the consecutive case it is easy to see that the maximum number of possible patterns is $\sum_{k=1}^{n} \min\{k!, (n-k+1)\}$. This bound *can* be attained for $1 \leq n \leq 12$ (data not provided).

Consecutive Case

- In the consecutive case it is easy to see that the maximum number of possible patterns is $\sum_{k=1}^{n} \min\{k!, (n-k+1)\}$. This bound *can* be attained for $1 \leq n \leq 12$ (data not provided).
- BUT, more importantly, can the *expected* value of X, the number of distinct subpatterns, be close to the maximum value $\sum_{1 \leq k \leq n} \min\{k!, (n-k+1)\} = \frac{n^2}{2}(1-o(1))$ as it does for $n \leq 12$ (data not provided)?

Results for Consecutive Case

With $X = X_n$ denoting the number of distinct consecutive patterns in a random permutation

Theorem
(9PP)

$$\mathbb{E}(X_n) \geq \frac{n^2}{2}\left(1 - 200\frac{\ln n}{n}\right).$$

Theorem
(Swickheimer and G)

$$\mathbb{E}(X) \geq \frac{n^2}{2}\left(1 - 17\frac{\ln n}{n}\right).$$

X, Y, Z

The difference between these two $200 \ln n$ and $17 \ln n$ papers is not just cosmetic "improved analysis".

X, Y, Z

The difference between these two $200 \ln n$ and $17 \ln n$ papers is not just cosmetic "improved analysis". It is a matter of technique.

X, Y, Z

The difference between these two $200 \ln n$ and $17 \ln n$ papers is not just cosmetic "improved analysis". It is a matter of technique. In the first paper, the authors didn't quite know how to work with X and so they introduced Y and Z.

X, Y, Z

The difference between these two $200 \ln n$ and $17 \ln n$ papers is not just cosmetic "improved analysis". It is a matter of technique. In the first paper, the authors didn't quite know how to work with X and so they introduced Y and Z.
*$Y = \sum_k Y_k =$ the number of repeated patterns, so that if 132 occurs 4 times, its contribution to Y_3 is 3.

X, Y, Z

The difference between these two $200 \ln n$ and $17 \ln n$ papers is not just cosmetic "improved analysis". It is a matter of technique. In the first paper, the authors didn't quite know how to work with X and so they introduced Y and Z.
*$Y = \sum_k Y_k =$ the number of repeated patterns, so that if 132 occurs 4 times, its contribution to Y_3 is 3.
*$Z = Z_k$ equals the number of pairs of isomorphic patterns, so that the contribution of 132 to $Z_3 = \binom{4}{2} = 6$ in the above example.

X, Y, Z

The difference between these two $200 \ln n$ and $17 \ln n$ papers is not just cosmetic "improved analysis". It is a matter of technique. In the first paper, the authors didn't quite know how to work with X and so they introduced Y and Z.

*$Y = \sum_k Y_k =$ the number of repeated patterns, so that if 132 occurs 4 times, its contribution to Y_3 is 3.

*$Z = Z_k$ equals the number of pairs of isomorphic patterns, so that the contribution of 132 to $Z_3 = \binom{4}{2} = 6$ in the above example.

*Clearly, $X_k = \binom{n}{k} - Y_k \leq \binom{n}{k} - Z_k$.

X, Y, Z

The difference between these two $200\ln n$ and $17\ln n$ papers is not just cosmetic "improved analysis". It is a matter of technique. In the first paper, the authors didn't quite know how to work with X and so they introduced Y and Z.

*$Y = \sum_k Y_k$= the number of repeated patterns, so that if 132 occurs 4 times, its contribution to Y_3 is 3.

*$Z = Z_k$ equals the number of pairs of isomorphic patterns, so that the contribution of 132 to $Z_3 = \binom{4}{2} = 6$ in the above example.

*Clearly, $X_k = \binom{n}{k} - Y_k \leq \binom{n}{k} - Z_k$.

However in the second paper, the authors recognized that a pattern was distinct if it occurred at least once; thus

X, Y, Z

The difference between these two $200 \ln n$ and $17 \ln n$ papers is not just cosmetic "improved analysis". It is a matter of technique. In the first paper, the authors didn't quite know how to work with X and so they introduced Y and Z.
*$Y = \sum_k Y_k =$ the number of repeated patterns, so that if 132 occurs 4 times, its contribution to Y_3 is 3.
*$Z = Z_k$ equals the number of pairs of isomorphic patterns, so that the contribution of 132 to $Z_3 = \binom{4}{2} = 6$ in the above example.
*Clearly, $X_k = \binom{n}{k} - Y_k \le \binom{n}{k} - Z_k$.
However in the second paper, the authors recognized that a pattern was distinct if it occurred at least once; thus

$$\mathbb{E}(X_k) = \sum_j \mathbb{P}(N_j \ge 1)$$

where N_j is the number of occurrences of the jth pattern of length k.

DISTRIBUTIONS

- In the consecutive case, Hannah Swickheimer showed that $L(N_j) \approx \mathrm{Po}(\lambda)$, where $L(\cdot)$ is the distribution of $\cdot$ and $\mathrm{Po}(\lambda)$ denotes the Poisson r.v. with parameter $\lambda = \frac{n-k+1}{k!}$, which is the expected number of consecutive occurrences of any pattern of length k.

DISTRIBUTIONS

- In the consecutive case, Hannah Swickheimer showed that $L(N_j) \approx \mathrm{Po}(\lambda)$, where $L(\cdot)$ is the distribution of $\cdot$ and $\mathrm{Po}(\lambda)$ denotes the Poisson r.v. with parameter $\lambda = \frac{n-k+1}{k!}$, which is the expected number of consecutive occurrences of any pattern of length k.
- In the non-consecutive case, the distribution of X appears to be quite hard and certainly worth further investigation à la Tracy-Widom etc. Here $\lambda = \frac{\binom{n}{k}}{k!}$.

DISTRIBUTIONS

- In the consecutive case, Hannah Swickheimer showed that $L(N_j) \approx \mathrm{Po}(\lambda)$, where $L(\cdot)$ is the distribution of $\cdot$ and $\mathrm{Po}(\lambda)$ denotes the Poisson r.v. with parameter $\lambda = \frac{n-k+1}{k!}$, which is the expected number of consecutive occurrences of any pattern of length k.
- In the non-consecutive case, the distribution of X appears to be quite hard and certainly worth further investigation à la Tracy-Widom etc. Here $\lambda = \frac{\binom{n}{k}}{k!}$.
- Our work in this area using CLTs, martingale methods such as Azuma's inequality, etc have borne no fruit. In a nutshell,

DISTRIBUTIONS

- In the consecutive case, Hannah Swickheimer showed that $L(N_j) \approx \mathrm{Po}(\lambda)$, where $L(\cdot)$ is the distribution of $\cdot$ and $\mathrm{Po}(\lambda)$ denotes the Poisson r.v. with parameter $\lambda = \frac{n-k+1}{k!}$, which is the expected number of consecutive occurrences of any pattern of length k.
- In the non-consecutive case, the distribution of X appears to be quite hard and certainly worth further investigation à la Tracy-Widom etc. Here $\lambda = \frac{\binom{n}{k}}{k!}$.
- Our work in this area using CLTs, martingale methods such as Azuma's inequality, etc have borne no fruit. In a nutshell,
- The dependencies amongst the summands in

$$N_k = \sum_{r \in \binom{n}{k}} I_r$$

where I_r is one if the rth k-pattern occurs in non-consecutive positions, are too extreme.

Opera

- This is a hard problem, work on this is in progress;

Opera

- This is a hard problem, work on this is in progress;
- At the current time, this project seems like a bad opera with few conclusions and a handful of results (two to be specific). It is truly "all over the place";

Opera

- This is a hard problem, work on this is in progress;
- At the current time, this project seems like a bad opera with few conclusions and a handful of results (two to be specific). It is truly "all over the place";
- The thought is to submit at the end of the Fall after proving at least one new result from among the directions that will be mentioned in the rest of the talk;

Opera

- This is a hard problem, work on this is in progress;
- At the current time, this project seems like a bad opera with few conclusions and a handful of results (two to be specific). It is truly "all over the place";
- The thought is to submit at the end of the Fall after proving at least one new result from among the directions that will be mentioned in the rest of the talk;
- There are two strategies, to continue with the X, Y, Z trifecta in the non-consecutive case; and to invoke the theory of subadditivity;

Opera

- This is a hard problem, work on this is in progress;
- At the current time, this project seems like a bad opera with few conclusions and a handful of results (two to be specific). It is truly "all over the place";
- The thought is to submit at the end of the Fall after proving at least one new result from among the directions that will be mentioned in the rest of the talk;
- There are two strategies, to continue with the X, Y, Z trifecta in the non-consecutive case; and to invoke the theory of subadditivity;
- A perfect long term agenda would be to find the mean, variance, distribution, tightness of concentration etc.

The Non-Consecutive Case, Summary of Results

- Can the expected value of X be close to the maximum, i.e., 2^n, as determined by Miller? (Recall that $\mathbb{E}(X) \sim \max X$ is true in the consecutive case).

The Non-Consecutive Case, Summary of Results

- Can the expected value of X be close to the maximum, i.e., 2^n, as determined by Miller? (Recall that $\mathbb{E}(X) \sim \max X$ is true in the consecutive case).
- Here is a summary of results, which we will address for the rest of the talk:

The Non-Consecutive Case, Summary of Results

- Can the expected value of X be close to the maximum, i.e., 2^n, as determined by Miller? (Recall that $\mathbb{E}(X) \sim \max X$ is true in the consecutive case).
- Here is a summary of results, which we will address for the rest of the talk:
- (Jackson and LeBlanc) have shown that $\mathbb{E}(X) \sim c^n$ for some $1 < c \leq 2$.

The Non-Consecutive Case, Summary of Results

- Can the expected value of X be close to the maximum, i.e., 2^n, as determined by Miller? (Recall that $\mathbb{E}(X) \sim \max X$ is true in the consecutive case).
- Here is a summary of results, which we will address for the rest of the talk:
- (Jackson and LeBlanc) have shown that $\mathbb{E}(X) \sim c^n$ for some $1 < c \leq 2$.
- We are close, using methods of Borras-Serrano, Byrne and Veimau, to proving that $c = 2$, failing which we will try to show that

The Non-Consecutive Case, Summary of Results

- Can the expected value of X be close to the maximum, i.e., 2^n, as determined by Miller? (Recall that $\mathbb{E}(X) \sim \max X$ is true in the consecutive case).
- Here is a summary of results, which we will address for the rest of the talk:
- (Jackson and LeBlanc) have shown that $\mathbb{E}(X) \sim c^n$ for some $1 < c \leq 2$.
- We are close, using methods of Borras-Serrano, Byrne and Veimau, to proving that $c = 2$, failing which we will try to show that
- $c \geq c_0$, perhaps $c_0 = 1.73$ (Jackson and LeBlanc)

The Covariance

Recall Z_k counts the number of *pairs* of isomorphic patterns. Thus

$$\mathbb{E}(Z_k) = \sum_{\eta_1} \sum_{\eta_2} P(\pi_1 \simeq \pi_2),$$

where η_1 and η_2 are two sets of k positions and $\pi_1 \simeq \pi_2$ if the patterns in these positions are isomorphic.

The Covariance

Recall Z_k counts the number of *pairs* of isomorphic patterns. Thus

$$\mathbb{E}(Z_k) = \sum_{\eta_1} \sum_{\eta_2} P(\pi_1 \simeq \pi_2),$$

where η_1 and η_2 are two sets of k positions and $\pi_1 \simeq \pi_2$ if the patterns in these positions are isomorphic.

This agenda is needed if we are to employ the X, Y, Z trifecta in the non-consecutive case (Borras-Serrano et al.)

The Covariance

Recall Z_k counts the number of *pairs* of isomorphic patterns. Thus

$$\mathbb{E}(Z_k) = \sum_{\eta_1} \sum_{\eta_2} P(\pi_1 \simeq \pi_2),$$

where η_1 and η_2 are two sets of k positions and $\pi_1 \simeq \pi_2$ if the patterns in these positions are isomorphic.

This agenda is needed if we are to employ the X, Y, Z trifecta in the non-consecutive case (Borras-Serrano et al.)

Clearly for $P(\pi_1 \simeq \pi_2) > 0$, we must have the overlap positions in π_1, π_2 to be isomorphic.

Bounding the Covariance

Lemma

The probability that two sets of k-positions that overlap in r specific spots contain isomorphic patterns satisfies

$$\mathbb{P}(\pi_1 \simeq \pi_2) \leq \frac{\binom{k}{r}^2 r!(k-r)!2^{2k-2r}}{(2k-r)!}.$$

Bounding the Covariance

Lemma

The probability that two sets of k-positions that overlap in r specific spots contain isomorphic patterns satisfies

$$\mathbb{P}(\pi_1 \simeq \pi_2) \leq \frac{\binom{k}{r}^2 r!(k-r)!2^{2k-2r}}{(2k-r)!}.$$

The lengthy proof of the lemma consists of bounding the number of ways in which we can assign $2k - r$ numbers to π_1 and π_2, so that $\pi_1 \simeq \pi_2$. However, it sweeps under the rug the fact that two patterns that are isomorphic in their overlaps need not be isomorphic in their totality due to a poor alignment. This is a flaw!

The lemma fails to deliver

Unfortunately this key lemma gave up too much and does not prove to be useful to bound $\mathbb{E}(X)$ as in the consecutive case. What occurs is that for k's around $n/2$, we get $\sum \mathbb{E}(Z_k) = 2^{n(1+o(1))}$ rather than $\sum \mathbb{E}(Z_k) = 2^{n(1-o(1))}$. However,

Pairs of non-isomorphic patterns

Theorem

The expected number $\mathbb{E}(\Delta_n)$ of pairs of non-isomorphic patterns of all lengths is at least

$$C \cdot \frac{2^{2n}}{\sqrt{n}} \left(1 - \frac{3}{n^2}\right).$$

Pairs of non-isomorphic patterns

Theorem

The expected number $\mathbb{E}(\Delta_n)$ of pairs of non-isomorphic patterns of all lengths is at least

$$C \cdot \frac{2^{2n}}{\sqrt{n}} \left(1 - \frac{3}{n^2}\right).$$

If we knew, however, e.g., that *most* pairs of non-isomorphic patterns were obtained by comparing two patterns from among those in the list of distinct patterns, then, we'd have $X_k \sim \sqrt{\Delta_k}$, and we'd be closer to our goal. However the theorem above is useful in its own right.

Subadditivity and Fekete's lemma

A real sequence is subadditive if

$$a_{n+m} \leq a_n + a_m$$

.

FEKETE'S LEMMA: If a sequence is subadditive then

$$\lim \frac{a_n}{n} = \inf \frac{a_n}{n}$$

exists in $[-\infty, \infty)$

Subadditivity and Fekete's lemma

A real sequence is subadditive if

$$a_{n+m} \leq a_n + a_m$$

.

FEKETE'S LEMMA: If a sequence is subadditive then

$$\lim \frac{a_n}{n} = \inf \frac{a_n}{n}$$

exists in $[-\infty, \infty)$

If it were the case that

$$\mathbb{E}(X_{n+m}) \geq \mathbb{E}(X_{1,\ldots,n}) \cdot \mathbb{E}(X_{n+1,\ldots,n+m}),$$

Continued...

Then we'd have

$$-\log_2 \mathbb{E}(X_{1,\ldots,n+m}) \le -\log_2 \mathbb{E}(X_{1,\ldots,n}) + (-\log_2 \mathbb{E}(X_{n+1,\ldots,n+m}))$$

Continued...

Then we'd have

$$-\log_2 \mathbb{E}(X_{1,\dots,n+m}) \le -\log_2 \mathbb{E}(X_{1,\dots,n}) + (-\log_2 \mathbb{E}(X_{n+1,\dots,n+m}))$$

- or, $\frac{-\log_2 \mathbb{E}(X_n)}{n} \to \ell$ (by Fekete)
- i.e., $\mathbb{E}^{\frac{1}{n}}(X_n) \to 2^{-\ell} := c$,

Our data shows that $\mathbb{E}(X_n)^{1/n}$ increases as 1, 1.414, 1.542, 1.592, 1.624, 1.650, 1.672, 1.693, 1.713, 1.730 till $n = 10$

so that a block argument would give $\mathbb{E}(X_n) \sim c^n$ for $c \in [1.73, 2]$.

Continued...

Then we'd have

$$-\log_2 \mathbb{E}(X_{1,\dots,n+m}) \leq -\log_2 \mathbb{E}(X_{1,\dots,n}) + (-\log_2 \mathbb{E}(X_{n+1,\dots,n+m}))$$

- or, $\frac{-\log_2 \mathbb{E}(X_n)}{n} \to \ell$ (by Fekete)
- i.e., $\mathbb{E}^{\frac{1}{n}}(X_n) \to 2^{-\ell} := c$,
- or $\mathbb{E}(X_n) \sim c^n$

Our data shows that $\mathbb{E}(X_n)^{1/n}$ increases as 1, 1.414, 1.542, 1.592, 1.624, 1.650, 1.672, 1.693, 1.713, 1.730 till $n = 10$

so that a block argument would give $\mathbb{E}(X_n) \sim c^n$ for $c \in [1.73, 2]$.

Unfortunately subadditivity does not hold, but we have instead

Unfortunately subadditivity does not hold, but we have instead

Theorem

$$\mathbb{E}(X_n) \cdot \mathbb{E}(X_{n+1}, \ldots, X_{n+m}) \leq (n+m)\mathbb{E}(X_{n+m})$$

Unfortunately subadditivity does not hold, but we have instead

Theorem

$$\mathbb{E}(X_n) \cdot \mathbb{E}(X_{n+1}, \ldots, X_{n+m}) \leq (n+m)\mathbb{E}(X_{n+m})$$

This helps due to the Theorem of Erdős and deBruijn which generalizes Fekete as follows

Unfortunately subadditivity does not hold, but we have instead

Theorem

$$\mathbb{E}(X_n) \cdot \mathbb{E}(X_{n+1}, \dots, X_{n+m}) \leq (n+m)\mathbb{E}(X_{n+m})$$

This helps due to the Theorem of Erdős and deBruijn which generalizes Fekete as follows

Theorem

(DeBruijn-Erdős) Let $\phi(t)$ be positive and increasing for $t > 0$, and assume

$$\int_1^\infty \phi(t) t^{-2} dt < \infty$$

Then,
if the sequence a_n satisfies $a_{n+m} \leq a_n + a_m + \phi(n+m)$ for $\frac{1}{2}n \leq m \leq 2n$, then $\frac{a_n}{n} \to L$ for $L \in [-\infty, \infty)$

Consequences of Near-Subadditivity

$$\mathbb{E}(X_{n+m}) \geq \frac{\mathbb{E}(X_n)\mathbb{E}(X_m)}{n+m},$$

so that (all logs are to base 2)

$$\log \mathbb{E}(X_{n+m}) \geq \log(\mathbb{E}(X_n) + \log(\mathbb{E}(X_m)) - \log(n+m),$$

or

$$-\log(\mathbb{E}(X_{n+m}) \leq -\log(\mathbb{E}(X_n) - \log(\mathbb{E}(X_m) + \log(n+m),$$

which shows that $-\log \mathbb{E}(X_n)$ is near-subadditive. Fekete's lemma (in its improved Erdős-DeBruijn form with $\phi(n) = \log n$ yields that

$$-\frac{\log \mathbb{E}(X_n)}{n} \to c.$$

Since

$$1 \leq \mathbb{E}(X_n) \leq 2^n,$$

we must have

$$-1 \leq c \leq 0.$$

Consequences Continued....

Thus

$$\log \frac{1}{\mathbb{E}(X_n)^{1/n}} \to c,$$

or

$$\mathbb{E}(X_n)^{1/n} \to 2^{-c},$$

which proves that

$$\mathbb{E}(X_n) \sim 2^{-nc} = 2^{dn},$$

where

$$0 \leq d \leq 1.$$

Theorem

We thus have

Theorem

$$\mathbb{E}(X_n)^{1/n} \to c \in (1, 2].$$

In other words no oscillatory behavior is possible, even in the 4th decimal place!

Issues and Continuing Work

- Is $\mathbb{E}(X_n)^{1/n}$ monotone in n? This would give $c \geq 1.73$.

Issues and Continuing Work

- Is $\mathbb{E}(X_n)^{1/n}$ monotone in n? This would give $c \geq 1.73$.
- Can a weaker version of subadditivity, not quite as strong as

$$\mathbb{E}(X_n) \cdot \mathbb{E}(X_{n+1}, \ldots, X_{n+m}) \leq (n+m)\mathbb{E}(X_{n+m})$$

be proved and still yield the conclusion that

$$\mathbb{E}(X_n) \geq (1.73)^n?$$

We know that

$$\mathbb{E}(X_{n+m}) \geq \frac{\mathbb{E}(X_n)\mathbb{E}(X_m)}{n+m},$$

Similarly, we can prove that

$$\mathbb{E}(X_{n+m}) \leq \binom{n+m}{m} \frac{\mathbb{E}(X_n)\mathbb{E}(X_m)}{n+m},$$

so that

$$\frac{\mathbb{E}(X_n)\mathbb{E}(X_m)}{n+m} \leq \mathbb{E}(X_{n+m}) \leq \binom{n+m}{m} \frac{\mathbb{E}(X_n)\mathbb{E}(X_m)}{n+m},$$

However this is too weak to give two sided estimates that allow rates of convergence in Erdős-deBruijn (Steele, Hammersley) to be applied to yield 1.73 or better. Can the exponential factor $\binom{n+m}{m}$ be improved?.

Other tries have included

- The use of Kleitman's lemma and other tools from lattice theory/the theory of correlation inequalities;

Other tries have included

- The use of Kleitman's lemma and other tools from lattice theory/the theory of correlation inequalities;
- ignoring and enumerating "inconvenient permutations", i.e., those for which are those for which $X_{n+m} \geq X_{1,\ldots,n} X_{n+1,\ldots,n+m}$ does not hold. Examples include the two monotone permutations and $123 \cdots (n-2)(n)(n-1)$. Hopefully these exceptions will still allow $\mathbb{E}(X_{n+m}) \geq \mathbb{E}(X_{1,\ldots,n})\mathbb{E}(X_{n+1,\ldots,n+m})$.

Quotes

(Steele) "The determination of the limiting constant is often difficult. In fact, there are fewer than a handful of cases where we are able to calculate the limiting constant obtained by a subadditivity argument; even good approximations of the constants often require considerable ingenuity."

(Steele) "By and large subadditivity offers only elementary tools, but on remarkably many occasions such tools provide the key organizing principle in the attack on problems of nearly intractable difficulty."