A General Purpose Algorithm for Counting Permutations with Consecutive Sequences

Thomas Becker

March 13, 2017

Abstract

1 Introduction

Notation 1.1 If n is an integer with $n \geq 1$, we write $\mathcal{P}_n$ for the set of all permutations of the integers $1, 2, \ldots, n$.

Definition 1.2 Let $P \in \mathcal{P}_n$, and let $k \geq 2$. A **consecutive sequence** of length k in P is a contiguous subsequence of k consecutive integers in p, that is, a contiguous subsequence of the form $(i, i+1, \ldots, i+k-1)$. A consecutive sequence is called **maximal** if it is not a subsequence of a consecutive sequence of greater length.

There is a vast body of work regarding the count of permutations that have a specified configuration of consecutive sequences, such as permutations having a certain number of consecutive pairs or triples [1]. In this paper, we state, prove the correctness of, and discuss the complexity of two general purpose algorithms for counting permutations whose maximal consecutive sequences are described in certain ways. The need for these algorithms, which are available under the MIT license [**?**], arose from the author's curiosity about the behavior of random shuffle mode on today's music players [**?**].

In Theorem 2.5, we give an auxiliary algorithm on which the two main algorithms are based. It calculates the number of permutations that satisfy a specification of maximal consecutive sequences by initial elements and lengths.

Theorem 3.3 provides an algorithm to count the permutations whose maximal consecutive sequences have been specified by stating how many exactly of each length there should be. Using that algorithm, one may, for example, calculate the number of permutations that have exactly three consecutive pairs, none of which are joined to form a consecutive triple (specify "three maximal consecutive sequences of length two, zero maximal consecutive sequences of any other length").

Building on top of that, Theorem 3.3 describes a generic, customizable algorithm that iterates over all possible configurations of maximal consecutive sequences by length and count. A user-supplied function decides if permutations with a given configuration should be included in the count or not. For example, the user-supplied function could accept any configuration that specifies a non-zero count for maximal sequences of length two and a zero count for all other lengths, and reject all others configurations. The result would be the number of permutations that have any number of consecutive pairs, but no adjacent pairs that form larger consecutive sequences.

2 Specifying Consecutive Sequences by Lengths And Initial Elements

Definition 2.1 Let n be an integer with $n \geq 1$. An **MCS-specification by lengths and initial elements** for n is a set of pairs of integers

$$\{ (a_1, k_1), (a_2, k_2), \ldots, (a_m, k_m) \}$$

with the following properties:

(i) $a_i \geq 1$ and $k_i \geq 2$ for $1 \leq i \leq m$,

(ii) $a_i + k_i \leq a_{i+1}$ for $1 \leq i \leq m-1$,

(iii) $a_m + (k_m - 1) \leq n$.

Notation 2.2 If S is an MCS-specification by lenghts and initial elements for n as in Definition 2.1 above, we write $\mathcal{Q}_{(n,S)}$ for the set of all permutations $P \in \mathcal{P}_n$ with the following property: for each i with $1 \leq i \leq m$, P has a maximal consecutive sequence of length k_i that starts with the integer a_i, and P has no other maximal consecutive sequences.

Purpose of this Section Present an auxiliary algorithm, to be used in later sections, for calculating $|\mathcal{Q}_{(n,S)}|$ from n and S.

The following technical lemma will be needed when we use induction on m in connection with MCS-specifications by lengths and initial elements.

Lemma 2.3 Let S be an MCS-specification by lengths and initial elements for n as in Definition 2.1 above, and assume that $m \geq 1$. Then $n - (k_m - 1) \geq 1$, and

$$T = \{ (a_1, k_1), (a_2, k_2), \ldots, (a_{m-1}, k_{m-1}) \}$$

is an MCS-specification by lengths and initial elements for $n - (k_m - 1)$.

Proof From (i) and (iii) of Definition 2.1, we may conclude that

$$1 \le a_m \le n - (k_m - 1),$$

which proves the first claim of the lemma. If $m = 1$, the second claim is trivial since the empty set is an MCS-specifications by lengths and initial elements for any positive integer. Now let $m > 1$. It is clear that T has properties (i) and (ii) of Definition 2.1. Moreover, we have

$$\begin{aligned} a_{m-1} + (k_{m-1} - 1) &\le a_m - 1 \\ &\le n - k_m \\ &\le n - (k_m - 1), \end{aligned}$$

and thus T satisfies (iii) of Definition 2.1 as well. □

Notation 2.4 We let $\mathcal{U}_n$ denote the subset of $\mathcal{P}_n$ that consists of all permutations with no consecutive sequences.

It is well-known (see e.g. [2]) that the cardinality of $\mathcal{U}_n$ satisfies the recurrence relation

$$|\mathcal{U}_n| = (n-1) \cdot |\mathcal{U}_{n-1}| + (n-2) \cdot |\mathcal{U}_{n-2}|.$$

The theorem below provides the desired algorithm for calculating $|\mathcal{Q}_{(n,S)}|$ by reducing the problem to the calculation of $|\mathcal{U}_r|$ for a certain r.

Theorem 2.5 Let $n \ge 1$, let $S = \{\, (a_1, k_1), (a_2, k_2), \ldots, (a_m, k_m) \,\}$ be an MCS-specification by lengths and initial elements for n, and let $k = \sum_{i=1}^{m} k_i$. Then $|\mathcal{Q}_{(n,S)}| = |\mathcal{U}_{n-(k-m)}|$.

Proof We will prove the theorem by showing that there is a bijection between $\mathcal{Q}_{(n,S)}$ and $\mathcal{U}_{n-(k-m)}$. For this, it suffices to show that there are maps

$$f : \mathcal{Q}_{(n,S)} \to \mathcal{U}_{n-(k-m)} \quad \text{and} \quad g : \mathcal{U}_{n-(k-m)} \to \mathcal{Q}_{(n,S)}$$

such that $g \circ f$ is the identity on $\mathcal{Q}_{(n,S)}$ and $f \circ g$ is the identity on $\mathcal{U}_{n-(k-m)}$. Intuitively speaking, f is obtained by throwing out all elements of maximal consecutive sequences except for the initial ones, then adjusting greater elements of the permutation downward to close the gaps. The map g is the reverse operation of that. For a formal proof of the existence of these maps, we proceed by induction on m. For $m = 0$, the claim is trivial as

$$\mathcal{U}_{n-(k-m)} = \mathcal{U}_n = \mathcal{Q}_{(n,S)}$$

in that case. Now let $m > 0$, and let

$$T = \{\, (a_1, k_1), (a_2, k_2), \ldots, (a_{m-1}, k_{m-1}) \,\}.$$

By Lemma 2.3, T is an MCS-specification by lengths and initial elements for $n-(k_m-1)$. This together with the induction hypothesis implies that there is a bijection between

$$\mathcal{Q}_{(n-(k_m-1),T)} \quad \text{and} \quad \mathcal{U}_{(n-(k_m-1))-((k-k_m)-(m-1))} = \mathcal{U}_{n-(k-m)}.$$

Therefore, it suffices to construct maps

$$f : \mathcal{Q}_{(n,S)} \to \mathcal{Q}_{(n-(k_m-1),T)} \quad \text{and} \quad g : \mathcal{Q}_{(n-(k_m-1),T)} \to \mathcal{Q}_{(n,S)}$$

such that $g \circ f$ is the identity on $\mathcal{Q}_{(n,S)}$ and $f \circ g$ is the identity on $\mathcal{Q}_{(n-(k_m-1),T)}$. For $P \in \mathcal{Q}_{(n,S)}$, let $f(P)$ be the integer sequence that is obtained from P as follows:

1. Strike the elements $a_m+1, a_m+2, \ldots, a_m+(k_m-1)$ from P.
2. In the remaining sequence, replace every element a that is greater than a_m with $a-(k_m-1)$.

For $Q \in \mathcal{Q}_{(n-(k_m-1),T)}$, first note that the integer a_m occurs in the sequence Q because $a_m \leq n-(k_m-1)$ by Definition 2.1 (iii). Now let $g(Q)$ be the integer sequence that is obtained from Q by reversing the procedure that defines f:

1. Replace every element a in Q that is greater than a_m with $a+(k_m-1)$.
2. Augment the resulting sequence by inserting the sequence $(a_m+1, a_m+2, \ldots, a_m+(k_m-1))$ following the element a_m.

It is easy to see that $f(P)$ contains exactly the integers $1, 2, \ldots, n-(k_m-1)$, and $g(Q)$ contains exactly the integers $1, 2, \ldots, n$, and therefore,

$$f(P) \in \mathcal{P}_{n-(k_m-1)} \quad \text{and} \quad g(Q) \in \mathcal{P}_n.$$

Also, it is immediate from the definition of f and g that $g \circ f$ is the identity on $\mathcal{Q}_{(n,S)}$ and $f \circ g$ is the identity on $\mathcal{Q}_{(n-(k_m-1),T)}$. It remains to show that

$$f(\mathcal{Q}_{(n,S)}) \subseteq \mathcal{Q}_{(n-(k_m-1),T)} \quad \text{and} \quad g(\mathcal{Q}_{(n-(k_m-1),T)}) \subseteq \mathcal{Q}_{(n,S)}.$$

So let $P \in \mathcal{Q}_{(n,S)}$. To show that $f(P) \in \mathcal{Q}_{(n-(k_m-1),T)}$, we must prove that $f(P)$ has precisely the maximal consecutive sequences that T specifies. Before delving into that argument, it may be helpful to visualize how $f(P)$ is obtained from P. Under the action of f, an element of the sequence P may be removed, change its position, change its value, change both position and value, or change neither position nor value. The elements $a_m+1, a_m+2, \ldots, a_m+(k-1)$, which we know are positioned consecutively, get removed. The elements that are positioned to the right of that subsequence, all the way to the end of P, move k_m-1 positions to the left. Finally, those elements are greater than a_m—and the only ones that are left are actually greater than $a_m+(k-1)$—are decremented by $k-1$. You may also want to remind yourself that the subscript m on a_m is not indicative of position in P or $f(P)$. It stems from the MCS-specification S.

Now imagine the sequence $f(P)$ being split in two, with the cut being after the element a_m. Let's call these two pieces P_1 and P_2. All the integers that are members of the $m-1$ maximal consecutive sequences in P starting with $a_1, a_2, \ldots, a_{m-1}$ are less than a_m. Therefore, their values are not changed under the action of f, and neither are their relative positions. Therefore, each of these sequences is present as a consecutive sequence in either P_1 or P_2. As for the elements in between and around those sequences, in P_1 or P_2, they are either less than or equal to a_m, in which case their value is unchanged under f, or they are greater than a_m, in which case they are the result of decrementing in lockstep, by the same amount, namely, $k_m - 1$. Moreover, no relative positions have changed among any of these under the action of f. It follows that none of these elements have joined any of the maximal consecutive sequences of P, and the only new consecutive pair that could have formed among them would be $(a_m, a_m + 1)$, but that's impossible since a_m sits at the end of P_1. We see that the maximal consecutive sequences that we find in P_1 and P_2 are precisely those that are specified by T.

It remains to show that no consecutive pair forms between the last element of P_1 and the first element of P_2 as we join the two to form $f(P)$. The last element of P_1 is a_m. The first element of P_2 is the result of the effect that f had on the first element following the maximal consecutive sequence $a_m, a_m+1, \ldots, a_m+(k-1)$ in P. That element was either less than or equal to a_m, in which case its value is unchanged, or it was greater than a_m and unequal to $a_m + k$, in which case its value was changed to something not equal to $a_m + 1$. In either case, no consecutive pair forms at the juncture of P_1 and P_2. This concludes the proof that $f(P) \in \mathcal{Q}_{(n-(k_m-1),T)}$ and thus $f(\mathcal{Q}_{(n,S)}) \subseteq \mathcal{Q}_{(n-(k_m-1),T)}$. We leave the proof of $g(\mathcal{Q}_{(n-(k_m-1),T)}) \subseteq \mathcal{Q}_{(n,S)}$ to the reader, as it is little more than the argument that we just made in reverse. $\square$

Since we know how to calculate the cardinality of $\mathcal{U}_n$ for any n, the theorem above gives us an algorithm to calculate the number of permutations of n integers that have maximal consecutive sequences of specified lengths with specified initial elements. However, judging from experience, that algorithm isn't very interesting. The description of consecutive sequences is just too specific. What one wants is being able to count the permutations with consecutive sequences or maximal consecutive sequences that are specified by length and count, as in, "exactly x number of consecutive triples," or, "exactly x number of consecutive triples and no longer consecutive sequences," or some such thing. This will be achieved in the next two sections.

As for the complexity of the algorithm of Theorem 2.5, it is clear that the classical recurrence relation for $\mathcal{U}_n$ that we stated preceding the theorem can be rewritten as a bottom-up multiplication that calculates $\mathcal{U}_n$ in constant space and linear time. Therefore, the complexity of the algorithm of Theorem 2.5 is $O(n)$.

As an aside, let us mention that Theorem 2.5 continues to hold if instead of specifying maximal consecutive sequences by initial element and count, we specify them by initial position and count. This follows from the fact that for $n \geq 1$, the map that exchances value and position is a bijection on $\mathcal{P}_n$. Here, the permutation $(a_1, a_2, \ldots, a_m)$ maps to the permutation where i is the element at position a_i for $1 \leq i \leq n$. Under this map, maximal consecutive sequences of length k with initial element a map to maximal consecutive sequences of length k that start at position a and vise versa.

3 Specifying Consecutive Sequences by Lengths And Counts

Definition 3.1 Let n be an integer with $n \geq 1$. An **MCS-specification by lengths and counts** for n is a set of pairs of integers

$$\{ (k_1, c_1), (k_2, c_2), \ldots, (k_m, c_m) \}$$

with the following properties:

(i) $k_i \geq 2$ and $c_i \geq 1$ for $1 \leq i \leq m$, and

(ii) $\sum_{i=1}^{m} c_i \cdot k_i \leq n$.

Notation 3.2 If T is an MCS-specification by lenghts and counts for n as in Definition 3.1 above, we write $\mathcal{R}_{(n,T)}$ for the set of all permutations $P \in \mathcal{P}_n$ with the following property: for each i with $1 \leq i \leq m$, P has exactly c_i maximal consecutive sequence of length k_i, and P has no other maximal consecutive sequences.

Purpose of this Section Present an algorithm for calculating $|\mathcal{R}_{(n,T)}|$ from n and T.

It is clear from Definitions 2.1 and 3.1 and the corresponding Notations 2.2 and 3.2 that $\mathcal{R}_{(n,T)}$ is the disjoint union of certain $\mathcal{Q}_{(n,S)}$, namely, those where S ranges over all those MCS-specifications by lengths and initial elements that are of the form

$$S = \{ (a_1, l_1), (a_2, l_2), \ldots, (a_p, l_p) \}$$

with the properties

(i) $p = \sum_{i=1}^{m} c_i$, and

(ii) for $1 \leq i \leq m$, there are exactly c_i many j with $1 \leq j \leq p$ and $l_j = k_i$.

So if we denote the set of all MCS-specifications by lengths and initial elements that satisfy (i) and (ii) above by $\mathcal{S}_T$, then we have, as a first step towards our algorithm for calculating $|\mathcal{R}_{(n,T)}|$,

$$|\mathcal{R}_{(n,T)}| = \sum_{S \in \mathcal{S}_T} |\mathcal{Q}_{(n,S)}|. \tag{1}$$

Theorem 2.5 tells us how to calculate $|\mathcal{Q}_{(n,S)}|$, and moreover, the algorithm for doing so uses only n, p, and $\sum_{j=1}^{p} l_j$. It is immediate from properties (i) and (ii) above that

$$p = \sum_{i=1}^{m} c_i \quad \text{and} \quad \sum_{j=1}^{p} l_j = \sum_{i=1}^{m} c_i \cdot k_i.$$

So if we let $c = \sum_{i=1}^{m} c_i$ and $k = \sum_{i=1}^{m} c_i \cdot k_i$, we can extend equation (1) above to the following second step towards our algorithm for calculating $|\mathcal{R}_{(n,T)}|$:

$$|\mathcal{R}_{(n,T)}| = |\mathcal{S}_T| \cdot |\mathcal{U}_{n-(k-c)}|. \tag{2}$$

Therefore, all that remains to do is to figure out what $|\mathcal{S}_T|$ is: how many MCS-specifications by lengths and initial elements are there that satisfy (i) and (ii) above? That number is fairly easy to describe: it is the number of ways in which one can choose subsets $A_1, A_2, \ldots, A_m \subset \{1, 2, \ldots, n\}$ such that

(a) $|A_i| = c_i$ for $1 \le i \le m$, and

(b) the elements of the A_i are far enough apart so that each $a \in A_i$ can be the initial value of a maximal consecutive sequence of length k_i.

At first glance, it may seem rather tricky to figure out the number of ways in which the A_i can be chosen. The key to making it easy lies in going back the proof of the equality $|\mathcal{Q}_{(n,S)}| = |\mathcal{U}_{n-(k-c)}|$, which we just used to pass from equation (1) to equation (2). This equality (Theorem 2.5) was proved by exhibiting a bijection between $\mathcal{Q}_{(n,S)}$ and $\mathcal{U}_{n-(k-c)}$. We mapped permutations with maximal consecutive sequences to shorter permutations without any consecutive sequences by striking from all maximal consecutive sequences all elements except for the first one, then renumbering the remaining elements to close the resulting gaps. The inverse operation consisted of starting with a permutation with no consecutive sequences, then blowing up the specified initial elements to consecutive sequences by inserting and renumbering elements. At the risk of being accused of a hand-waving argument, we'll say that it is now clear that picking the subsets $A_1, A_2, \ldots, A_m \subset \{1, 2, \ldots, n\}$ with properties (a) and (b) above is equivalent to picking subsets $B_1, B_2, \ldots, B_m \subset \{1, 2, \ldots, n-(k-c)\}$ with just property (a). The formal proof by induction parallels the proof of Theorem 2.5 and is simpler than the latter. Counting the ways in which the B_i can be selected is elementary. The answer is

$$\prod_{i=1}^{m} \binom{n-(k-c)-\sum_{j=1}^{i-1} c_j}{c_i},$$

or, equivalently,

$$\frac{(n-(k-c))!}{c_1! \cdot c_2! \cdot \ldots \cdot c_m! \cdot c!},$$

or, equivalently,

$$\frac{(n-(k-c)) \cdot (n-(k-c)-1) \cdot \ldots \cdot (n-(k-c)-c+1)}{c_1! \cdot c_2! \cdot \ldots \cdot c_m!}.$$

We have thus proved the following theorem, which provides the desired algorithm for calculating $|\mathcal{R}_{(n,T)}|$ from n and T.

Theorem 3.3 Let $n \geq 1$, let $T = \{ (k_1, c_1), (k_2, c_2), \ldots, (k_m, c_m) \}$ be an MCS-specification by lengths and counts for n, let $k = \sum_{i=1}^{m} c_i \cdot k_i$, and let $c = \sum_{i=1}^{m} c_i$. Then

$$|\mathcal{R}_{(n,T)}| = |\mathcal{U}_{n-(k-c)}| \cdot \prod_{i=1}^{m} \binom{n-(k-c)-\sum_{j=1}^{i-1} c_j}{c_i},$$

or, equivalently,

$$|\mathcal{R}_{(n,T)}| = |\mathcal{U}_{n-(k-c)}| \cdot \frac{(n-(k-c))!}{c_1! \cdot c_2! \cdot \ldots \cdot c_m! \cdot c!},$$

or, equivalently,

$$\begin{aligned} |\mathcal{R}_{(n,T)}| &= |\mathcal{U}_{n-(k-c)}| \cdot \\ &\quad \frac{(n-(k-c)) \cdot (n-(k-c)-1) \cdot \ldots \cdot (n-(k-c)-c+1)}{c_1! \cdot c_2! \cdot \ldots \cdot c_m!}. \quad \square \end{aligned}$$

It is clear that the complexity of the algorithm of 3.3 is $O(m \cdot n)$, which, depending on how the k_i are defined, can be anything between $O(n)$ and $O(n^2)$.

4 Iterating over Specifications by Lengths And Counts

Purpose of this Section Present a generic algorithm for counting the permutations that meet certain specifications by lengths and counts, where a client-supplied function performs the selection of specifications to be included in the count.

Now that we know how to calculate $|\mathcal{R}_{(n,T)}|$, that is, the number of permutations that meet a given specification by lengths and counts, it is an obvious and rather trivial thing to write an algorithm that performs an in-place creation of every specification by lengths and counts for a given n and lets a user-provided function decide which ones should be included in the count. Therefore, the following theorem requires no further proof.

Theorem 4.1 Let $n \geq 1$, let $\mathcal{T}$ be the set of all MCS-specifications by lengths and counts for n, and let f be a function from $\mathcal{T}$ to the set $\{0, 1\}$. Then the expression

$$\sum_{\{T \in \mathcal{T} \mid f(T)=1\}} |\mathcal{R}_{(n,T)}|$$

amounts to an algorithm for calculating the number of permutations that meet exactly those MCS-specifications by lengths and counts for n on which the function f returns 1. □

Since the algorithm of theorem 4.1 iterates over all MCS-specifications by lengths and counts for n, it is clear that it is exponential in n. The implementation of the algorithm at [**?**] organizes the process of accepting and rejecting specifications by lengths and counts in such a way that in many cases, not every specification needs to be looked at.

References

[1] Even if the author's mathematical specialty were combinatorics, which it is not, it would be foolish to attempt an overview or an even remotely complete set of references in a short article like this. A good place to learn about existing results and start finding references is the Online Encyclopedia of Integer Sequences, specifically the entries A010027, A002628, and A0000255.

[2] This proof by Jed Yang on Quora is short, elegant, and self-contained.