Why You Should Remove Zeros From Data Before Guessing

Manuel Kauers* and Thibaut Verron†
Institute for Algebra
Johannes Kepler University
Linz, Austria
[manuel.kauers|thibaut.verron]@jku.at

Abstract

A common advice in automated guessing is that when the input is a sequence like $2, 0, 0, 7, 0, 0, 17, \ldots$, one should remove the zeros before passing the data to the guesser. On this poster, we explain why this approach is sound, and which problem may arise if one does not take this step.

1 Introduction

Guessing is a well-known and popular technique in experimental mathematics [6, 2, 4]. The task is to derive from the given first few terms of an infinite sequence a plausible hypothesis about relations that the sequence may satisfy. In this poster, we have a closer look at what happens when the sequence under consideration is interlaced with zeros.

For example, consider a sequence (a_n) starting with the 16 terms

$$2, 0, 0, 7, 0, 0, 17, 0, 0, 29, 0, 0, 41, 0, 0, 53, \ldots$$

In order to find a recurrence, we may set up an equation template $\sum_{j=0}^{3} \sum_{i=0}^{2} c_{i,j} n^i a_{n+j}$ with 12 undetermined coefficients $c_{i,j}$, instantiate it with $n = 0, 1, 2, \ldots, 12$, and produce a linear system in the $c_{i,j}$'s. Note that we cannot instantiate at $n \geq 13$, because it would require knowing the values of a_{16} and beyond. Anyhow, the resulting linear system is

$$\begin{pmatrix} 2 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 7 & \ldots & 0 \\ 7 & 21 & 63 & 0 & \ldots & 153 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 41 & 492 & 5904 & 0 & \ldots & 7632 \end{pmatrix} \begin{pmatrix} c_{0,0} \\ c_{0,1} \\ c_{0,2} \\ c_{1,0} \\ \vdots \\ c_{3,2} \end{pmatrix} = 0,$$

and since it has more equations than variables, it is generically not expected to have a nonzero solution. The fact that it has the solution $(3701376, -121173, 62423, 0, 0, 0, 0, 0, 0, -1057536, -31575, -50387)$ would typically be interpreted as evidence in favor of the dubious hypothesis that (a_n) satisfies the recurrence

$$(50387n^2 + 31575n + 1057536)a_{n+3} - (62423n^2 + 121173n - 3701376)a_n = 0.$$

*Supported by the Austrian FWF grant P31571-N32
†Supported by FWF grant F5004.

On the other hand, if we remove the zeros from the sequence, that is we consider the sequence $(\bar{a}_n)$ starting with

$$2, 7, 17, 29, 41, 53, \ldots,$$

and if we make an ansatz for a recurrence of order 1 and degree 2, we are led to the *underdetermined* system

$$\begin{pmatrix} 2 & 0 & 0 & 7 & 0 & 0 \\ 7 & 7 & 7 & 17 & 17 & 17 \\ 17 & 34 & 68 & 29 & 58 & 116 \\ 29 & 87 & 261 & 41 & 123 & 369 \\ 41 & 164 & 656 & 53 & 212 & 848 \end{pmatrix} \begin{pmatrix} c_{0,0} \\ c_{0,1} \\ c_{0,2} \\ c_{1,0} \\ c_{1,1} \\ c_{1,2} \end{pmatrix} = 0$$

with 6 variables and 5 equations, and we cannot expect its solution to tell us anything about the infinite sequence. Typical implementations of guessing [3, 5] raise an error when the number of variables exceeds the number of equations. On this poster, we propose a slightly refined condition that also catches the situation encountered in the previous example, and we show that no relations can be missed by deleting zeros. For simplicity, we analyze phenomenon from the perspective of the linear algebra guessing approach sketched above, but the effects are not specific to this particular approach and appear also when, for example, Hermite/Pade approximation [1] is used applied instead of linear algebra.

2 Can we lose anything by removing the zeros?

Let $m \in \mathbb{N}$, consider a sequence (a_n) with $a_n \neq 0 \Rightarrow m \mid n$, and let $(\bar{a}_n)$ be defined by $\bar{a}_n = a_{mn}$. In that situation, common wisdom suggests removing the zeros and working with $\bar{a}_n$ instead.

Clearly, any recurrence of order r and degree d for $(\bar{a}_n) = (a_{mn})$ can be translated into a recurrence of order mr and degree d for (a_n). In order to be sure that no relations are lost when considering $(\bar{a}_n)$ instead of (a_n), we need the following converse.

Proposition 1. *Let $m \in \mathbb{N}$, and $(a_n) \in K^{\mathbb{N}}$ be a sequence such that $a_n = 0$ whenever n is not divisible by m. Assume that (a_n) satisfies a recurrence relation of order r and degree d,* i.e.*, there exists polynomials $f_0, \ldots, f_r \in K[n]$ with degree d, not all zero, such that for all $n \in \mathbb{N}$,*

$$\sum_{i=0}^{r} f_i(n) a_{n+i} = 0. \tag{1}$$

Let $(\bar{a}_n)$ be the sequence defined, for all $n \in \mathbb{N}$, by $\bar{a}_n = a_{mn}$. Then there exists $k \in \{0, \ldots, m-1\}$ such that:

- *the sequence $(\bar{a}_n)$ satisfies the recurrence relation of order $\bar{r} := \lfloor r/m \rfloor$ and degree d*

$$\sum_{i=0}^{\bar{r}} f_{mi+k}(mn+k) \bar{a}_{n+i} = 0; \tag{2}$$

- *the sequence (a_n) satisfies the recurrence relation of order $m\bar{r} = m \lfloor r/m \rfloor$ and degree d*

$$\sum_{i=0}^{\bar{r}} f_{mi+k}(n-k) a_{n+mi} = 0. \tag{3}$$

Proof. At least one of $f_0, \ldots, f_r$ is non-zero, say f_s, and let k be the remainder of dividing s by m. Consider the relation of Eq. (1) at index $mn+k$, for $n \in \mathbb{N}$. Reordering the terms $f_i(mn)a_{mn-k+i}$ according to the remainder of i modulo m yields

$$\sum_{\substack{0 \le i \le r \\ i \bmod m = k}} f_i(mn+k)a_{mn+k+i} + \sum_{\substack{0 \le i \le r \\ i \bmod m \neq k}} f_i(mn+k)a_{mn+k+i} = 0.$$

Since all a_{mn+k+i} in the second summand are zero, that sum is zero. The first summand can be rewritten, with the change of variable $i \leftarrow \frac{i+k}{m}$, as

$$0 = \sum_{i=0}^{\lfloor r/m \rfloor} f_i(mn+k)a_{m(n+i)} = \sum_{i=0}^{\bar{r}} f_i(mn+k)\bar{a}_{n+i}$$

which is the wanted relation (2) for $(\bar{a}_n)$.

From the above, the recurrence (3) is satisfied for all n which are divisible by m. If n is not divisible by m, all involved values a_{n+mi} are zero, and the recurrence (3) is again satisfied. So the sequence (a_n) satisfies recurrence (3) as a whole. □

Besides recurrence equations of a sequence (a_n), we can also guess polynomial equations of the corresponding formal power series $a(x) = \sum_{n=0}^{\infty} a_n x^n$. From the first N coefficients of such a series, we can create N linear equations for the unknown coefficients in an ansatz $\sum_{i=0}^{r}\sum_{j=0}^{d} c_{i,j}x^j a(x)^i = 0$ for a polynomial equation. If (a_n) and $(\bar{a}_n)$ are related as before, then $\bar{a}(x) = \sum_{n=0}^{\infty} \bar{a}_n x^n$ and $a(x)$ are related through $a(x) = \bar{a}(x^m)$, and we can ask how the polynomial equations of these series are related to one another. The answer is as follows.

Proposition 2. *Assume that $a(x)$ satisfies a polynomial relation with degree d in t and r in $a(x)$, then:*

- *$\bar{a}(x)$ satisfies a polynomial relation $\bar{P}(x, \bar{a}(x)) = 0$ with degree $\lfloor d/m \rfloor$ in x and r in $\bar{a}(x)$;*
- *$a(x)$ satisfies the polynomial relation $\bar{P}(x^m, a(x)) = 0$ with degree $m\lfloor d/m \rfloor$ in x^m and r in $a(x)$.*

3 Refined Confidence Conditions

In general, if we want to guess a recurrence of order $\bar{r}$ and degree d for an arbitrary sequence $(\bar{a}_n)$, then from the first $\bar{N}$ terms of the sequence we can get $\bar{N} - \bar{r}$ linear equations, and since there are $(\bar{r}+1)(d+1)$ variables in the ansatz, we need to have at least $\bar{N} \ge (\bar{r}+1)(d+2)$ in order to ensure that the system is overdetermined.

Now let $m \in \mathbb{N}$ and consider the sequence (a_n) defined by $a_{mn} = \bar{a}_n$ and $a_n = 0$ if $n \nmid m$. As shown in the previous section, (a_n) satisfies a recurrence of order $r = m\bar{r}$ and degree d if and only if $(\bar{a}_n)$ satisfies a recurrence of order $\bar{r}$ and degree d. But if we search directly for the recurrence of (a_n), we have $N = m\bar{N}$ terms at our disposal, from which we can generate $N - r = m(\bar{N} - \bar{r})$ equations for the $(r+1)(d+1) = (m\bar{r}+1)(d+2)$ variables, and we encounter an overdetermined system as soon as $\bar{N} \ge \frac{1}{m}(m\bar{r}+1)(d+2)$.

The two bounds are not exactly the same. For example, for $m = 3, \bar{r} = 2, d = 2, \bar{N} = 6$, we have $\bar{N} \ge \frac{1}{m}(m\bar{r}+1)(d+2)$ but not $\bar{N} \ge (\bar{r}+1)(d+2)$. This is the situation we observed in the introduction. On the other hand, we have seen in Prop. 1 that the two guessing problems are essentially equivalent. At first glance, this seems to suggest that instead of removing zeros, we should rather add zeros, because this can turn a setting corresponding to an underdetermined system into a setting for an overdetermined system. But of course, these overdetermined systems will not be generic, and they will have "fake" solutions like in the example.

A more sensible course of action is to take into account this defect of genericity when counting how many values are needed for guessing a relation for (a_n). Namely, in order to guess a recurrence of order r and degree d for (a_n), one should ensure to have at least N terms, for some N such that

$$N \geq m\left(\left\lfloor \frac{r}{m} \right\rfloor + 1\right)\Big(d+2\Big).$$

Similarly, if $a(x) = \bar{a}(x^m)$ and we want to guess a polynomial equation of order r and degree d for $a(x)$, without removing zeros, one should use at least N terms, for some N such that

$$N \geq m\Big(r+1\Big)\left(\left\lfloor \frac{d}{m} \right\rfloor + 1\right).$$

4 What about differential equations?

Interestingly, the case of differential equations is slightly different. If a power power series $a(x) = \sum_{n=0}^{\infty} a_n x^n$ satisfies a certain linear differential equation $\sum_{i=0}^{r}\sum_{j=0}^{d} c_{i,j}x^j a^{(i)}(x) = 0$, then neither removing nor adding zeros may be a good idea (unless we write the differential equation in terms of the Euler derivation rather than the standard derivation, but this is cheating because it changes the degree). As shown by the following example, there is no immediate counterpart of Propositions 1 and 2.

Example 3. *Consider the formal power series* $\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} + \dots$. *It satisfies the differential equation*

$$\cos''(x) + \cos(x) = 0$$

of order 2 *and degree* 0. *However,* $a(x) = \cos(x^2)$ *does not satisfy any differential equation with degree* 0 *or* 1, *the smallest equations have order* 2 *and degree* 3, *or order* 3 *and degree* 2*:*

$$xa''(x) - a'(x) + 4x^3a(x) = a^{(3)}(x) + 4x^2a'(x) + 12xa(x) = 0.$$

Removing the zeros in the data of $\cos(x)$ *also leads to a larger equation for* $b(x) = \cos(\sqrt{x})$, *at its smallest annihilating equation has order* 2 *and degree* 1*:*

$$4xb''(x) + 2b'(x) + b(x) = 0.$$

References

[1] B. Beckermann and G. Labahn. A uniform approach for the fast computation of matrix-type Padé approximants. *SIAM Journal on Matrix Analysis and Applications*, 15(3):804–823, 1994.

[2] W. Hebisch and M. Rubey. Extended Rate, more GFUN. *Journal of Symbolic Computation*, 46(8):889–903, 2011.

[3] M. Kauers. Guessing handbook. Technical Report 09-07, RISC-Linz, 2009.

[4] M. Kauers. The holonomic toolkit. In *Computer Algebra in Quantum Field Theory: Integration, Summation and Special Functions*, pages 119–144. Springer, 2013.

[5] M. Kauers, M. Jaroschek, and F. Johansson. Ore polynomials in Sage. In *Computer Algebra and Polynomials*, LNCS 8942, pages 105–125. Springer, 2014.

[6] B. Salvy and P. Zimmermann. Gfun: a Maple package for the manipulation of generating and holonomic functions in one variable. *ACM Transactions on Mathematical Software*, 20(2):163–177, 1994.