A Lower Bound and an Approximation for the Van der Waerden Number Function $W(c, k)$

Eli Maynard*

Advised by William Gasarch of the University of Maryland

November 2018

*With help from Noah Gleason and Ryan Tse of Montgomery Blair High School

Abstract

We deductively derived two lower bounds and empirically derived an approximation of the Van der Waerden number function $W(c, k)$ at $c = 2$. $W(c, k)$ gives the lowest $n \in \mathbb{N} \mid \forall c$coloring C of size n, $\exists$ some monochromatic arithmetic subsequence of size k in C. To arrive at the approxmiation, we programmatically generated and then analyzed approximations of probabilities that some arbitrary 2-coloring of size n contains a monochromatic arithmetic subset of size k. We iteratively found the lowest n for each k to produce a 100% probability; approximating this n for a given k is approximating $W(2, k)$.

1 Introduction

1.1 c-Colorings

A *c-coloring* of size n is a sequence of n elements, where each element is one of c distinct items, or "colors". For instance, a 3-coloring of size 4 may be "Red Red Blue Purple". In this paper, c will never exceed 9, so c-colorings will be represented as a number in which each digit corresponds to one color. The previous coloring would be written as just "0012"; 0 for Red, 1 for Blue, and 2 for Purple.

1.2 Monochromatism and Monochromatic Arithmetic Subsequences

A *monochromatic arithmetic subsequence* of size k, or *MAS(k)*[1], is a selection of items from a c-coloring such that all items are the same distance apart and the same color. For instance:

Coloring: 230112320103231133

MAS(4): 3 3 3 3

This paper will use 1-indexing. So, the indicies of the 3s are $2, 7, 12,$ and 17, respectively. Since there is a fixed distance between each item, i.e., $17 - 12 = 12 - 7 = 7 - 2 = 5$, the subsequence is arithmetic; since each item is the same color, i.e., $3 = 3 = 3 = 3$, it is monochromatic; therefore, it is a MAS; since it has 4 items, it is a MAS(4).

1.3 Van der Waerden's Theorem

Van der Waerden's Theorem states that

$$\forall k, c \in \mathbb{N}, \exists n \in \mathbb{N} \mid \forall c\text{-coloring } C \text{ of size } n, C \text{ has some MAS}(k).$$

The smallest satisfactory n for a given c, k is denoted $W(c, k)$; this function is the "Van der Waerden number function".

That is, given a number of colorings c and subsequence size k, there exists some $W(c, k)$; any c-coloring of the size $W(c, k)$ has a monochromatic arithmetic subsequence of size k. Furthermore, this property doesn't hold for any Natural number less than $W(c, k)$.

Consider an example. Choose a number... Unfortunately, since this is a paper, I cannot know the number you chose. I'll assume it's 10. So let $k = 10$. Then consider the natural numbers $\mathbb{N}$ in which we classify each number as either a prime or a composite. This is two categories, so $c = 2$. Van der

[1] Also known as a *size-k arithmetic progression*

Waerden's theorem guarantees that we can always find $k = 10$ numbers evenly spaced which are either all prime or all composite in any subsequence of $\mathbb{N}$ of size $\geq W(2, 10)$.

Note that, as well as colorings smaller than $W(c, k)$ *not* being guaranteed a MAS(k) (due to $W(c, k)$ being the minimal value), all colorings greater than $W(c, k)$ *are* guaranteed a MAS(k). This is because all colorings larger than $W(c, k)$ contain a coloring of the size $W(c, k)$, which has a guaranteed MAS(k).[2]

1.4 Motivation for Research

Information about $W(c, k)$ does not immediately lend itself into any particular application. It is, however, part of Ramsey Theory, which continues to find applications, both in the theoretical: "Logic [...] Concrete Complexity [...] Complexity Classes [...] Parallelism [...] Algorithms [... and] Computational Geometry" (**?**), and the concrete: "communications, information retrieval, [...] and decisionmaking" (**?**). Due to the continued application of Ramsey Theory as a whole, we have faith that Van der Waerden's theorem will eventually find practical use.

1.5 Some Facts

The number of colorings for some given c, n, k is

$$\#\text{col} = c^n$$

since each of n positions in the coloring may be colored one of c ways.

The probability that a given arbitrary arithmetic subseqence of some arbitrary coloring is monochromatic is

$$\%\text{ AS is MAS} = \frac{c}{c^k} = c^{1-k}$$

since of c^k ways that the subsequence may have been colored, only c of them are monochromatic: one for each color.

The number of arithmetic subsequences in a c-coloring of size n is

$$\#\text{AS/col} = \frac{n^2 - n}{2(k-1)} + \frac{2-k}{2} - \epsilon \text{ for } \epsilon := \sum_{p_0=1}^{n-k+1} \frac{mod(n - p_0, k-1)}{k-1}$$

This may be understood by discussing the process of choosing an arithmetic subsequence from a given

[2] As a sidenote, this means that, $W(c, k)$ "splits" $\mathbb{N}$ into two contiguous sections: numbers for which c-colorings of that size are not guaranteed a MAS(k), and numbers for which c-colorings of that size are guaranteed to have a MAS(k).

coloring. One way to do this is with two steps, as follows:

1. Choose the starting point p_0 of the subsequence
2. Choose the distance between each item of the subseqnce

Step (1.) has $n-k+1$ options, from position 1 to position $n-k+1$. Step (2.), for a given p_0, has $\left\lfloor \frac{n-p_0}{k-1} \right\rfloor$ options. As such, the total number of arithmetic subsequences in a c-coloring of size n is

$$\sum_{p_0=1}^{n-k+1} \left\lfloor \frac{n-p_0}{k-1} \right\rfloor$$

And since

$$\left\lfloor \frac{a}{b} \right\rfloor = \frac{a - mod(a,b)}{b}$$

then we may replace the sum with

$$\sum_{p_0=1}^{n-k+1} \frac{n-p_0}{k-1} - \sum_{p_0=1}^{n-k+1} \frac{mod(n-p_0, k-1)}{k-1}$$

Calling the second sum ϵ, Wolfram|Alpha tells us that this is equal to

$$\frac{n^2-n}{2(k-1)} + \frac{2-k}{2} - \epsilon.$$

Note that since

$$0 \leq mod(n-p_0, k-1) \leq k-2$$

then

$$0 \leq \epsilon \leq \frac{k-2}{k-1}(n-k+1)$$

1.6 A Lower Bound on $W(c,k)$

Take

$$n < \frac{1}{2}\left(1+\sqrt{8(k-1)c^{k-1}+(3-2k)^2}\right)$$
$$n < \frac{1}{2}\left(1+\sqrt{1-12k+4k^2+8(k-1)c^{k-1}+8}\right)$$
$$n < \frac{1}{2}\left(1+\sqrt{1-4(3k-k^2-2(k-1)c^{k-1}-2)}\right)$$
$$n < \frac{1}{2}\left(1+\sqrt{1-4(2k-k^2-2kc^{k-1}-2+k+2c^{k-1})}\right)$$
$$n < \frac{1}{2}\left(-(-1)+\sqrt{(-1)^2-4(1)(k-1)(2-k-2c^{k-1})}\right)$$
$$0 \le n < \frac{1}{2}\left(-(-1)+\sqrt{(-1)^2-4(1)(k-1)(2-k-2c^{k-1})}\right) \text{ since } n \in \mathbb{N}^+$$
$$n^2-n+(k-1)(2-k-2c^{k-1}) < 0 \text{ by the quadratic formula}$$
$$(2-k)(k-1)+n^2-n < 2c^{k-1}(k-1)$$
$$(2-k)(k-1)+n^2-n < 2c^{k-1}(k-1)$$
$$\frac{2-k}{2}+\frac{n^2-n}{2(k-1)} < c^{k-1}$$
$$\left(\frac{2-k}{2}+\frac{n^2-n}{2(k-1)}\right)c^{1-k} < 1$$
$$\left(\frac{2-k}{2}+\frac{n^2-n}{2(k-1)}-\epsilon\right)c^{1-k} < 1 \text{ since } \epsilon \ge 0$$
$$(\#\text{AS/col})(\%\text{ AS is MAS}) < 1$$
$$(\#\text{AS/col})(\%\text{ AS is MAS})(\#\text{col}) < \#\text{col}$$
$$\#\text{MAS} < \#\text{col}$$
$$\text{Some coloring has no MAS}$$
$$n < W(c,k)$$

Thus

$$L(c,k) = \frac{1}{2}\left(1+\sqrt{8(k-1)c^{k-1}+(3-2k)^2}\right)$$

is a lower bound of $W(c,k)$.[3]

1.6.1 Comparison to Existing Bounds

The first (**?**) lower bound on $W(c,k)$, presented by **?**, was $\sqrt{2(k-1)c^{k-1}}$. Our bound, though stronger, asymptotically approaches Erdös' and Rado's bound as $k \to \infty$.

[3]Motivation for the steps of the proof may be seen by reading the proof backwards.

There are certainly better bounds. Called the "best known (asymptotic) lower bound" by **?**, **?** presents that $\forall \epsilon > 0,\ \forall$ large enough $k,\ W(2,k) \geq \frac{2^k}{k^{-\epsilon}}$. **?** presents that for prime k, $W(2,k) > (k-1)2^{k-1}$. **?** presents "the best known bound for a large interval of" c that for prime k and $2 \leq c \leq k \leq k$, $W(2,k) > (k-1)^{c-1}$.

Though of these bounds are stronger than our bound, they have constraints, respectively that k is large enough, that k is prime, and that p is prime and $2 \leq c \leq p \leq k$. Since our bound has no such constraints then though it is weaker, it applies more generally and with more ease.

1.7 The ζ and ζ_a Functions

We define the ζ function to be

$$\zeta(c,n,k) := \text{The probability of a random } c\text{-coloring of size } n \text{ to have a MAS}(k).$$

We may approximate ζ by generating some number a ('a' for 'attempts') of c-colorings; the ratio between the number of generated colorings with a MAS(k) and the number of generated colorings is an approximation of $\zeta(c,n,k)$. We call this approximation $\zeta_a(c,n,k)$.

Note that $\zeta(c,n,k) = 1 \longrightarrow \zeta(c,n+1,k) = 1$. We generalized this to $\zeta_a(c,n,k) = 1 \longrightarrow \zeta_a(c,n+1,k)$. **This is wrong**. It is possible that $\zeta_a(c,n,k) = 1$ "by coincidence", i.e., despite that $\zeta(c,n,k) \neq 1$. In this case, we know nothing about $\zeta(c,n+1,k)$ and therefore nothing about $\zeta_a(c,n+1,k)$. We adopted this assumption anyway in order to speed up the program. This makes the generated data a somewhat worse approximation.[4]

1.8 The Shape of ζ

... The thrilling mathematical sequel to the 2017 movie *The Shape of Water*.

It is reasonable to express ζ as "the probability that it is *not* the case that *every* arithmetic subsequence of the given coloring is *not* monochromatic"[5]. It is then *tempting* to express this as

$$\begin{aligned}\zeta &= 1 - (1 - \%\text{ AS is MAS})^{\#\text{AS/col}} \\ &= 1 - (1 - c^{1-k})^{\frac{n^2-n}{2(k-1)} + \frac{2-k}{2} - \epsilon}\end{aligned}$$

However, this is not quite correct. This expression uses the probabalistic multiplicaiton rule, which

[4]Some measures are taken against this issue and will be acknowledged in the Discussion section.
[5]i.e., $\neg\forall AS,\ \neg mono(AS) \longleftrightarrow \exists AS \mid mono(AS)$

may be applied to two *independent* events. However, one AS being monochromatic is *not* independent from another AS being monochromatic. We can also see that this is incorrect because though 1 is in ζ's range, it is not in the range of this expression.

Despite this, this expression gives us some insight. It leads us to suspect that ζ looks somewhat like this expression in some capacity[6]. Specifically, it will be important to consider the shape of ζ with respect to n. Observe that as $n \to \infty$, this expression $\to 1$ in an asymptotic manner. Note that since ζ reaches 1 eventually, ζ is not *actually* asymptotic, so we call it ¨textitnear-asymptotic".

1.9 Approximating $W(c, k)$

If we generate $\zeta_a(c, n, k)$ approximations iterating over n, then we may approximate $W(c, k)$ as the first n for which $\zeta_a(c, n, k) = 1$. We call this result V. Note that this is not a "mathematical variable" but rather the result of an algorithm which may change each time the algorithm is run.

2 Materials and Methods

The program was roughly the following:

```
1: function GENERATE-SUCCESS-COUNT(c, n, k, a)          ▷ Approximates a · ζ_a(c, n, k)
2:     successes ← 0
3:     loop a times
4:         coloring ← MAKE-RANDOM-COLORING(c, n)
5:         if coloring contains a MAS(k) then
6:             successes ← successes + 1
7:     return successes
8:
9: function TRIALS
10:    for c in [2] do                                  ▷ Only concerned with c = 2
11:        for k in [1, 2, ...] do                      ▷ Unboundedly increment k
12:            for a in [5k, 10k, 15k, ..., 500k] do    ▷ 'k' denoting "thousand"
13:                for n in [k, k + 1, k + 2, ...] do   ▷ Unboundedly increment n
14:                    successes ← GENERATE-SUCCESS-COUNT(c, n, k, a)
15:                    if successes = a then            ▷ If 100% success rate
16:                        RECORD-DATA(c, k, a, n)      ▷ Record V = n for c, k, a
17:                        skip to next a ▷ ζ_a(c, n, k) = 1 so know ∀n′ > n , ζ_a(c, n′, k) = 1 so skip all n′
```

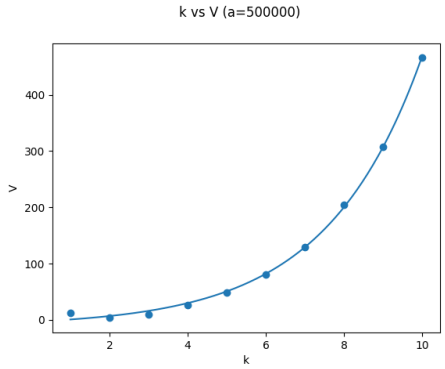
The actual source code is available online at `https://github.com/Quelklef/random-algorithms/tree/b50e7b0176bdc6c140ac0db2f2497fcdbef3ef4d`.

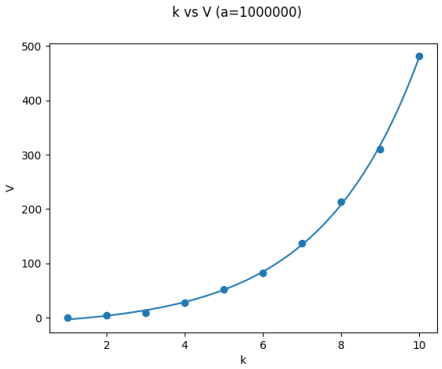
[6] And, though we no longer have the graphs to support so, we'll state that this is in fact a good approximation.

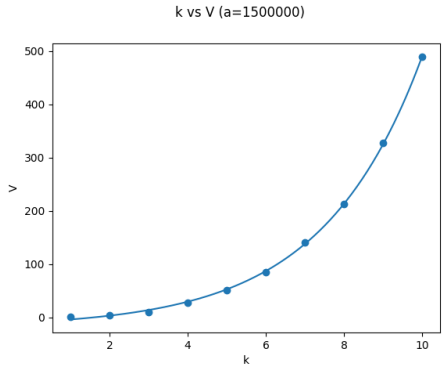
3 Results

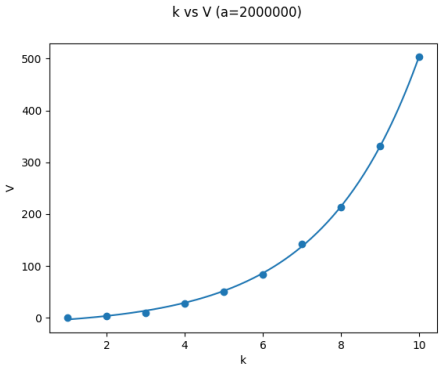
The following are graphs of k vs V from the trial data. Each graph was approximated with an exponential curve

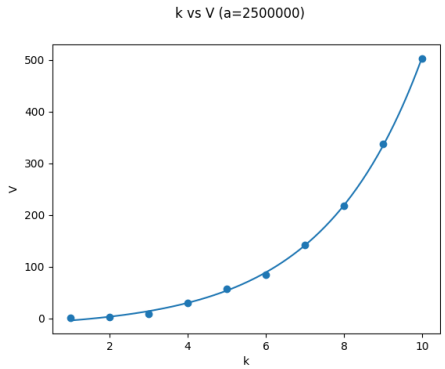
$$E(k) := y_0 + A \cdot e^{q \cdot (k - x_0)}$$

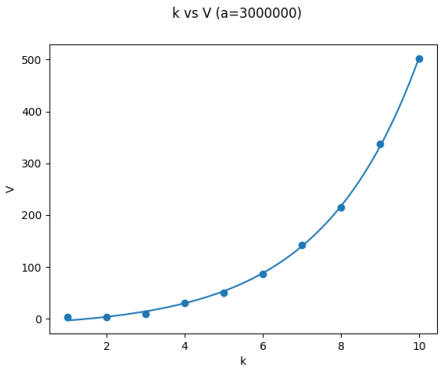


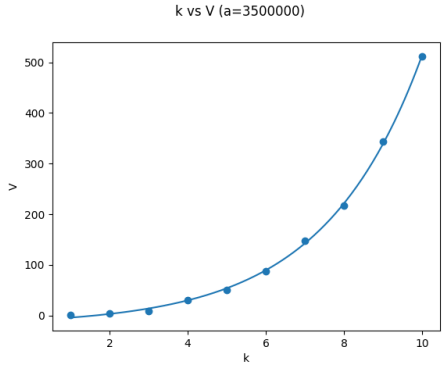


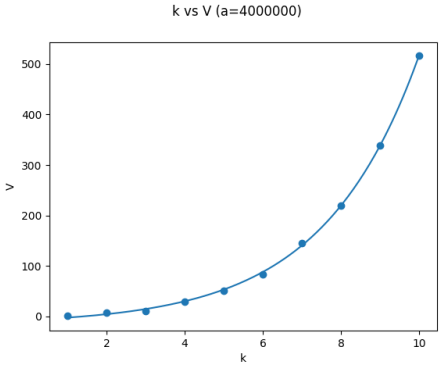


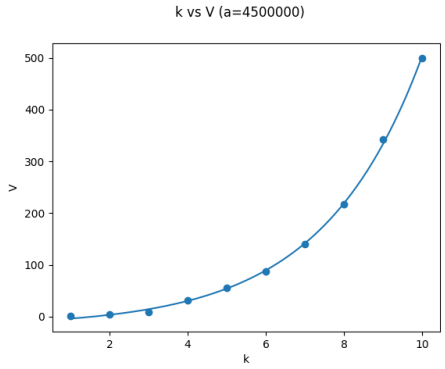


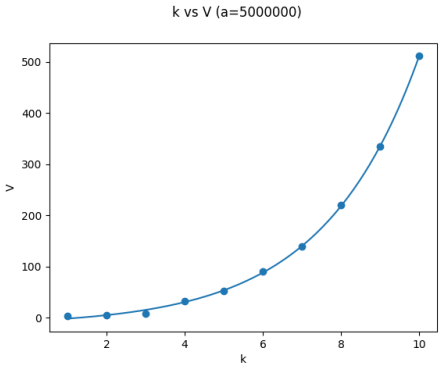


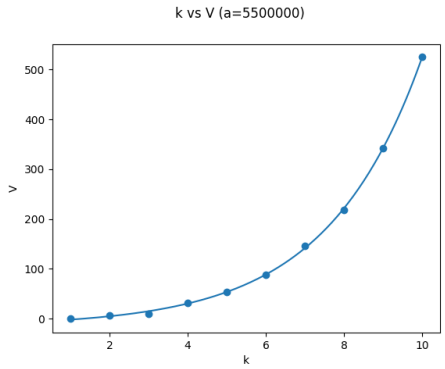


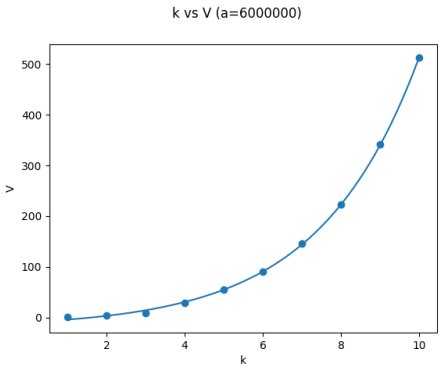

k vs V (a=3500000)
V
k

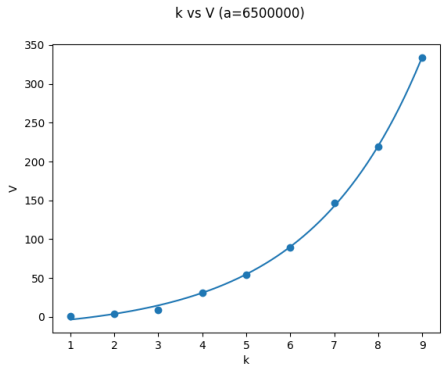

k vs V (a=4000000)
V
k

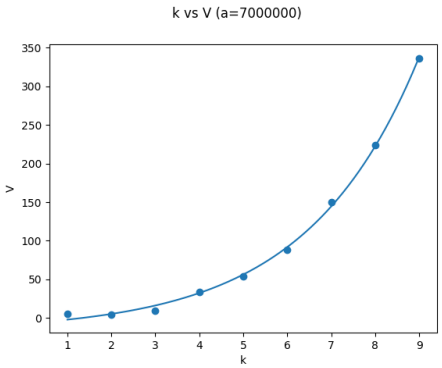

k vs V (a=4500000)
V
k


k vs V (a=5000000)
V
k


k vs V (a=5500000)
V
k


k vs V (a=6000000)
V
k


k vs V (a=6500000)
V
k


k vs V (a=7000000)
V
k

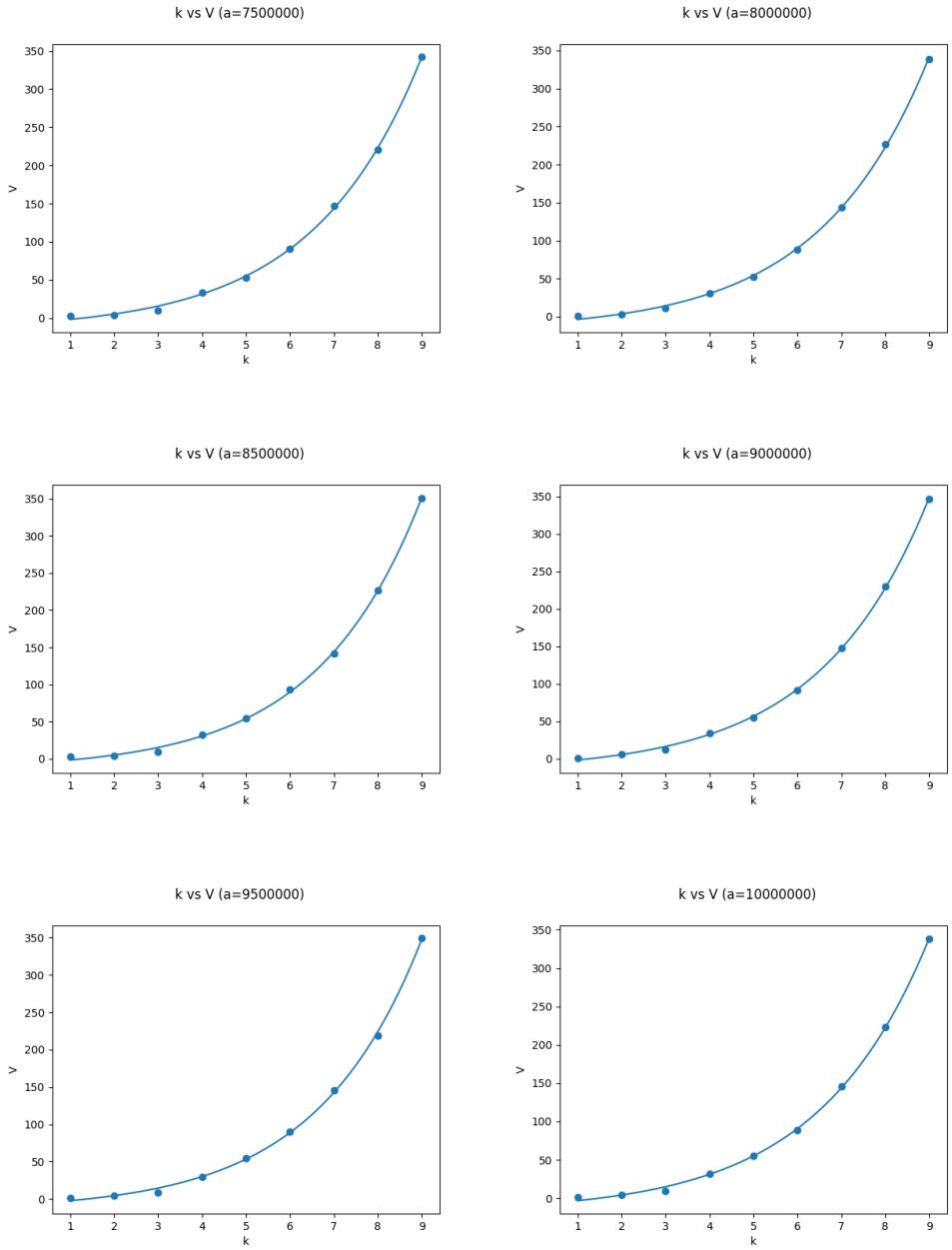
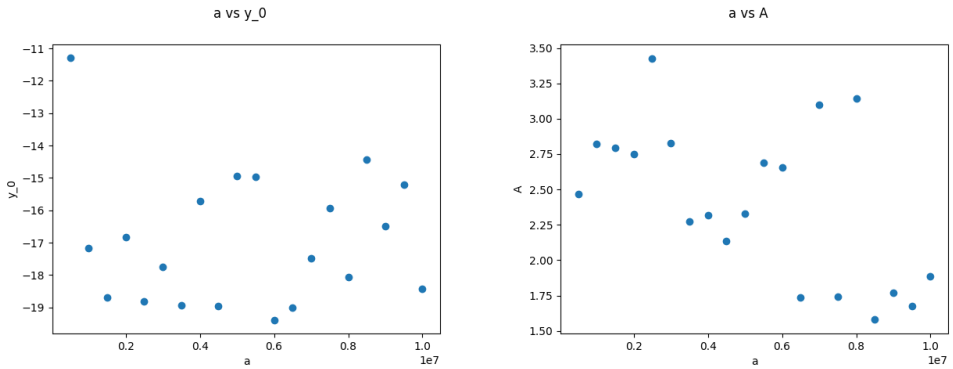


Figure 1: k vs V curves for fixed values of a, with an exponential fit.

The found values of y_0, A, q, and x_0 versus a are shown below.



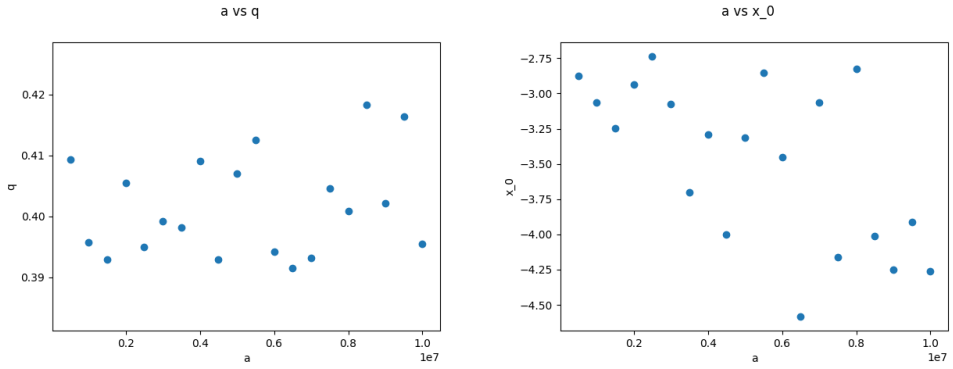


Figure 2: a versus parameter values for the previous exponential fittings.

4 Discussion

We expect V to only be a mediocre approximation of $W(c, k)$. This is because of ζ's near-asymptotic behavior. Since a ζ value *close to* 1 may result in a ζ_a value *of* 1, and since ζ is near-asymptotic, and therefore has many consecutive values close to 1 before actually reaching 1, then a ζ_a value may be 1 long before ζ is 1.

This issue is further exacerbated by our assumption that $\zeta_a(c, n, k) = 1 \longrightarrow \zeta_a(c, n + 1, k) = 1$; if we did not assume this, then we may, after finding an $n \mid \zeta_a = 1$, find an $n' > n \mid \zeta_a \neq 1$, thus showing that $n < W(c, k)$ since certainly $n' < W(c, k)$. Using this method, these lower "false positive" ns for which $\zeta_a = 1$ could be detected.[7]

The hope was that the generation of trials for several a values could fix this issue. Note that a can be seen as a kind of "confidence value". As a increases, we'd expect the "correctness" of the results to increase as well, since there are more trials and therefore random variance should lessen. We hoped that, though *each* a would have the issues mentioned above, we could see a pattern emerging over the iteration of *many* as, and this pattern would reveal the true curve of V.

However, this did not happen. The graphs of a against the four parameters y_0, A, q, and x_0 do not have a strong enough shape to justify any extrapolation. This is possibly due simply to not having enough trials; however, it could also be because an exponential fit is inappropriate. This would not be surprising as fitting exponentially was a total guess; a more informed fit would require knowing the shape of $W(c, k)$, which we don't know: if we did, then we wouldn't have had to do this research.[8]

The second-best options are to approximate these parameters either with the value given by the highest a, or an average over all as. We approximate them with an average because the graphs seem to indicate that different as actually have a minimal effect on the "correctness" of the values, so we may as

[7] A notable downside to this method is that it does not give a condition for stopping iteration over n.

[8] One may object that an approximation is nearly useless without *already* knowing the shape of the curve. This is valid and a major pitfall of this paper.

well use them all.

Figure 3: Point estimates of parameters to exponential fits of V

y_0	-16.925
A	2.406
q	0.402
x_0	-3.481

Thus, an approximation of W is:

$$W(2,k) \approx -16.925 + 2.406e^{0.402\cdot(k--3.481)}$$

We may compare this to known W(2,k) numbers

Figure 4: Comparison of known $W(2,k)$ values to estimated values.

k	Known value	Approximation	|Difference|
3	9	15.58	6.58
4	35	31.64	3.36
5	178	55.65	122.35
6	1132	91.53	1040.47

. . . and see that the approximation is not very good. Since it overestimates at first and underestimates lower on, the shape of the approximation seems to be too shallow. The error should only increase for higher ks, which is unfortunate because higher ks are of more significance[9]. We view this result as evidence that the actual shape of W is not exponential despite the k vs V graphs fitting seductively well to exponential curves.

[9]Due to W being easier to find for lower ks