A Computational Investigation of Bolzano's View of the Infinite

Max Rapp, student n° 11404310

July 6, 2017

1 Introduction

"Paradoxien des Unendlichen" - "paradoxes of the infinite" is the title of Bernard Bolzano's main work on the infinite. The paradoxes he is refering to stem from the joint inconsistency of the following three principles:

1. **Euclid's Principle:** If A is a proper subset of B then $|A| \neq |B|$.
2. **Cantor's Principle:** If there exists $f : A \to B$ s.t. f is a bijection then $|A| = |B|$
3. **Bolzano's Principle:** There exists a set A s.t. $|A| > n$ for all $n \in \mathbb{N}$

Namely, as Bolzano shows, infinite sets defined in terms of (3.) have the property that they admit bijections with some of their infinite proper subsets [1].

Anachronistically speaking, three roads have been taken to avoid these paradoxes: before Bolzano and Cantor, the infinite was considered ill-defined if its existence was not outright denied:

> I protest against the use of infinite magnitude as something completed, which is never permissible in mathematics. Infinity is merely a way of speaking, the true meaning being a limit which certain ratios approach indefinitely close, while others are permitted to increase without restriction.

wrote Carl Friedrich Gauss in a letter to a fellow mathematician in 1831, thus denying (3.). That the leading mathematician of his time thought that infinity was not even a valid object of mathematical investigation whereas it is well-understood today shows how bold and revolutionary the work of scholars of infinity such as Bolzano and Cantor was at the time.

Bolzano believed the infinite was worth studying but took a conservative approach, seeking to preserve the received Euclidian principle (2.). Thus he wrote that

> [Zwei Mengen] können trotz [der Existenz einer Bijektion zwischen ihnen] ein Verhältnis der Ungleichheit in ihren Vielheiten haben, so daß die eine derselben sich als ein Ganzes, davon die andere ein Teil, herausstellen kann. Auf eine Gleichheit dieser Vielheiten wird erst geschlossen werden dürfen, wenn irgendein anderer Grund noch dazukommt, wie etwa, daß beide Mengen ganz gleiche Bestimmungsgründe, z. B. eine ganz gleiche Entstehungsweise haben.
>
> [Two sets] may in spite [of the existence of a bijection between them] have a relation of inequality with respect to their multitudes [Vielheiten], such that it may turn out that one is a whole of which the other is a part. It may only be concluded that they are equal if some additional reason is present, such as, that both sets [Mengen] have completely equal defining conditions, e.g. a completely equal mode of generation. [My translation]

Finally, Cantor took the then revolutionary step to define equinumerosity in terms of bijections, thereby giving up the Euclidian principle for the infinite.

[1] Bolzano believed this to be a defini

2 Bolzano's change of mind?

Close to the end of his life, in a letter to a pupil, Bolzano appears to have repudiated his earlier view and endorsed (2.) instead of (1.). In earlier writings, he had maintained that due to the superset relation $\mathbb{N} \supset \mathbb{N}^2 \supset \mathbb{N}^4 \supset \mathbb{N}^8$..., $|\mathbb{N}| > |\mathbb{N}^2| > |\mathbb{N}^4| > |\mathbb{N}^8|$. However, in his letter Bolzano writes

> Die Sache ist nicht nur unklar vorge- tragen, sondern wie ich soeben zu erkennen anfange, ganz falsch. [...] [S]oll durch das Zeichen n jede beliebige ganze Zahl vorgestellt werden, so ist damit schon entschieden, welche (unendliche) Menge von Gegenständen dies Zeichen vorstelle. An dieser ändert sich nicht das geringste dadurch daß wir durch Zusatz eines Exponenten wie n2,n4, n8, n16 . . . . verlangen, daß jede dieser Zahlen jezt auf die zweite, jezt auf die vierte . . . Potenz erhoben werden soll. Die Menge der Gegenstände welche das n vorstellt, ist genau immer noch dieselbe wie vorhin, obgleich die Gegenstände selbst, die n2 vorstellt, nicht eben die nemlichen sind, welche n vorstellt.(to be found in the dataset in the folder "REIHE II_NACHLASS_BAND 12_VERMISCHTE PHILOSOPHISCHE_UND PHYSIKALISCHE SCHRIFTEN 1832–184\ 8_ZWEITER TEIL_9783772804717")
>
> The example is not only unclear in exposition but, as I now begin to realise, completely false. If the sign n shall denote an arbitrary whole number then it is thereby already decided which (infinite) set of objects (Gegenständen) this sign represents. This set is not changed in the slightest if by adding an exponent such as n2, n4, n8, n16 . . . . we demand that each of these numbers shall be raised raised to, be it the second, the fourth . . . power. The set of objects represented by n is still exactly the same as before, albeit the objects themselves, represented by n2, are not the same ones represented by n.[My translation]

Technically, Bolzano gives here clear preference to the existence of the bijection $n^2 : \mathbb{N} \to \mathbb{N}^2$ as the criterion of equinumerosity over the Euclidian principle. Has he changed his mind? Or is this inconsistency due to his old age as of the writing of this letter?

Bolzano scholar Jan Berg maintains that far from a fluke, Bolzano's letter constitutes "second achievement of major importance in Bolzano's [...] investigation of the infinite"Bolzano et al. (2012). However, Mancosu (2009) calls this "whig history" and "completely anachronistic". He maintains "While this could be argued for sets of natural numbers, the claim strikes me as implausible, if not downright false, when it comes to Bolzano's handling of infinite sets in geometrical contexts."

3 Can a Neural Network help?

Thus two experts on the issue disagree. Unfortunately, Bolzano's "recantation" is only preserved as a one piece fragment. Thus it gives little indication how far reaching Bolzano would have considered his insight in the letter and whether it would have affected any of his views outside the confined area of infinite sequences.

However, Mancosu's hypothesis depends on the implict assumption that Bolzano treats the infinite in a different way when it comes to geometry as opposed to arithmetic. Namely, the "Cantorian" notion of bijection should be less pronounced or absent in geometric contexts. Can we determine whether such a difference really exists in Bolzano's writing?

The classical way of approaching this problem would be close reading of Bolzano's work performed by experts on the topic.

In this experiment, a different approach from the emerging paradigm of computational philosophy is chosen: machine learning techniques, namely a neural network is applied to the problem.

3.1 How can a Neural Network predict?

Under a naive covering law view an expert investigation of the hypothesis "Bolzano treats the infinite differently in geometric and arithmetic contexts respectively" might be seen to have the following strengths:

- detailed elaboration of the hypothesis: "what would constitute a difference in treatment? what is a geometric, what an arithmetic context?
- Careful and context sensitive predictions: e.g. the expert will have detailed knowledge of Bolzano's terminology and might thus predict "given the hypothesis, term x should not appear in context y".
- Diligent appraisal of evidence: e.g. "is term x in the scope of a negation or intension? Does x at all pertain to the notion of infinity in this context?".

The neural network used in this experiment will be very bad with respect to all of these aspects. It will use an extremely rudimentary statistic to specify the hypothesis. It will test blunt and general predictions. It will not in any way address subtleties of meaning in the evidence.

However, it will be much better than an expert in the following respects:

- number of predictions: the neural network can keep in memory and simultaneously check a great number of different predictions
- amount of evidence: the neural network can test the hypothesis on an immense amount of evidence, in this case all the evidence there is, i.e.a complete corpus of Bolzano's writing.
- speed: it can test predictions within a few hours thus enabling repeated testing of different specifications of the hypothesis.

3.2 Idealizations

These advantages are tempting. Yet, how serious should we take the neural network's conclusion. This comes down to our trust in idealizations: if we regard an expert naively and reductively as a very complex neural network, then we have to trust that this very complex neural network would arrive at similar findings as our simple neural network if applied to the same corpus of data.

In the case of this experiment trusting the idealizations taken here is likely too much to ask. However, later I shall suggest some improvements that might render an undertaking like this more credible.

4 Dataset & Statistics

The dataset used in this experiment is a comprehensive corpus of Bolzano's known writings including books, manuscripts and letters. The material has been digitalized and in its majority not been processed for text analysis. Thus most of the material is not split into content units and writings by third parties (editors, commenters, etc.) have not been removed. In general (with a few exceptions), there is one .txt and .djbu file per book. Another problem is the presence of large amounts of meaningless digitalization artifacts resulting from formulas and symbols which could not be handled by the OCR algorithm.

4.1 Data Parsing

In order to make the data accessible to a machine learning algorithm several steps had to be taken. Firstly, for the present hypothesis, the document units (books, letters, ...) were regarded as irrelevant. Thus in a first step, all the individual text files in the datasets subdirectories were concatenated (without respecting order) into one big .txt file.

A reading function was then applied to this .txt file returning one long string of data. While reading I encountered a difficulty with non-unicode signs present in the data created mainly by faulty digitalization of formulas. The reading function was therefore adapted to ignore these non-unicode signs and converting the rest of the data to utf-8.

4.2 Specifying Context

Next a first choice regarding the specification of the hypothesis had to be made: what should one regard as a "context" for present purposes?

Luckily, a cursory viewing of the data revealed an obvious choice for the scope of contexts: in most of his writing, Bolzano adhered to the then common convention of using paragraphs to demarcate sections. Most of these paragraphs are between one half and one and a half pages long and they usually pertain to one specific problem or more general content unit. Thus they provide a simple yet powerful way to specify contexts. I therefore decided to split the datastring along the paragraph signs.[2]

Unfortunately, this method is not perfect: firstly, Bolzano sometimes refers back to earlier paragraphs in his writing resulting in an undesirable split. Even more frequently, third parties refer to paragraphs in Bolzano's writing. However, the latter issue is actually beneficial: the frequent splits tend to leave much smaller third party contexts then contexts by Bolzano. As we shall see, very small contexts are likely to affect our algorithm much less than longer contexts, reducing undesirable third party noise.

4.3 Geometric, Arithmetic and Cantorian Contexts

Next, with the help of secondary literature I performed a careful reading of some key sections in Bolzano's work relating to the infinite. These included mainly §20-22 in "Paradoxes of the Infinite", Bolzano's letter to Zimmermann containing his "recantation" and a small section of "Reine Zahlenlehre" dealing with the definition of arithmetic. From these readings I extracted a selection of key words characterizing three kinds of contexts: arithmetic contexts, geometric contexts, Cantorian contexts. The following are the respective keyword lists:

- **Arithmetic Contexts:** "ziffer", "arithmet", "zahl", "unbenannt", "konkret", "vielheit", "reihe", "begriff","zeichen", "gegenst"
- **Geometric Contexts:** "geomet", "gerade", "linie", "ebene", "körper", "raum", "entfernung", "entfernt", "punkt", "figur"
- **Cantorian Contexts:** "paar", "zählbar", "weite", "bestimmungsgründe", "enstehungsweise", "gleich", "verbindung", "regel", "verglich"

Note that since the data was kept as a string, no issues of word or line breaks had to be addressed. In addition a keyword such as "gleich" would also result in a positive hit for e.g. "Vergleichung" which Bolzano frequently uses. Finally, the datastring was transfomed into all lowercase so as to avoid issues with case sensitivity.

I will justify here only a few of the choices I made: arithmetic according to "Reine Zahlenlehre" is the study of "konrete unbenannte Zahlen". In addition, Bolzano regards sequences as "Reihen von Begriffen" where "Begriff" is somewhat analogous to an intensional definition of a set - e.g. "Der Begriff jeder beliebigen ganzen Zahl" corresponds to the set of whole numbers in modern parlance - with "Gegenstände" as the elements of sets.

Geometrical language is pervasive in Bolzano's writing so the key words selected here really form a subset of a wealth of options.

The most difficult task was identifying keywords for what I anachronistically call "Cantorian contexts". "Cantorian" as they pertain to the Cantorian principle of equinumerosity in that they treat what is today called bijections. Unfortunately Bolzano did not develop a terminology for bijections but rather used quite laborious definitions which employ very general terms such as "Paar", "Regel", "Verbindung", "gleich". These are so general that unfortunately some false positives are to be expected here.

4.4 Statistical Observations

A context was regarded as Arithmetic, (Geometric, Cantorian) if it contained four or more of the Arithmetic, (Geometric, Cantorian) keywords. In a first step the numbers of such contexts in the data were counted and compared. This was performed first on all 16414 contexts extracted from the data then on the subset of contexts containing the string "unendlich" ("infinite"). The latter step was taken since contexts that don't pertain to the infinite were deemed irrelevant for present purposes. Thus the number of contexts could be brought down to 1400 - promising to speed up computations a lot. Table 1 contains the results that were obtained by the counts.

[2] In the instances where the data had already been split into sections within the dataset the paragraph signs were removed. To deal with this an extra paragraph sign was added at the end of each file when reading.

Table 1: Contexts in Bolzano's Writings

#Contexts	16414	16414	1400
#Arithmetic	590	185	89
#Geometric	396	235	132
#Cantorian	538	155	64
#Arithmetic∩Geometric	190	91	60
#Arithmetic∩ Cantorian	221	64	45
#Geometric∩ Cantorian	192	67	39

Table 1 reveals a high degree of overlap between contexts. Since the context selection is rather restrictive (no category was judged to apply to more than one tenth about the contexts) this does not seem to be due to too general choices of key words.

Rather, it is in line with what one sees during cursory reading of Bolzano: examples from geometry and arithmetic frequently appear as pairs, often to illustrate the shape the problem discussed takes in either area. For example the discussion of infinite sets bijective to their subsets in Paradoxes of Infinity §21 contains an arithmetic and a geometric example.

4.5 Hypothesis representation

For application of a neural network a more fine grained approach was chosen: the hypothesis was represented as a function from the frequency of the key words within a given context to the presence or absence of a Cantorian context: $H : (N) \times (N) \times (N)... \times (N) \rightarrow \{0, 1\}$. In other words, if Mancosu is right then the neural network should return parameters that yield a hypothesis that predicts 1 whenever the context contains a high number of arithmetic key words and 0 else. Especially, geometric key words should have close to zero or negative parameter values.

For this purpose the independent variable values were derived as before. However, with respect to the dependent variables, the frequency of every geometric or arithmetic key word was now regarded as a feature. Thus the dependent variable values were derived by counting the instances of each arithmetic and geometric key word in a given context.

After having been prepared in this way the data was ready for the application of a neural network.

5 Application of Neural Network

The neural network was implemented from scratch for practice purposes (alternatively, one of the many readymade packages could have been used). In what follows I will give an overview of the algorithm that was implemented, describe the implementational choices made and then list the parameters that were used to train the network.

5.1 The Algorithm

In what follows, scalars are denoted by lowercase letters, vectors by boldfaced lowercase letters and matrices by boldfaced uppercase letters. A neural network algorithm consists of three distinct subalgorithms:

- **Forward Propagation**
- **Backward Propagation**
- **Optimization**

Forward Propagation takes a predictor value x and assigns to each value x in x an input unit. In addition a bias unit of value 1 is prepended leading to the input layer l_1.

The "hidden" layer is then computed in the following way: given a number of hidden units, the *activations* a of each hidden layer j are the result of a logistic equation

$$\frac{1}{1 - e^{\Theta^{j-1} a^{j-1}}}$$

where Θ^{j-1} is a matrix of parameters $\theta_{k,n}$ from every unit n in the previous layer $j-1$ to every unit k of j. Note that again a bias unit is prepended to the hidden layer j. Its activation a_0^j is set to be 1. Then unit $k's$ activation value is given by

$$a_k^j = \frac{1}{1 - e^{\theta_{k,0}a_0^{j-1} + \theta_{k,1}a_1^{j-1} + \theta_{k,2}a_1^{j-1} \dots}}$$

Finally, the output layer is computed in the same way as the hidden layer yielding the hypothesis

$$h_\Theta(x) = \frac{1}{1 - e^{\theta_{k,0}a_0^{j} + \theta_{k,1}a_1^{j} + \theta_{k,2}a_1^{j} \dots}}$$

Thus a neural network constitutes a complex non-linear function of the dependent variables.

Backward Propagation aims at improving the predictions of the network. For this purpose a cost function $J(\Theta)$ is computed. The goal will be to minimize the cost incurred by making wrong predictions on a training set. To achieve this, the output "error" will be backwards distributed to the earlier layers, "punishing" those units that contributed most.

The error $\delta^{\mathbf{j+1}}$ at the output level is computed as

$$\delta^{\mathbf{j+1}} = \mathbf{h_\Theta(x)} - \mathbf{y}$$

where y is the independent value corresponding to x. Errors at the hidden levels are given by

$$\delta^{\mathbf{j}} = \mathbf{\Theta^j}^T * a^j * (1 - a^j)$$

where $*$ denotes the Hadamard product. Finally, it can be shown that the gradient of the cost function is given by

$$\frac{\partial}{\mathbf{\Theta}} J(\mathbf{\Theta}) = \frac{1}{m} \sum_{i=1}^{m} \delta_{\mathbf{m}}^{\mathbf{j+1}} \mathbf{a_m^j}^{\mathbf{T}}$$

where m is the number of datapoints in the dataset.

Optimization takes the partial derivatives computed by backward propagation and updates the parameters at each layer as follows

$$\Theta^j := \Theta^j - \alpha \frac{\partial}{\mathbf{\Theta^j}} J(\Theta)$$

The overall algorithm can then be summarized by while #iterations$\leq n$: for $\mathbf{x}$ in data: Set $\mathbf{a^1} = \mathbf{x}$ Perform forward propagation to compute activations $\mathbf{a^j}$ Set $\delta^{j+1} = y$ Perform backward propagation to compute derivatives $\frac{\partial}{\mathbf{\Theta^j}} J(\Theta)$ Perform one step of gradient descent

5.2 Training

The parameters of the neural network were chosen as follows: the network has three layers. Each feature (i.e. keyword) is represented by an input unit, yielding 20 input units. The hidden layer was set to have 40 units. For implementational reasons the number of output layers was chosen to be two although strictly speaking a binary classification task may just as well be implemented with just one output unit.

For the purpose of training the dataset was randomly divided into a dataset of 840 datapoints, a test set and a cross-validation set containing 280 datapoints respectively. For 10000 training iterations, the best prediction results were achieved when no regularization was used, a dynamic learning rate for gradient descent was employed ranging from 0.3 to 30. A gradient check was performed to ensure correct implementation.

5.3 Connection Weight Method

Given the trained neural network, the question is how one assesses the contribution of each feature to the output. For neural networks this is not as easy as for simple regression models but that does not mean that they are the “black boxes” as which they are sometimes treated. One possibility to assess the relative importance of a feature for the output is the connection weight method Olden et al. (2004).

Connection weights are a measure of the overall "stream of information" from an input to an output node. We have seen that the hypothesis in a neural network is computed as

$$h_\Theta(x) = \frac{1}{1 - e^{\theta_{k,0}a_0^j + \theta_{k,1}a_1^j + \theta_{k,2}a_1^j \ldots}}$$

If we denote the sigmoid function by s(x), rewrite the formula in terms of vectors and ignore the bias units, we get

$$h_\Theta = s(\Theta^j s(\Theta^{j-1}\mathbf{x}))$$

Ignoring the sigmoid-functions we end up with

$$h_\Theta \approx \Theta^j \Theta^{j-1}$$

as an approximation of the hypothesis. We call this approximation the *connection weights* of the neural network. In order to get an impression of the scale of these connection weights, we use the sum of all connection weights for a given output unit as "normalization"-factor. The results of applying this method to our trained neural network are displayed in Table 2.

5.4 Results

We have already seen that the high degree of overlap between arithmetic, geometric and Cantorian contexts casts doubt on Mancosu's hypothesis: it gives the impression that Bolzano had a highly integrated view of geometry and arithmetic. It appears thus less likely in light of this statistic that Bolzano might have viewed his late insight on the equinumerosity of sequences in isolation from infinity in geometric contexts.

The results of the neural network training are largely in line with this scepticism. The training resulted in a neural network that was somewhat able to predict Cantorian contexts based on the key words: the trained neural networks puts out no false positives and correctly predicts 25 out of 39 Cantorian contexts in the 840 datapoint training set. For the test set of 280 datapoints, it finds 7 out of 18 Cantorian contexts correctly and incurs 6 false positives. Finally, on the cross-reference set with 280 datapoints of 7 out of 18 Cantorian contexts are identified correctly versus 4 false positives. Overall, this performance is rather weak and indicates that this is not an easy classification task.

Still, the results of the connection weight method should give some insight into the plausibility of Mancosu's hypothesis. However, the picture is very mixed. The four positively most influential key words are "körper", "zeichen", "gegenst", "gerade" - key words from both the geometric and arithmetic domain. Notably, the four least influential key words are all arithmetic while the four negatively msot influential keywords are all geometric. Overall this result does not support Mancosu's hypothesis.

6 Conclusions

In this exercise in computational philosophy a neural network was applied to the question whether Bolzano changed his mind on his conception of the infinite towards the end of his life. The hypothesis that such a change of mind could only have pertained to the arithmetic context was put to the test by the selecting a list of 30 key words pertaining to airthmetic, geometric and Cantorian contexts respectively. Rudimentary statistics of these key words were created from a comprehensive corpus of Bolzano's writings; a neural network was implemented and trained on these statistics. Finally, the hypothesis was put to the test by assessing the relative importance of the key words using the connection weight method.

The result was a negative one. However, this may well be due to the deficiencies of the present study. Future investigations should involve experts in the selection of key words; be based on more sophisticated statistics and employ more advanced methods of neural network training. Still I feel that this kind of investigation is a promising new approach in philosophy that should be pursued further.

Table 2: Connection Weight Method

Connection Weight 0	Connection Weight 1	Key Word	Relative Importances
-45.390	45.392	ziffer	0.020/0.021
56.473	-56.425	arithmet	0.026/0.026
56.729	-56.647	zahl	0.026/0.026
32.875	-32.846	unbenannt	0.015/0.015
-227.605	227.406	konkret	0.105/0.105
-182.559	182.423	vielheit	0.084/0.084
-184.556	184.377	reihe	0.085/0.085
-155.450	155.253	begriff	0.072/0.072
-360.609	360.319	zeichen	0.167/0.167
-359.098	358.820	gegenst	0.166/0.166
-190.875	190.692	geomet	0.088/0.088
-301.234	300.930	gerade	0.139/0.139
199.239	-199.104	linie	0.092/0.092
-218.905	218.668	ebene	0.101/0.101
-475.571	475.217	körper	0.220/0.220
59.308	-59.292	raum	0.027/0.027
170.186	-170.045	entfernung	0.079/0.079
-189.489	189.227	entfernt	0.088/0.088
216.676	-216.522	punkt	0.100/0.100
-63.438	63.369	figur	0.029/0.029

References

Bolzano, B., Terrell, B., and Berg, J. (2012). *Theory of Science: A Selection, with an Introduction*. Synthese Historical Library. Springer Netherlands.

Mancosu, P. (2009). Measuring the size of infinite collections of natural numbers: was cantor's theory of infinite number inevitable? *The Review of Symbolic Logic*, 2(4):612–646.

Olden, J. D., Joy, M. K., and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3):389 – 397.