Number Theory Meets Linguistics: Modelling Noun Pluralisation Across 1497 Languages Using 2-adic Metrics

Gregory Baker and **Diego Molla-Aliod**

Macquarie University

4 Research Park Drive

gregory.baker2@hdr.mq.edu.au and diego.molla-aliod@mq.edu.au

Abstract

A simple machine learning model of pluralisation as a linear regression problem minimising a p-adic metric substantially outperforms even the most robust of Euclidean-space regressors on languages in the Indo-European, Austronesian, Trans New-Guinea, Sino-Tibetan, Nilo-Saharan, Oto-Meanguean and Atlantic-Congo language families. There is insufficient evidence to support modelling distinct noun declensions as a p-adic neighbourhood even in Indo-European languages.

1 Introduction

In this paper, we study whether p-adic metrics are a useful addition to the toolkit of computational linguistics.

It has been known in the mathematical community since 1897 —— although only clearly since (Hensel, 1918) — that there is an unusual and unexpected family of distance metrics based on prime numbers which can be used instead of Euclidean metrics, which have infinitesimals (to support calculus), the triangle inequality (to support geometry), and other useful properties all the while maintaining mathematical consistency. They are known as the p-adic metrics. (Gouvea, 1997) provides a valuable and readable introduction to p-adic analysis.

Given a prime number p it is possible to define a 1-dimensional distance function d as:

$$d_p(r, r) = 0$$

$$d_p(r, q) = \begin{cases} 1 & \text{if } p \nmid (r - q) \\ \dfrac{1}{p} d_p\left(\dfrac{r}{p}, \dfrac{q}{p}\right) & \text{otherwise} \end{cases}$$

(Where $x \nmid y$ means "x does not divide y")

For example, if $p = 3$ then $d_3(1, 4) = \frac{1}{3}$ and $d_3(2, 83) = \frac{1}{81}$.

In particular, if $p = 2$, the authors have found that the 2-adic distance is a surprisingly useful measure for grammar morphology tasks. In many of
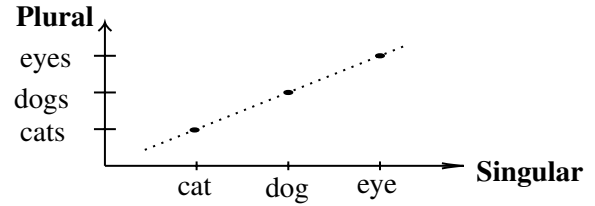

Figure 1: Pluralisation as a linear regression problem with solution $y = 2^{32}x + 116$

the languages in this study we found that identifying the grammar rules for pluralisation turned into a problem of finding a linear regressor which minimised a p-adic metric.

2 Pluralisation as linear regression

In this paper we use a simple and naive approach for converting vocabulary words into vectors: use whatever the unicode bit sequence for the word would be; this bit sequence can also be viewed as an integer vector with one element. This is of course extremely arbitrary and subject to the whims of the unicode consortium, but it is the most common way to represent text from any human language on a computer.

Note that in this naive encoding scheme words like "sky", "fry" and "butterfly" are very close using a 2-adic metric — the last 32 bits are the same, meaning that the distance between them is less than or equal to than 2^{-32}. Using a Euclidean metric "butterfly" is at least $\left(2^{32}\right)^6 = 2^{192}$ apart from the other two words. A little exploration will observe that noun declensions in many languages — especially ones in the Indo-European family — have this property that they consist of words that form tight 2-adic clusters.

This odd correspondence between 2-adic geometry and grammar morphology extends to declension rules for case and number where they exist. Consider that the first two rules in Figure 2 have the property that in the naive UTF-32 encoding they

1. If the singular form ends in "y", replace the "y" with "ies".

2. For singulars ending in "o" or "i" or "ss" append "es".

3. There are irregular nouns: "person" $\mapsto$ "people", "sheep" $\mapsto$ "sheep"

4. If no other rule applies, append "s".

Figure 2: A simplified and incomplete set of rules for forming plurals in English

Algorithm	Neigh-bourhood Metric	Number of neigh-bours	Regr-essor
Global p-adic	N/A	N/A	p-adic
Global Siegel	N/A	N/A	Siegel
Local p-adic	p-adic	$3\ldots20$	p-adic
Local Siegel	Euclidean	$3\ldots20$	Siegel
Hybrid Siegel	p-adic	$3\ldots20$	Siegel

Table 1: Enumeration of algorithms and configurations tested, as discussed in Section 3.

can all be accurately modelled using a linear regression performed on points in the local 2-adic neighbourhood. The fourth rule is illustrated in Figure 1, with singulars and plurals of "cat", "dog" and "eye" plotted. They lie on the straight line $y = 2^{32}x + 116$.

2.1 Mathematical Challenges

Unfortunately, finding the line through a set of points that minimises the sum of the p-adic measure of the residuals is harder than finding the line that minimises the sum of the square of the residuals. Having chosen a prime p, the formulation looks similar: given a set of points $\{(x_i, y_i), i \in \{1\ldots N\}\}$, find m and b to minimise $f(m,b) = \sum_{i=1}^{N} |y_i - (mx_i + b)|_p$ where $|\cdots|_p$ is the p-adic measure described in section 1. But, there is no guarantee that there is a unique (m,b) that minimises f. Consider the data set $\{(0,0),(1,0),(1,1),(1,2),(1,3)\}$. The 2-adic sum of distances from those points is $\frac{5}{2}$ for $y = 0, y = x, y = 2x$ and $y = 3x$.

The derivatives of f with respect to m and b are also unhelpful: there are an infinite number of inflection points for any non-trivial data set.

Fortunately, it is possible to prove that the p-adic line of best fit — unlike the Euclidean line of best fit — must pass through two of the data points[1], which at least provides an $O(n^3)$ algorithm for finding optimal (m,b) values: draw a line through every pair of points and try them all. The proof is in Appendix A.

2.2 Data

The dataset of singular and plural forms we used in this research is the LEAFTOP dataset, as described in (Baker and Molla-Aliod, 2022). This consists

of singular and plural noun pairs from Bible translations in 1,480 languages[2] grouped by language family using the union of the Ethnologue (Eberhard et al., 2021) and Glottolog (Hammarström et al., 2021). Since they differ on the world's primary language families, and not every language can or should be assigned to a language family[3], there are overlaps and gaps in the LEAFTOP language families that are reflected in the results of this research.

For many languages in our data set[4] we believe no language morphology task has ever been run, and we thus set a baseline for these languages.

3 Experiment

The aim of this research is to identify whether or not using a p-adic metric space is likely to generate improvements on computational linguistics tasks.

A linear model will obviously not be able to capture irregular nouns. The 2-adic neighbourhood will not capture nouns that belong to different noun declensions but share the same ending. Comparing a linear regression model (even if it is operating over an unusual space) to a million-parameter neural network[5] where such subtleties can be captured is going to be uninformative in telling us about the usefulness of p-adic metrics. As a result we are comparing p-adic linear regression against methods that are clearly not the state-of-the-art, but are methods which can be legitimately compared.

---

[1]In this way, the p-adic line of best fit is similar to the line of best fit supplied by the Theil-Sen, Siegel or RANSAC algorithms.

[2]Section 4 reports results on 1,497 languages. In the LEAFTOP dataset, a language which has multiple orthographies is counted as one language (e.g. Chadian Arabic can also be written in a Roman alphabet), where in this paper each orthography has been counted as a separate language. Languages with significant geographic variations (such as Spanish or Portuguese) are also considered one language by LEAFTOP, and as multiple in this paper.

[3]Klingon, for example.

[4]Very little computational linguistics has been run on the Trans-New Guinea family of languages, for example.

[5]Assuming that there were computational resources and data available to perform this task on thousands of low-resource languages.
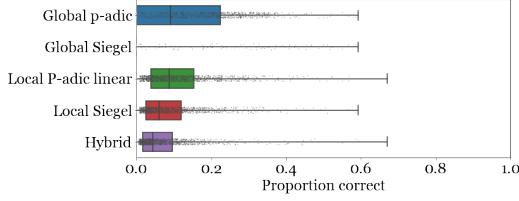
Figure 3: Strip and box plot of the proportions correct for each algorithm

family	Are p-adics better? Probability that we saw p-adics doing 'better' by chance (Bonferroni adjusted)			
	Global p-adic vs Global Siegel	Local P-adic linear vs Local Siegel	Local P-adic linear vs Hybrid	Hybrid vs Local Siegel
Indo-European	4.40e-04	1.34e-05	2.79e-02	5.55e-01
Austronesian	1.04e-25	1.38e-11	7.61e-32	1.00e+00
Trans New Guinea	1.58e-06	9.99e-01	3.37e-06	1.00e+00
Sino-Tibetan	2.52e-09	1.39e-01	4.57e-09	1.00e+00
Kra-Dai	1.00e+00	1.00e+00	1.00e+00	1.00e+00
Niger-Congo	1.15e-37	4.08e-21	1.85e-36	1.00e+00
Australian aboriginal	9.99e-01	1.00e+00	8.61e-01	1.00e+00
Afro-Asiatic	1.90e-01	9.92e-01	9.03e-01	1.00e+00
Nilo-Saharan	2.84e-03	2.05e-02	1.37e-05	1.00e+00
Oto-Manguean	1.94e-08	8.97e-01	1.25e-09	1.00e+00
Austroasiatic	5.58e-02	1.00e+00	1.22e-03	1.00e+00
Dravidian	8.29e-01	9.36e-01	1.00e+00	1.00e+00
Tupian	1.00e+00	1.00e+00	1.00e+00	1.00e+00
Atlantic-Congo	1.69e-35	2.90e-20	8.84e-35	1.00e+00
Pama-Nyungan	9.99e-01	1.00e+00	8.61e-01	1.00e+00
Arawakn	7.62e-02	8.79e-01	3.70e-01	1.00e+00
Mande	6.70e-01	1.00e+00	9.85e-01	1.00e+00
Unclassified	3.98e-56	4.00e-24	6.75e-63	1.00e+00
Unrecorded	9.99e-01	5.87e-01	3.71e-01	1.00e+00
All languages	3.13e-160	1.70e-67	1.11e-170	1.00e+00

Table 2: Experimental Results. Lighter colours indicate stronger statistical significance.

The choice of the Siegel regressor (Siegel, 1982) as the representative for Euclidean regression was forced by the need for robustness to a large number of outliers. The LEAFTOP data set is known to be only 72% accurate and any irregular nouns will also be outliers. Huber 1964, Theil-Sen 1950 and ordinary least squares regression are all ruled out by these criteria.

The Siegel and p-adic regressors were run in "global" mode (learn from as many examples as possible) and "local" mode (learning from a small number of nearby words). To identify the impact of the p-adic neighbourhood vs the impact of the p-adic linear regressor, local Siegel was run twice, once with a p-adic (a "hybrid" of a Euclidean regressor and a p-adic neighbourhood) and once with a Euclidean neighbourhood (labeled "local Siegel"). The complete set of algorithms and their configurations is listed in Table 1.

The only metric that can be used for this comparison is L0 — accuracy — since any other metric (e.g. L1 or L2 norms) will bias the results towards the metric space that they operate in. A leave-one-out cross validation was done for each algorithm for each language.

4 Results

A plot of results by algorithm is in Figure 3. Summary statistics for each language family and algorithm combination are shown in Table 3.

In all language families (and overall across all languages), p-adic approaches outperformed Euclidean ones, however the results were not all statistically significant. The differences in performance between algorithms on a language do not follow a normal distribution. Since the research question is simply "which is better?" the magnitude of the effect is unimportant, and a Wilcoxon signed-rank test can be used. The Pratt method was used for handling situations where the scores were identical and no sign can be calculated. The probability is

that of a one-sided result.

There are 80 statistical tests required to perform to confirm validity. There are 17 languages families in the Ethnologue and Glottolog plus another 3 pseudo-families from the LEAFTOP labelling (Unclassifed, Unrecorded and All). For each of these 20 families, there are 4 tests: global p-adic vs global Siegel; local p-adic vs local Siegel; local p-adic vs Siegel using a p-adic neighbourhood; Siegel with a Euclidean neighbourhood vs a p-adic neighbourhood. The correction to apply to the raw statistical test results is therefore $p \mapsto 1-(1-p)^{80}$. It is this latter (corrected) number[6] that is reported in Table 2.

There is strong evidence that noun pluralisation in languages in the Indo-European, Austronesian, Trans New Guinea, Sino-Tibetan, Niger-Congo, Nilo-Saharan, Oto-Meanguean and Atlantic-Congo families can be modelled better with p-adic linear regression than with Euclidean. This is also true for the unclassified languages in the LEAFTOP dataset.

Moreover, the data in Table 2 also support the hypothesis that a randomly chosen human language will model better using p-adic linear regression than Euclidean.

---

[6]For example, the test result for probability that global p-adic regression is equivalent to global Euclidean Siegel on Afro-Asiatic languages is 0.00263 — which would have been a very clear result! — but with 80 experiments, we would expect to see some low-probability results. Thus the probability of seeing a result as extreme as we saw for *at least one of the 80 experiments* by chance is much higher: 0.23.

Average proportion correct (+/- stddev) using each algorithm by language family

	Global p-adic	Global Siegel	Hybrid	Local Siegel	Local P-adic linear
Indo-European	0.105+/-0.139	0.056+/-0.136	0.135+/-0.126	0.109+/-0.117	0.154+/-0.126
Austronesian	0.203+/-0.141	0.038+/-0.118	0.103+/-0.084	0.142+/-0.102	0.172+/-0.103
Trans New Guinea	0.124+/-0.109	0.011+/-0.057	0.063+/-0.052	0.084+/-0.055	0.092+/-0.061
Sino-Tibetan	0.174+/-0.131	0.027+/-0.087	0.087+/-0.077	0.112+/-0.092	0.130+/-0.097
Kra-Dai	0.397+/-0.028	0.000+/-0.000	0.224+/-0.050	0.293+/-0.031	0.309+/-0.042
Niger-Congo	0.093+/-0.089	0.005+/-0.037	0.033+/-0.042	0.048+/-0.054	0.077+/-0.071
Australian aboriginal	0.033+/-0.071	0.000+/-0.000	0.027+/-0.030	0.030+/-0.043	0.038+/-0.038
Afro-Asiatic	0.039+/-0.077	0.000+/-0.002	0.037+/-0.042	0.040+/-0.035	0.051+/-0.051
Nilo-Saharan	0.076+/-0.105	0.000+/-0.002	0.039+/-0.043	0.054+/-0.057	0.078+/-0.062
Oto-Meanguean	0.145+/-0.124	0.003+/-0.019	0.070+/-0.044	0.112+/-0.060	0.125+/-0.067
Austroasiatic	0.294+/-0.137	0.086+/-0.186	0.151+/-0.107	0.227+/-0.108	0.244+/-0.118
Dravidian	0.082+/-0.087	0.009+/-0.030	0.090+/-0.063	0.072+/-0.048	0.106+/-0.072
Tupian	0.019+/-0.027	0.000+/-0.000	0.032+/-0.045	0.015+/-0.003	0.047+/-0.041
Atlantic-Congo	0.095+/-0.088	0.005+/-0.038	0.034+/-0.043	0.049+/-0.055	0.080+/-0.072
Pama-Nyungan	0.033+/-0.071	0.000+/-0.000	0.027+/-0.030	0.030+/-0.043	0.038+/-0.038
Arawakn	0.077+/-0.097	0.004+/-0.016	0.047+/-0.042	0.062+/-0.053	0.079+/-0.065
Mande	0.102+/-0.114	0.012+/-0.031	0.033+/-0.034	0.041+/-0.044	0.054+/-0.058
Unclassified	0.104+/-0.122	0.010+/-0.056	0.059+/-0.067	0.074+/-0.075	0.095+/-0.086
Unrecorded	0.092+/-0.142	0.061+/-0.149	0.092+/-0.147	0.083+/-0.129	0.115+/-0.141
All languages	0.121+/-0.127	0.016+/-0.075	0.067+/-0.075	0.085+/-0.085	0.109+/-0.094

Table 3: Average proportion correct for each combination of language family and algorithm. Darker values indicate higher accuracy.

Language family	Bonferroni-adjusted p-value of test
Austronesian	$2.39 * 10^{-6}$
Trans New Guinea	0.032
Sino-Tibetan	$1.92 * 10^{-5}$
Niger-Congo	$8.76 * 10^{-7}$
Atlantic-Congo	$2.44 * 10^{-6}$
Unclassified	0.0048
All languages	$2.69 * 10^{-13}$

Table 4: p-values of Wilcoxon tests for global p-adic regression versus local regression

4.1 How much does a p-adic neighbourhood pre-filter help?

There are many language families where training on the vocabulary in the p-adic neighbourhood produced a better average correctness score: Indo-European, Afro-Asiatic, Nilo-Saharan, Dravidian, Tupian and Arawakan. Because of the discrepancies between the Ethnologue and Glottolog on the categorisation of Australian languages, it appears that there are two other language familiies ("Australian aboriginal" and "Pama-Nyungan") where p-adic neighbourhoods are useful for predicting the plural of a word. In addition, languages where LEAFTOP has no language family information ("Unrecorded") also appear to benefit from p-adic neighbourhoods.

Unfortunately, none of these results hold up. The raw p-value of the Wilcoxon test comparing global versus local p-adic methods on Indo-European languages is $5.98 * 10^{-3}$, but given that there are 9 tests to perform, the Bonferroni adjustment tells us that the probability of seeing a result like that is 0.053. Close, but not compelling proof. None of the other language families passed significance testing either.

Turning it around, and looking at the other 11 language families (including "All" and "Unclassified"), 7 of these show a statistically significant difference between the local and global versions of p-adic linear regression. P-values for these experimental results are in Table 4. This can be interpreted to mean that either these language families do not generally have noun declensions, or that using p-adic distance is a poor way of separating those noun declensions.

Note also that the Hybrid algorithm (Siegel regressor trained on a p-adic neighbourhood) also underperforms a Euclidean-trained Siegel regressor.

5 Related Work

Murtagh (e.g. his overview paper Murtagh, 2014) and Bradley (e.g. Bradley, 2009, Bradley, 2008) have written the most on p-adic metrics in machine learning, having explored clustering and support vector machines in some depth. (Khrennikov and Tirozzi, 2000) provides an algorithm for training a neural network. An extensive literature search has failed to find any other p-adic adaptions of traditional machine learning algorithms. This paper is the first to discuss p-adic linear regression.

Expanding the literature search more broadly, we find that there have been very few side-by-side comparisons of Euclidean metrics versus strongly mathematically-formulated non-Euclidean metrics for tasks in computational linguistics.

(Nickel and Kiela, 2017), (Tifrea et al., 2018) and (Saxena et al., 2022) performed their learning of word embeddings on a non-Euclidean metric, choosing a Poincaré hyperbolic space. Calculating derivatives and finding minima of a function in a Poincaré space is substantially more complex both mathematically and computationally than for a Euclidean space. p-adics are simpler in both regards, but give rise to a space with similar hyperbolic properties. We believe that this may be a fruitful area of future research.

6 Conclusion

We demonstrated superiority over Euclidean methods on languages in the Indo-European, Austronesian, Trans New-Guinea, Sino-Tibetan, Nilo-Saharan and Oto-Meanguean and Atlantic-Congo

Algorithm	Seconds per run	Total runs	Approx CPU days
Global p-adic	8814.6	8643	881.8
Global Siegel	32.7	8643	3.3
Local Siegel	0.368	155574	0.66
Local p-adic	10.1	155574	18.2
Hybrid Siegel	0.398	155574	0.72

Table 5: Computation time

language families.

Based on this, we expect that substituting p-adic metrics for Euclidean metrics in other computational linguistics tasks and machine learning methods may be an exciting area of research.

Acknowledgements

References

Gregory Baker and Diego Molla-Aliod. 2022. The construction and evaluation of the leaftop dataset of automatically extracted nouns in 1480 languages. *Processings of the 13th Language Resources and Evaluation Conference, LREC 2022, Marseille, France*.

Patrick Erik Bradley. 2008. Degenerating families of dendrograms. *Journal of Classification*, 25:27–42.

Patrick Erik Bradley. 2009. On p-adic classification. *p-Adic Numbers, Ultrametric Analysis, and Applications*, 1:271–285.

David Eberhard, Gary Simons, and Charles Fennig. Ethnologue: Languages of the world [online]. 2021.

Fernando Q. Gouvea. 1997. *Foundations*, chapter 2. Springer.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. Glottolog 4.5. Max Planck Institute for Evolutionary Anthropology.

Kurt Hensel. 1918. Eine neue theorie der algebraischen zahlen. *Math Z*, 2:433–452.

Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Andrei Khrennikov and Brunello Tirozzi. 2000. Learning of p-adic neural networks. *Can. Math. Soc. Proc. Ser*, 29:395–401.

Fionn Murtagh. 2014. Thinking ultrametrically. In *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, pages 3–14. Springer Science and Business Media.

NAACL [online]. 2021. [link].

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30.

Chandni Saxena, Mudit Chaudhary, and Helen Meng. 2022. Cross-lingual word embeddings in hyperbolic space. *arXiv preprint arXiv:2205.01907*.

Andrew F. Siegel. 1982. Robust regression using repeated medians. *Biometrika*, 69(1):242–244.

H Theil. 1950. A rank-invariant method of linear and polynomial regression analysis. *Nederlandse Akademie van Wetenschappen*, pages 386–392.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincaré glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*.

1. $\forall i, j, i \neq j, \frac{y_i - y_j}{x_i - x_j} \in \mathbb{Z}$

2. It contains the origin $(x_0, y_0) = (0, 0)$ and one of the optimal lines of best fit passes through the origin, and can therefore be written as $y = mx$

3. $i \neq 0 \Rightarrow x_i \neq 0$

4. The data set is sorted such that $\left| \frac{y_1 - mx_1}{x_i} \right|_p \leq \left| \frac{y_i - mx_i}{x_i} \right|_p$ for all i where $i > 1$

Table 6: Constraints on the data set for the proof in subsection A.2

A Proof that the p-adic line of best fit passes through at least two points in the dataset

The proof is in three sections:

1. A proof that a p-adic line of best fit must pass through at least one point. (Subsection A.1).

2. A proof that for a data set with some strong restrictions, that if a p-adic line of best fit passes through one particular point in a dataset that it must pass through a second point. (Subsection A.2).

3. A set of short proofs that every data set which doesn't satisfy those restrictions is related to a data set which does satisfy them, and that the p-adic lines of best fit can be calculated directly from them.

The phrase "optimal line" will be used to mean "one of the set of lines whose p-adic residual sum is equal to the minimum residual sum of any line through that data set".

The notation $\mathrm{Res}_p(\{(x_i, y_i)\}, y = mx + b)$ will be used for "the sum of the p-adic residuals of the line $y = mx + b$ on the set $\{(x_i, y_i)\}$.

A.1 p-adic best-fit lines must pass through one point

Proof. Suppose that there exists one or more lines that are optimal for a given data set of size s, and suppose further that none of these lines passes though any point in the data set.

Let one of these optimal lines be $y = mx + b$.

Order the points (x_i, y_i), in the dataset by their residuals (smallest first) for this line:

$$|y_i - \hat{y}_i|_p \leqslant |y_{i+1} - \hat{y}_{i+1}|_p$$

Since $y = mx + b$ does not pass through any point in the dataset, $|\hat{y}_0 - y_0|_p > 0$, and we can write the residual $|\hat{y}_0 - y_0|_p$ as ap^n for some non-zero value of a (satisfying $|a|_p = 1$) and some value (possibly zero) of n. The ordering criteria means that $|ap^n| \leq |y_i - \hat{y}_i|_p$ for all i.

Consider the line $y = mx + b - ap^n$. Its residual sum is

$$\mathrm{Res}_p(\{(x_i, y_i)\}, y = mx + b - ap^n)$$
$$= \sum_{i=0}^{s} |\hat{y}_i - ap^n - y_i|_p$$
$$= |\hat{y}_0 - ap^n - y_0|_p + \sum_{i=1}^{s} |\hat{y}_i - ap^n - y_i|_p$$
$$= 0 + \sum_{i=1}^{s} |\hat{y}_i - ap^n - y_i|_p$$
$$\leq \sum_{i=1}^{s} \max(|\hat{y}_i - y_i|_p, |ap^n|_p)$$
$$= \sum_{i=1}^{s} |\hat{y}_i - y_i|_p$$
$$< \sum_{i=0}^{s} |\hat{y}_i - y_i|_p$$
$$= \mathrm{Res}_p(\{(x_i, y_i)\}, y = mx + b)$$

As this final line is the residual sum for the line $y = mx + b$, and the first line is strictly less than the final, $y = mx + b - ap^n$ is a more optimal line than $y = mx + b$, contradicting the premise. $\square$

A.2 p-adic best-fit lines must pass through two points

Consider a data set $\{(x_i, y_i)\}$ of size s with the properties listed in Table 6. Then the chosen optimal line which passes through the origin also passes through another point in the dataset.

Proof. Suppose that the chosen optimal line passes through only one point in the data set.

Let $m' = m + \frac{y_1 - mx_1}{x_1}$ and consider the residual sum of the line $y = m'x$ (which passes through both (x_0, y_0) and (x_1, y_1)).

$$\text{Res}_p(\{(x_i, y_i)\}, y = m'x)$$

$$= \sum_{i=0}^{s} \left|(m + \frac{y_1 - mx_1}{x_1})x_i - y_i\right|_p$$

$$= |0| + \left|(m + \frac{y_1 - mx_1}{x_1})x_1 - y_i\right|_p$$

$$\quad + \sum_{i=2}^{s} \left|(m + \frac{y_1 - mx_1}{x_1})x_i - y_i\right|_p$$

$$= |mx_1 + y_1 - mx_1 - y_i|_p$$

$$\quad + \sum_{i=2}^{s} \left|(m + \frac{y_1 - mx_1}{x_1})x_i - y_i\right|_p$$

$$= 0 + \sum_{i=2}^{s} \left|(m + \frac{y_1 - mx_1}{x_1})x_i - y_i\right|_p$$

$$= \sum_{i=2}^{s} \left|mx_i - y_i + \frac{y_1 - mx_1}{x_1})x_i\right|_p$$

$$\leq \sum_{i=2}^{s} \max(|mx_i - y_i|_p, \left|\frac{y_1 - mx_1}{x_1}x_i\right|_p)$$

$$= \sum_{i=2}^{s} \max(|mx_i - y_i|_p, \left|\frac{y_1 - mx_1}{x_1}\right|_p \cdot |x_i|_p)$$

$$\leq \sum_{i=2}^{s} \max(|mx_i - y_i|_p, \left|\frac{y_i - mx_i}{x_i}\right|_p \cdot |x_i|_p)$$

$$= \sum_{i=2}^{s} \max(|mx_i - y_i|_p, |mx_i - y_i|_p)$$

$$= \sum_{i=2}^{s} |mx_i - y_i|_p$$

$$< 0 + |y_1 - mx_1|_p + \sum_{i=2}^{s} |mx_i - y_i|_p$$

$$= \sum_{i=0}^{s} |mx_i - y_i|_p$$

$$= \text{Res}_p(\{(x_i, y_i)\}, y = mx)$$

The last term is the residual sum from the line $y = mx$ (a line which was supposed to be optimal for the data set), which is strictly larger than the residual sum from $y = m'x$. This contradicts the premise. $\quad\square$

A.3 Loosening the criteria

This subsection loosens the criteria of the proof in subsection A.2.

The first three arguments (and the last half of the fourth argument) have a common structure.

They start with a data set of points D and find a way of taking an arbitrary linear function f and performing a non-singular (invertible) linear transformation to turn them into a set D' and f' where the residuals of the two functions are also invertibly linearly transformed, with the transformation coefficients solely based on the contents of D.

That is, there will be a set-transformation function of the form $T_d(x, y) = (t_0 x + t_1, t_2 y + t_3)$, a function transformation $T_f(f) : T_f(f(x, y)) = f(t_4 x + t_5, t_6 y + t_7)$, and a residual transformation $T_r(\text{Res}_p(D, f)) = \text{Res}_p(D', f') = t_8\text{Res}_p(D, f)) + t_9$. The coefficients $t_0 \ldots t_9$ are dependent only on D, and $t_0, t_2, t_4, t_6 + t_8$ are all non-zero.

Thus, if a line f is optimal for D, then the line f' will be optimal for D' and vice versa. As a result, the interesting property of the optimal line f' of D' (that f' must pass through two points in D' if it is optimal) will also apply to D and f.

Scaling of y. Given two datasets, $D = \{(x_i, y_i)\}$ and $D' = \{(x_i, \alpha y_i)\}$ and a line $y = mx + b$ with a residual r on D, there is another line $y = \alpha mx + \alpha b$ with a residual $|\alpha|_p r$ on D' (and vice versa). This is a straightforward consequence of factorisation:

$$\text{Res}_p(\{(x_i, \alpha y_i)\}, y = \alpha mx + \alpha b)$$

$$= \sum_i |\alpha mx_i + \alpha b - (\alpha y_i)|_p$$

$$= |\alpha|_p \cdot \sum_i |mx_i + b - y_i|_p$$

$$= |\alpha|_p \text{Res}_p(\{(x_i, y_i)\}, y = mx + b)$$

$\quad\square$

Scaling of x. Likewise, there are relationships between data sets with scaled x values. If $D = \{(x_i, y_i)\}$ and $D' = \{(\alpha x_i, y_i)\}$, then the residual of the line $y = mx + b$ on D is the same as the residual of the line $y = \frac{m}{\alpha}x + b$ on D'.

$$\text{Res}_p(\{(\alpha x_i, y_i)\}, y = \frac{m}{\alpha}x + b)$$

$$= \sum_i \left| \frac{m}{\alpha}(\alpha x_i) + b - y_i \right|_p$$

$$= \sum_i |mx_i + b - y_i|_p$$

$$= \text{Res}_p(\{(x_i, y_i)\}, y = mx + b)$$

$\square$

Therefore, a data set having some rational (non-integer) coefficients can be transformed into a data set with integral coefficients where the optimal lines are similarly transformed with only a constant multiplier effect on each residual sum simply by multiplying through by the product of all denominators.

Moreover, if $D = \{(x_i, y_i)\}$ has integer coordinates, then $D' = \{\alpha x_i, y_i)\}$ where α is the product $\prod_{j,k,j<k}(u_j v_k - u_k v_j)$ will not only have integer coordinates, but every line between two points in D' will have an integer gradient (and therefore an integer y-intercept).

This generalises the result from subsection A.2 even when condition (1) from Table 6 is not satisfied.

Translation in the plane. Similar mechanisms apply for translation by a fixed offset in the (x, y) plane: by adding a constant to all x or y values. Given $D = \{(x_i, y_i)\}$ and $D' = \{(x_i + a, y_i + c)\}$, the line $y = mx + b$ has the same residual sum on D as $y = mx + (b + c - ma)$ does on D'.

$$\text{Res}_p(\{(x_i + a, y_i + c)\}, y = mx + (b + c - ma))$$

$$= \sum_i |m(x_i + a) + (b + c - ma) - (y_i + c)|_p$$

$$= |mx_i + b - y_i|_p$$

$$= \text{Res}_p(\{(x_i, y_i)\}, y = mx + b)$$

$\square$

This generalises the result from subsection A.2 to cover data sets where condition (2) from Table 6 is not satisfied.

When $x_i = 0$ for some or all i. If condition (3) from Table 6 is violated, then there are two sub-cases to handle.

Firstly, if $x_i = 0$ for all i then the optimal line is a vertical line along the y-axis, which has the property of passing through two points in the data set.

Alternatively, if $x_i \neq 0$ for some i, then define Z as being the set of points of D where $x_i = 0$, and $D' = (D \setminus Z) \cup (0, 0)$ where $\setminus$ is the set difference operator.

Then for any function $f(x)$ defined as $y = mx + b$,

$$\text{Res}_p(D, f) = \text{Res}_p(D', f) + \text{Res}_p(Z, f)$$

$$= \text{Res}_p(D', f) + \sum_{z \in Z} b - y_z$$

The last term is a constant that only depends on the elements of D, not f, thus defining an invertible linear transformation between the residuals. $\square$

Condition (4) from Table 6 can be achieved by sorting the dataset.

B NAACL Reproducibility Checklist

This appendix responds to the request for reproducibility from (NAACL, 2021).

NAACL requirements are shown in a **bold font**. **For all reported experimental results:**

- **A clear description of the mathematical setting, algorithm, and/or model** Details in section 2.

- **A link to a downloadable source code, with specification of all dependencies, including external libraries** https://github.com/solresol/thousand-language-morphology and https://github.com/solresol/padiclinear

- **A description of computing infrastructure used** A little over half the computation was run on a 48-cpu node in the Gadi supercomputing facility. The remainder was done on Arm64 virtual machines running Ubuntu 21.10 at Amazon, the author's M1 Macbook Air and the author's x64-based Ubuntu 22.10 Linux system.

- **The average runtime for each model or algorithm, or estimated energy cost** On the

author's x64-based Ubuntu system (where it was possible to guarantee no contention), the average run times are given in Table 5.

- **The number of parameters in each model** Global P-adic and Global Siegel have no parameters. Local Siegel, Local P-adic Linear and Hybrid have one parameter: the number of neighbours to include in the training set.

- **Corresponding validation performance for each reported test result** There are not separate validation and test sets in this paper.

- **A clear definition of the specific evaluation measure or statistics used to report results.** As discussed in section 3, the only metric which can be used is accuracy.

For all results involving multiple experiments, such as hyperparameter search:

- **The exact number of training and evaluation runs** For the Local Siegel, Local P-adic Linear and Hybrid algorithms, 18 different neighbourhoods were explored.

- **The bounds for each hyperparameter** Minimum 3, maximum 20. Anything below 3 makes no sense, and with an $O(n^3)$ algorithm, growing beyond 20 starts to become computationally infeasible.

- **The hyperparameter configurations for best-performing models** Attached as a data file.

- **The method of choosing hyperparameter values (e.g. manual tuning, uniform sampling, etc.) and the criterion used to select among them (e.g. accuracy)** There was no need for hyperparameter selection as it was possible to cover the entire solution space.

- **Summary statistics of the results (e.g. mean, variance, error bars, etc.)** Detailed in section 4

Answers about all datasets used: See (Baker and Molla-Aliod, 2022) — https://github.com/solresol/leaftop