A Natural Law of Succession

Eric Sven Ristad¹ Mnemonic Technology, Inc. Princeton, NJ 08540 USA ristad@mnemonic.com

Abstract — We present a new solution to multinomial estimation and demonstrate that our solution outperforms standard solutions both in theory and in practice. The novelty of our approach lies in our use of combinatorial priors on strings.

I. NATURAL STRINGS

An alphabet represents the set of logically possible events. In this world, all strings are finite and most are very short. For this basic reason, natural strings do not include all the symbols in the alphabet. This claim is tautological for short strings, but it is also true for long strings.

To model this phenomenon, we propose a uniform prior on the cardinalities of all nonempty subsets of the alphabet. Such a prior on an alphabet of size k entails the probability

$$p_N(x^n | n) = \left[\min(k, n) \begin{pmatrix} k \\ q \end{pmatrix} \begin{pmatrix} n-1 \\ q-1 \end{pmatrix} \begin{pmatrix} n \\ \{n_i\} \end{pmatrix}\right]^{-1}$$

for strings x^n of length n with cardinality q.

This probability is not Kolmogorov compatible. To obtain a conditional probability, we must use $p(i|x^n, n+1)$ instead of the more obvious $p(i|x^n, n)$. Algebraic manipulation yields the following conditional, which we call the natural law,

$$p_N(i|\{n_i\}, n) = \begin{cases} (n_i+1)/(n+k) & q=k\\ (n_i+1)(n+1-q)/(n^2+n+2q) & q< k \land n_i > 0\\ q(q+1)/(k-q)(n^2+n+2q) & \text{otherwise} \end{cases}$$

where n_i is the frequency of the i^{th} symbol. The natural law reduces to Laplace's law (1775) on the attested symbols.

Unlike Laplace's law, or its popular generalization $p_{\lambda}(i|\{n_i\}, n) = (n_i + \lambda)/(n + k\lambda)$, the amount of probability $q(q+1)/(n^2 + n + 2q)$ assigned to novel events by our natural law decreases quadratically in the number n of trials. More importantly, the probability of novel events is independent of the alphabet size k. There is no penalty for large alphabets.

II. WASTED PROBABILITY

Let A be the universe of possible symbols and let B be the actual alphabet of the source, $B \subset A$, and |B| = b.

The total probability assigned to B^n by $p_{\lambda}()$ is

$$p_{\lambda}(B^{n}|n) \approx \Theta(\left(\frac{1}{n-1}\right)^{\lambda(k-b)})$$

Unless $\lambda = 0$ and b = 1, $p_{\lambda}(B^n|n)$ rapidly approaches zero. This profound flaw is typically disguised in the literature by analyzing convergence to an underlying source. Our analysis shows that such convergence is at best a feeble optimality. For the natural law,

$$p_N(B^n|n) = \frac{b!(k-b)!}{(k+1)!}$$

which is minimum at b = k/2. At that point, $p_N(B^n|n) > 2^{-k/2}$ irrespective of n. Thus, the natural law without prior knowledge of B performs at most a constant factor worse than *any* other probability function with prior knowledge of B.

III. RATIO OF ESTIMATES

The penalty for using $p_{\lambda}()$ instead of the natural law $p_N()$ can grow without bound:

$$p_N(x^n|n)/p_\lambda(x^n|n) = \Theta(\lambda^{(1-k)/2}n^{k\lambda-q}\prod_{i=1}^k n_i^{1-\lambda})$$

The λ^{-1} term is a reminder that $\lambda = 0$ will result in an infinite advantage for the natural law when q > 1. This ratio depends on the parameter λ and the observed $\{n_i\}$.

For the widely advocated $\lambda = 1/2$ [2, 3], the penalty grows without bound when $q \leq k/2$. The case q > k/2 depends on the symbol frequencies. The natural law will loose only in the unnatural case of a large q and a low empirical entropy.

IV. EXPERIMENTS

The prediction of naturally-occurring sequences poses a difficult test for all multinomial estimators. The task is to predict each symbol in a sequence on the basis of the frequencies of the preceding symbols. Our benchmark is the Calgary corpus, which includes a wide range of ASCII as well as non-ASCII files [1]. Each file is generated by a different source and represents a distinct prediction problem.

All prediction results are in whole bytes, relative to the empirical entropy $nH\{n_i/n\}$. Each byte represents a factor of 256 in probability.

The natural law assigns 256^{1106} times more probability to the Calgary corpus than Laplace's law $p_L()$, and 256^{903} times more than the popular $p_{\frac{1}{2}}()$. The natural law also outperforms Methods A-D from the text compression community [4].

References

- BELL, T. C., CLEARY, J. G., AND WITTEN, I. H. Text Compression. Prentice Hall, Englewood Cliffs, NJ, 1990.
- [2] JEFFREYS, H. An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. (London) A 186 (1946), 453-461.
- [3] KRICHEVSKII, R. E., AND TROFIMOV, V. K. The performance of universal coding. *IEEE Trans. Inform. Theory IT-27*, 2 (1981), 199-207.
- [4] RISTAD, E. S. A natural law of succession. Tech. Rep. CS-TR-495-95, Department of Computer Science, Princeton University, Princeton, NJ, May 1995.

¹This work was first presented on May 15, 1995 at Johns Hopkins University [4]. It was partially supported by NSF Young Investigator Award IRI-9258517 to the author.