Quantized Indexing: Beyond Arithmetic Coding

Ratko V. Tomic
1stWorks Corporation (www.1stWorks.com)

*Quantized Indexing (QI) is a fast and space-efficient form of enumerative (combinatorial) coding [1] (EC). The arithmetic precision, execution time, table sizes and coding delay of EC are all reduced by a factor O(**n**) at a redundancy below* $\log(e)/2^{g-1}$ *bits/symbol (for **n** input symbols and a g-bit QI precision). Due to its tighter enumeration, QI redundancy is below that of arithmetic coding (AC). The relative compression gains increase for shorter outputs and for less predictable data. Simultaneously, QI is significantly faster than the fastest arithmetic coders [2], from factor 6 in high to over 200 in low entropy limit ('typically' 10-20 times faster). These speedups are the result of using no coding operations on more probable symbol, fewer and simpler operations on less probable symbol and intrinsically cleaner division of labor between the coder and modeler.*

QI solves optimally [3] the four decade old problem of unlimited arithmetic precision of exact EC (which was solved only partially and sub-optimally by AC). The exact EC is first enhanced by identifying new *general recurrences* for enumerative addends (eq. (5), [3]), which are then used to construct the EC index (eq. (15), [3]) and to quantize the addends bottom-up (eq. (21), [3]) via *sliding window integers* (new numeric type of wider applicability, similar to floating point, except that its arithmetic operations are formally decoupled from rounding). The resulting index (eq. (23), [3]) is the closest integer to the exact EC index for any limited precision addends which satisfy Kraft inequality (eq. (20), [3]). The QI *modeling interface* (N7 p. 9, [3]) uses precise finite sequence parameters, which are a richer, sharper and more flexible language for finite sequences than the infinite sequence limit values such as probabilities (e.g. of the single next symbol). Due to its optimality and finer-grained parametrization, the QI index for fixed sequence parameters has a *precisely* fixed length, to the exact bit fraction, and its length is available *without decoding* (which is useful for optimal packing, fast traversals and random access of compressed records).

Below are sample ***test results*** of QI vs. AC [2] for input sizes N and given # of 1's (binary order 0 coders; $\mathbf{K} \equiv 2^{10}$ bits; 500 random inputs/result; columns **N**: output ***size %*** (A/Q-1)∗100, **Speed**: coding times ***ratio*** A/Q; array **Vary** ≡ int32 {...,-2,-1,0,+1,+2,...} illustrates the resilience of QI's *descriptive* (MDL) and fragility of AC's *predictive* modeling for 'unpredictable' inputs).

#1's	**N:4**K	**Speed**	**N:8**K	**Speed**	**N:32**K	**Speed**	**N:128**K	**Speed**
8	6.846	**68**.3	6.421	**112**.8	5.447	**199**.6	5.966	**247**.5 x
16	4.175	59.7	3.830	78.5	3.389	138.1	3.730	168.0
32	2.297	49.7	2.090	58.9	2.096	95.9	2.220	117.2
N/64	1.370	40.3	0.606	41.0	0.186	41.7	0.073	42.5
N/32	0.897	30.8	0.343	33.7	0.123	34.2	0.049	34.5
N/16	0.505	21.8	0.197	25.3	0.084	24.6	0.040	24.8
N/8	0.359	14.4	0.155	16.7	0.069	16.8	0.045	16.8
N/4	0.288	9.2	0.138	10.8	0.083	10.6	0.068	10.5
N/2	0.509	6.6	0.445	6.6	0.367	6.4	0.332	6.4
Vary	**110**.899%	21.9	**96**.736	19.6	**71**.308	16.5	**52**.580	14.1

[1] T.M. Cover *Enumerative source coding* IEEE Trans. Inf. Th., 19 (1), 73, 1973.
[2] A. Moffat, R.M. Neal, I.H. Witten *Arithmetic coding revisited* ACM Trans. Inf. Sys., 16 (3), 256, 1998.
[3] R.V. Tomic *Quantized Indexing: Beyond Arithmetic Coding* arXiv.org cs.IT/0511057, Nov 2005.
Extended technical reports and QI source code are available at: www.1stworks.com/ref/qi.htm