Error Containment in Compressed Data Using Sync Markers

Aaron Kiely*, Sam Dolinar*, Matthew Klimesh*, and Adina Matache*
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive, Mail Stop 238-420, Pasadena, CA 91109
e-mail: {aaron, sam, klimesh, matache}@shannon.jpl.nasa.gov

Abstract — **We examine a specific strategy of using sync markers for error containment in compressed data, using a model that separates the data compression and error containment stages.**

I. A Simple Error Containment Scheme

Consider a data compressor that maps blocks of source symbols to variable length binary sequences. At the output of the compressor, we assume that zeros and ones are equally likely and that an error at this point would be undetectable to the decompressor. To limit the effect of channel errors we use a special sequence called a sync marker. The length m sync marker s is inserted between compressed blocks so that block boundaries can be identified following a channel error. To prevent the chance occurrence of the sync marker within the compressed data sequence, we insert a bit whenever we observe the first $m-1$ bits of the sync marker.

Our aim is to determine how to choose sync markers and analyze the impact on overall performance of this strategy. A complete version of these results is available in [2]. Related work includes [1, 3].

II. Error Containment Performance

The bit insertion procedure can be described using a state diagram that is essentially the same as that in [1]. The expected total number of bits $I_n(s)$ inserted in a block of n compressed bits can be computed from the transition matrix for the state diagram [2].

Two length m sync markers s and t are said to be *equivalent* if they give identical performance in the error containment scheme, i.e., when $I_n(s) = I_n(t)$ for all $n > 0$. Two equivalent sync markers can have very different state diagrams.

We define the overlap set of the bit string $s = s_1 s_2 \ldots s_m$ as $V(s) = \{1 \le i < m : s_1 \ldots s_i = s_{m-i+1} \ldots s_m\}$. In other words, $i \in V(s)$ if s can be written twice with i identical bits overlapping. Let $V'(s)$ denote the overlap set of s with the last bit inverted.

Theorem 1 *If s and t are length m sync markers for which $V(s) = V(t)$ and $V'(s) = V'(t)$, then s and t are equivalent.*

The asymptotic growth rate of the average number of inserted bits $I_n(s)$ can be neatly evaluated for any sync marker s:

Theorem 2 *If s is a length m sync marker, then*

$$\lim_{n\to\infty} \frac{1}{n} I_n(s) = \left(2^{m-1} - 1 + \sum_{i\in V(s)} 2^{i-1} - \sum_{i\in V'(s)} 2^{i-1}\right)^{-1}.$$

*The work described was funded by the TMOD Technology Program and performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with the National Aeronautics and Space Administration.

Given a blocklength n, for any fixed sync marker length $m \le 8$, we have verified empirically (and we conjecture that it is true for all m) that the optimal sync marker must belong to one of three classes. These classes are those containing 10^{m-1}, $10^{m-2}1$, and 1^m, which we refer to as class-1, class-2, and class-3, respectively.

Theorem 3 *For class-1, class-2, and class-3 sync markers, the average number of inserted bits $I_n(s)$ takes the following form wherever it is nonzero:*

$$I_n(s) = \frac{n}{a(s)} + b(s) + c_n(s)2^{-n},$$

where $a(s)$, $b(s)$ are independent of n, and $c_n(s)$ is periodic in n with a short period on the order of the length of s. Expressions for $a(s)$, $b(s)$, and $c_n(s)$ are given in [2] for each of the three special classes.

To compare the performance of sync markers of different lengths, we compute the average *data expansion*, i.e., the average number of extra bits that are added to each data block for synchronization purposes. The average data expansion is $X_n(s) = |s| + I_n(s)$ where $|s|$ denotes the length of sync marker s. In Figure 1 we plot the difference between the globally optimum average data expansion $\min_s X_n(s)$ and $\log_2 n$.

For large n, class-1 and class-2 markers take approximately equal turns at being optimum, and the globally optimum average data expansion is confined to a tight range of values between $\log_2 n + 1.9$ and $\log_2 n + 2$. Class-3 markers, while asymptotically optimum for any *fixed* marker length m, are never globally optimum.

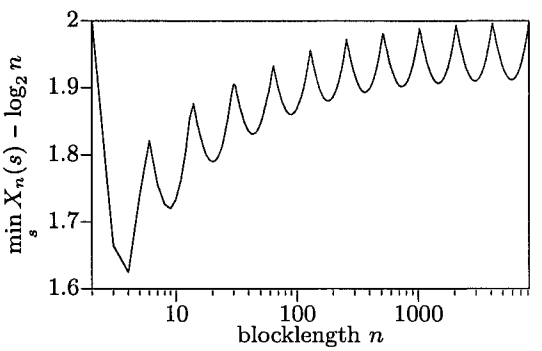

Figure 1: Globally minimum average data expansion, $\min_s X_n(s)$, plotted as a difference relative to $\log_2 n$.

References

[1] E. N. Gilbert, "Synchronization of Binary Messages," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 470–477, 1960.

[2] A. Kiely, S. Dolinar, M. Klimesh, A. Matache, "Synchronization Markers for Error Containment in Compressed Data," *TMO Progress Report 42-136*, Oct.-Dec. 1998, pp. 1-40, Feb. 15, 1999. tmo.jpl.nasa.gov/tmo/progress_report/42-136/136H.pdf

[3] J. J. Stiffler, *Theory of Synchronous Communications*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.