

Output distribution of the Burrows-Wheeler transform¹

Karthik Visweswariah
T. J. Watson Research Center
IBM, NY, USA
kvi@watson.ibm.com

Sanjeev Kulkarni
Dept. of Elec. Eng.
Princeton University, USA
kulkarni@ee.princeton.edu

Sergio Verdú
Dept. of Elec. Eng.
Princeton University, USA
verdu@ee.princeton.edu

Abstract — The Burrows-Wheeler transform is a block-sorting algorithm which has been shown empirically to be useful in compressing text data. In this paper we study the output distribution of the transform for i.i.d. sources, tree sources and stationary ergodic sources. We can also give analytic bounds on the performance of some universal compression schemes which use the Burrows-Wheeler transform.

I. INTRODUCTION

Burrows and Wheeler [2] proposed a lossless transformation which they showed (with empirical evidence) to be useful for the lossless compression of data. Recently there has been increasing interest in understanding and improving the performance of data compression algorithms using the Burrows-Wheeler transform (BWT). From empirical evidence [2] it appears that compression methods using this transform achieve better performance than Lempel-Ziv techniques, while not being computationally as intensive as compression methods using statistical modeling techniques. While there has been a large amount of empirical evidence to show the efficacy of the transform (e.g., [2], [3]), the analysis of the compression efficiency of methods based on the transform has received less attention. Sadakane [5], Arimura and Yamamoto [6], Balkenhol and Kurtz [4] and Effros [1] have provided the first steps in this direction.

In this paper we investigate the joint distribution at the output of the Burrows-Wheeler transform. For various classes of input sources, we show that the output distribution of the transform is approximately memoryless and piecewise stationary, in the sense that the normalized divergence between the output distribution and a memoryless and piecewise stationary distribution is small. Thus coding schemes that are good for memoryless, piecewise stationary sources can be used to give good coding performance. We also derive bounds on the coding rate for some data compression algorithms that use the BWT. The schemes that we analyze were also analyzed in [1] where bounds were obtained on average code length. The bounds we give are on individual sequences.

II. MAIN RESULT

We now introduce some notation so that we can precisely state our main result. We consider a Markov process \mathbf{X} which is a Markov source taking values in A and the set of states \mathcal{S} is a complete and prefix-free subset of A^* . Let $|\mathcal{S}| = k$ and label the states s_1, s_2, \dots, s_k in lexicographic order. We assume that the Markov source is irreducible and aperiodic. Let the steady state probability of a state $s \in \mathcal{S}$ be denoted by $\pi(s)$ and $P(a|s)$ denote the probability that $a \in A$ occurs when

we are in state $s \in \mathcal{S}$. Let $C(i) = \sum_{j=1}^i \pi(s_j)$. We will show that the divergence between the output distribution and a memoryless, piecewise stationary distribution with $k-1$ transitions is small. Let T_1, T_2, \dots, T_{k+1} be integers defined by $T_i = \lfloor C(i-1)n \rfloor + 1$. Note that $C(0) = 0$ and so $T_1 = 1$. Let us now define a memoryless distribution Q^n with $k-1$ changes in distribution, by

$$Q^n(y^n) = \prod_{j=1}^k \prod_{i=T_j}^{T_{j+1}-1} P(y_i|s_j).$$

We show that the output distribution is close to the distribution Q^n .

Theorem 1 Consider a tree source for which $P(a|s) > 0$ for all $a \in A, s \in \mathcal{S}$ with entropy rate H . Let X^n be the output of the tree source in steady state, $Y^n = \phi_{\text{BWT}}(\mathcal{R}(X^n))$ and P_{Y^n} denote the distribution of Y^n . Then

$$\frac{1}{n} D(P_{Y^n} || Q^n) \leq \frac{c}{\sqrt{n}}$$

for some constant c , where \mathcal{R} is a map from a string to its reverse and ϕ_{BWT} is a map from a string to the string part of its Burrows-Wheeler Transform.

The assumption that $P(a|s) > 0$ for all a, s can be removed and a result similar to the one above can be given. A result similar in spirit to the one above can also be shown for stationary ergodic sources.

Finally, we mention that we have also analyzed various methods to compress the the output of the BWT and obtained bounds on their performance. These results are like those in [1] except that we obtain results for individual sequences.

REFERENCES

- [1] M. Effros, "Universal lossless source coding with the Burrows-Wheeler transform," in *Proc. Data Compression Conference*, Snowbird, UT, 1999, pp. 178-187.
- [2] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," Tech. Rep. 124, Digital Systems Research Center, 1994.
- [3] M. Nelson, "Data compression with the Burrows-Wheeler transform," *Dr.Dobb's Journal*, pp. 46-50, September 1996.
- [4] B. Balkenhol and S. Kurtz, "Universal lossless data compression based on the Burrows Wheeler Transformation: Theory and Practice," Tech. Rep. 98-069, Universitat Bielefeld, 1998, <http://www.mathematik.uni-bielefeld.de/sfb343/preprints/>.
- [5] K. Sadakane, "On optimality of variants of block-sorting compression," in *Proceedings Symposium on Information Theory and its applications*, Matsuyama, Japan, December 1997, pp. 357-360.
- [6] M. Arimura and H. Yamamoto, "Asymptotic optimality of the block sorting data compression algorithm," *IEICE Transactions on fundamentals of electronics communications and computer sciences*, pp. 2117-2122, October 1998.

¹This work was partially supported by the National Science Foundation under Grants NYI Award IRI-9457645 and NCR 9523805