Generalized Context Modeling With Multi-Directional Structuring and MDL-Based Model Selection for Heterogeneous Data Compression

Wenrui Dai, *Member, IEEE*, Hongkai Xiong, *Senior Member, IEEE*, Jia Wang, Samuel Cheng, *Senior Member, IEEE*, and Yuan F. Zheng, *Fellow, IEEE*

***Abstract*—This paper proposes generalized context modeling (GCM) for heterogeneous data compression. The proposed model extends the suffix of predicted subsequences in classic context modeling to arbitrary combinations of symbols in multiple directions. To address the selection of contexts, GCM constructs a model graph with a combinatorial structuring of finite order combination of predicted symbols as its nodes. The estimated probability for prediction is obtained by weighting over a class of context models that contain all the occurrences of nodes in the model graph. Moreover, separable context modeling in each direction is adopted for efficient prediction. To find optimal class of context models for prediction, the normalized maximum likelihood (NML) function is developed to estimate their structures and parameters, especially for heterogeneous data with large sizes. Furthermore, it is refined by context pruning to exclude the redundant models. Such model selection is optimal in the sense of minimum description length (MDL) principle, whose divergence is proven to be consistent with the actual distribution. It is shown that upper bounds of model redundancy for GCM are irrelevant to the size of data. GCM is validated in an extensive field of applications, e.g., Calgary corpus, executable files, and genomic data. Experimental results show that it outperforms most state-of-the-art context modeling algorithms reported.**

***Index Terms*—Context modeling, heterogeneous data compression, minimum description length, model redundancy, model selection.**

Manuscript received January 09, 2015; revised June 21, 2015; accepted July 02, 2015. Date of publication July 20, 2015; date of current version September 21, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhengdao Wang. The work was supported in part by the NSFC, under grants 61425011, 61472234, 61271218, 61271221, and U1201255, NSF under Grant CCF 1117886, and "Shu Guang" project (13SG13).

W. Dai is with the Department of Biomedical Informatics, University of California, San Diego, CA 92093 USA, and also with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wed004@ucsd.edu).

H. Xiong and J. Wang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xionghongkai@sjtu.edu.cn; jiawang@sjtu.edu.cn).

S. Cheng is with the Department of Computer Science and Technology, Tongji University, Shanghai 201840, China, and the Department of Electrical and Computer Engineering, the University of Oklahoma, Tulsa, OK 74135 USA (e-mail: samuel.cheng@ou.edu).

Y. F. Zheng is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: zheng@ece.osu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2458784

I. Introduction

CONTEXT MODELING plays a significant role in most compression applications, which maps the context space into a parameter set constructed from statistics of the source. To exploit the statistical dependencies in the source, context modeling is parametrically represented in a probabilistic framework that concerns the structure of context models with probability assignment. In classical context modeling [1], [2], the optimal model is selected with regard to its order k, as each context is the suffix of predicted subsequences. However, these suffix-based techniques are greatly challenged in prediction and compression of heterogeneous data, e.g., image, video [3] and executable files [4]. Heterogeneous data are composed of numerous data streams with varying formats and partially unknown distributions, which are interlaced in a complex manner. It is insufficient to capture their statistics with classical sequential contexts. As a consequence, the class of extensive models with multi-directional properties and combinatorial structures are adopted. For such class of models, model selection with well-established information criterion is necessary to select its optimal subset for prediction.

Context modeling methods can be traced back to fixed-order predictors [5], [6], which perform asymptotically as well as the best Markov predictor with fixed order k. However, the fixed-order assumption is undesirable because of its strong dependency on the prior knowledge of k. Consequently, classical context modeling methods are typically based on the finite context statistical algorithms incorporating variable-order Markov models, where the optimal context model is selected by adaptively estimating its order. Among them, the state-of-the-arts are prediction by partial match (PPM) [7], [8], and context tree weighting (CTW) [9]. PPM utilized a set of finite order models and adaptively switched from the higher to the lower to fit the statistics of the source. While CTW estimated the weighted mixture over all finite-order models to asymptotically approximate the optimal one. Although these methods guarantee an asymptotic upper bound for prediction error, their performance is degraded in heterogeneous data compression. As an alternative, preprocessing and adjusting of the contexts are adopted, e.g., instruction rescheduling and split-stream for executable files [4], [10], dictionary-based preprocessing for genomic data [11], and approximate contexts for natural images [12]. However, these methods are restricted to specific sources, in that they depend on the prior knowledge of the statistics of sources.

As an improvement for heterogeneous data, multi-directional extension and combinatorial structuring of contexts are adopted to fully exploit the correlations of the interlaced data streams. Through extending CTW, [13] developed adaptive bidirectional context modeling for unknown discrete stationary sources. Importing the concept of margin, [14] casted prediction task as the problem of linear separation in Hilbert space and applied machine learning techniques and online optimization to the problem. Multi-directional extension of contexts are efficient to describe sources with multiple underlying distributions, but [15] demonstrated that CTW could not be extended to represent contexts with three or more directions. The other attempts focus on constructing contexts with arbitrary combination of predicted symbols. [16] extended CTW to recursive context weighting for arbitrary position splitting, where the set of all contexts were split at arbitrary position and a class of models could be generated to describe the sources. [17] and [18] improved CTW's recursive weighting scheme with switching distribution for the cases that the actual context model does not live in the model class. [19] enumerated sets of models with distinctive characterization and made an efficient estimation by mixing all these models. However, well-formed model selection methods are not utilized in such methods to determine the optimal class of models from a wider variety of context models. Consequently, their coding redundancies cannot be tightly approximated for increasing models led by combinatorial structuring of contexts.

For extensive context models led by multi-directional structuring, the optimal class of models should be selected based on the partial knowledge of heterogeneous data by specifying their structures as well as orders. Model selection methods, e.g., Akaike's Information Criterion (AIC [20]), Bayes Information Criterion (BIC [21]), and the Minimum Description Length (MDL [22], [23]) principle, can evaluate the measurement of prediction error and the cost of describing the model (stochastic complexity) to select the proper context model in the model class. Recently, MDL is desirable for prediction of heterogeneous data with varying formats and partially unknown distributions, as it selects the optimal models for prediction rather than estimate the actual distribution that generates the heterogeneous data. MDL selects the optimal model that leads to the best compression of the data by capturing the regularities and structures of the source, especially when the data size and number of source parameters are large enough [24]. Relating with Shannon codelength assignment [25], MDL solves model selection problem by estimating the parameters of probabilistic models in the penalized likelihood form. MDL is firstly proposed in a two-part form equivalent to BIC [26] by omitting the cost terms independent of sequence size, which was later demonstrated to be inconsistent under the noise with vanishing variances [27]. As an alternative, the insight of one-part coding was introduced to jointly encode the data sequence and the optimal model parameters for predicting. It was applied in the line of work of "MDL denoising [28], [29]". Although restricted to the closed-form maximum-likelihood estimation (MLE), the one-part coding was proven to be consistent even when the noise variance vanished [30]. For one-part coding, normalized maximum likelihood (NML) distribution can be estimated to find the optimal class of models [31]. MDL is prevalent in a variety of signal processing applications, e.g., wireless sensor array processing [32], autoregressive models [33], sparse coding [34], lossless image coding [35], and etc.

The interlaced correlations of heterogeneous data can be further exploited for high-performance compression. For example, horizontal, vertical, and spatial correlations in other directions in images can be considered for precise adaptive prediction. Executable files are composed of various data fields, e.g., instruction opcodes, displacements, and immediate data fields. In genomic data compression, the correlations among approximate repeats of DNA nucleotides and regular non-repeat regions can be jointly estimated. Therefore, improved context modeling techniques with sophisticated model selection strategies tend to benefit heterogeneous data compression.

The contribution of this paper is twofold. Firstly, the generalized context modeling (GCM) is proposed to establish contexts for modeling heterogeneous data with multi-directional structuring of predicted symbols. The proposed method extends the context from the suffix of predicted subsequences to the combination of arbitrary predicted symbols with arbitrary finite directions. For selection of contexts, model graph is constructed to represent the combinatorial structuring of contexts and regularize their occurrence in the context models. In GCM, the class of context models contains all the occurrences of nodes in the model tree. Consequently, the estimated probability for prediction is obtained by weighting over this model class. Separable context modeling in each direction is also developed for efficient context-based prediction.

Furthermore, MDL-based model selection for GCM is developed for the optimal class of context models. The selected class of models is optimal in the MDL sense, where complexity vanishes with the growth of data size. For the sequential prediction of heterogeneous data, sequentially normalized maximum likelihood (SNML) function is adopted to estimate the structures and parameters of contexts for each symbol. When the data size is large enough, SNML tends to asymptotically obtain optimal structures and parameters of multi-directional contexts. For compression, the model class is refined by context pruning to exclude the redundant models, and subsequently tuned into the optimal class of models. The additional model redundancy led by multi-directional structuring is proven to vanish with the growth of data size, which only depends on the number of directions and maximum order in each direction. For practical validation, GCM is applied into compression of the Calgary corpus and the specific heterogeneous data, e.g., executable files and genomic data.

The remainder of this paper is organized as follows. Section II provides the problem statement and introduces the notations we use in the sequel. In Section III, the formulation of the generalized context modeling is provided. In Section IV, separable GCM is developed and the optimal class of context models is selected in the sense of MDL principle. Section V develops upper bounds of model redundancy led by multi-directional extension and combinatorial structuring, respectively. To determine model class for prediction, a context pruning-based algorithm is developed in Section VI. Experimental results are shown to evaluate the compression performance in Section VII.

II. Problem Statement and Notation

In the remainder of this paper, we reserve normal symbols to scalar variables and bold face symbols to vector variables. Calligraphic symbols are used to represent sets of variables.

Consider heterogeneous data x_1^N generated by interlacing M data streams with an unknown random process $\Omega(t)$. Denote $x_1^{(j)} \cdots x_{N_j}^{(j)}$ the j-th data stream emitted from a stationary Markov source π_j with order d_j that takes its value in an alphabet $\mathcal{A} = \{a_1, \ldots, a_L\}$. Hereby, we define the heterogeneous data x_1^N by its generating process.

Definition 1 (Heterogeneous Data): Given the unknown random process $\Omega(t) \in \{1, \ldots, M\}$ varying in time, each symbol x_t of x_1^N is obtained from the j_t-th data stream $\{x_{n_{j_t}}^{(j_t)}\}_{j_t=1}^M$ for $t = 1$ to N by the following steps:

1) $j_t = \Omega(t)$,
2) $x_t = x_{n_{j_t}}^{(j_t)}$,
3) $n_{j_t} = n_{j_t} + 1$,

where $n_1, \ldots, n_M$ start from 1. Thus, it holds for x_1^N that

i) $\sum_{j=1}^M N_j = N$ and the depth $D = \max_{1 \le j \le M} d_j$;
ii) x_1^N is not wide sense stationary, as $\Omega(t)$ varies in time.

Thus, x_1^N is also non-stationary.

Definition 1 shows that heterogeneous data generalizes the formulation of piecewise stationary memoryless sources [36]–[38] by arbitrarily segmenting x_1^N into M data streams with $\Omega(t)$ rather than in a sequential manner. According to Definition 1, x_1^N can be predicted in a sequential procedure. In classical context modeling, context s for predicting x_t is arbitrary suffix of x_{t-D}^{t-1} with length $l(s)$. A valid context model $\mathcal{S}$ over the D-fold vector product of alphabet $\mathcal{A}$ is defined to satisfy the exhaustive and disjoint properties [15].

Definition 2 (Classical Context Set): Denote $\mathcal{C}(s)$ the set of subsequences whose suffix is s. $\mathcal{S}$ is a valid context model for the stationary Markov source π_i if it satisfies:

i) Exhaustive property:$\bigcup_{s\in\mathcal{S}} \mathcal{C}(s) = \mathcal{A}^D$
ii) Disjoint property: for any pair of contexts $s \neq s'$, $\mathcal{C}(s) \bigcap \mathcal{C}(s') = \emptyset$.

Definition 2 can be extended to M-directional case $\mathbf{s} = (s^{(1)}, s^{(2)}, \ldots, s^{(M)})$, but only uni-directional and bi-directional context models are available for tree representation. To exploit the interlaced correlations of heterogeneous data, classical context models are generalized with multi-directional structuring. Context s with combinatorial structuring is extended from the suffix of x_{t-D}^{t-1} to arbitrary combination of predicted symbols $x_{t-i_{l(s)}} \cdots x_{t-i_2} x_{t-i_1}$, $1 \le i_1 < \cdots < i_{l(s)} \le D$, which enables conditional dependencies based on arbitrary previous positions. To describe contexts with combinatorial structuring, index set is introduced to indicate the symbols contained in the contexts (or subsequences) in addition to their values. For example, the context s in the form of $x_{t-i_{l(s)}}, \ldots, x_{t-i_2}, x_{t-i_1}$ can be represented with an index set $\mathcal{I}(s) = \{i_1, i_2, \ldots, i_{l(s)}\}$. Consequently, $\mathcal{C}(s)$ is generalized to the set of subsequences whose index sets contain the one of context s.

$$\mathcal{C}(s) = \{x_m^n | \mathcal{I}(s) \subseteq \mathcal{I}(x_m^n)\}.$$

Letting $\mathcal{A}^* = \cup_{i \le D} \mathcal{A}^i$ be the set of all strings with a length not greater than D, arbitrary combination of predicted symbols can be represented by strings in $\mathcal{A}^*$. Consequently, a valid context model with combinatorial structuring is defined over $\mathcal{A}^*$.

Definition 3 (Context Set With Combinatorial Structuring): $\mathcal{S}$ is a valid context model with combinatorial structuring for the stationary Markov source π_i, if it satisfies:

i) Exhaustive property: for any subsequence x_n^m, there exists context s in $\mathcal{S}$ contained in it, or $\bigcup_{s\in\mathcal{S}} \mathcal{C}(s) = \mathcal{A}^*$.
ii) Disjoint property: for any subsequence x_n^m, only one context s in $\mathcal{S}$ can be found contained in it, or for any pair $s \neq s'$ in $\mathcal{S}$, $\mathcal{C}(s) \bigcap \mathcal{C}(s') = \emptyset$;

In comparison to classical context modeling, Definition 3 allows a wider variety of context models, as there are $D - i_{l(s)}$ groups of contexts with length $l(s)+1$ that contain s. To notify, we denote $\{sc_j | c_j \in \mathcal{A}\}$ the j-th group of contexts with index set $\mathcal{I}(sc_j) = \{i_1, \ldots, i_{l(s)}, i_{l(s)} + j\}$.

To fit the M interlaced sources $\{\pi_j\}_{j=1}^M$, Definition 3 can be further extended to multi-directional cases, where the set of contexts in each direction is a valid context model with combinatorial structuring. Denote $\mathbf{s} = (s^{(1)}, \ldots, s^{(M)})$ the M-directional context, where $s^{(j)}$ is arbitrary combination of D previously predicted symbols. Naturally, its index set is extended by

$$\mathcal{I}(\mathbf{s}) = \mathcal{I}\left(s^{(1)}\right) \times \mathcal{I}\left(s^{(2)}\right) \times \cdots \times \mathcal{I}\left(s^{(M)}\right),$$

where $\times$ denotes cartesian product. A valid M-directional context model $\mathcal{S} \subseteq (\mathcal{A}^*)^M$ is generalized with multi-directional structuring in Definition 4.

Definition 4 (Context set With Multi-Directional Structuring): $\mathcal{S}$ is a valid M-directional context model for M interlaced stationary Markov sources $\{\pi_i\}_{i=1}^M$, if it satisfies in each of its direction:

i) Exhaustive property: for any subsequence $x_{m_j}^{n_j} \in \mathcal{A}^*$ in j-th direction, there exists $\mathbf{s}$ in $\mathcal{S}$ such that $s^{(j)}$ is contained in it, or $\cup_{\mathbf{s}\in\mathcal{S}} \mathcal{C}\left(s^{(j)}\right) = \mathcal{A}^*$.
ii) Disjoint property: for any subsequence $x_{m_j}^{n_j} \in \mathcal{A}^*$ in j-th direction, only one $\mathbf{s}$ in $\mathcal{S}$ can be found such that $s^{(j)}$ is contained in it, or for any pair $\mathbf{s}$ and $\mathbf{s}'$ in $\mathcal{S}$, $\mathcal{C}\left(s^{(j)}\right) \cap \mathcal{C}\left(s'^{(j)}\right) = \emptyset$.

Context sets in Definition 4 are valid for multi-directional structuring in any finite M directions to utilize the Markovian property of the M interlaced sources. Here, contexts in each direction are utilized to exploit correlations in one data stream or a combination of several correlated data streams. Consequently, GCM constructs the class of context models with multi-directional extension and combinatorial structuring, as defined in Definition 4, to fully exploit the statistical correlations of the heterogeneous data. Under the assumption of finite-order context, model class $\mathcal{M}$ is usually the whole set of context models with maximum D-order contexts.

$$\mathcal{S}_k = \left\{\mathbf{s} | l\left(s^{(j)}\right) \le k, 1 \le j \le M\right\} \quad \mathcal{M} = \bigcup_{k=1}^{D} \mathcal{S}_k. \tag{1}$$

Thus, the main problem in GCM is to select an optimal subset of models in $\mathcal{M}$ to predict heterogeneous data x_1^N.

$$\hat{\mathcal{M}} = \arg \min_{\mathcal{M}' \subseteq \mathcal{M}} \mathcal{L}\left(x_1^N, \mathcal{M}'\right) \tag{2}$$

where $\mathcal{L}\left(x_1^N, \mathcal{M}'\right)$ is the measurement of predicting x_1^N with $\mathcal{M}'$ under the MDL principle.

III. Generalized Context Modeling

GCM adopts multi-directional structuring to establish extensive contexts with flexible structures for prediction. However, these contexts cannot be represented with splittable structures

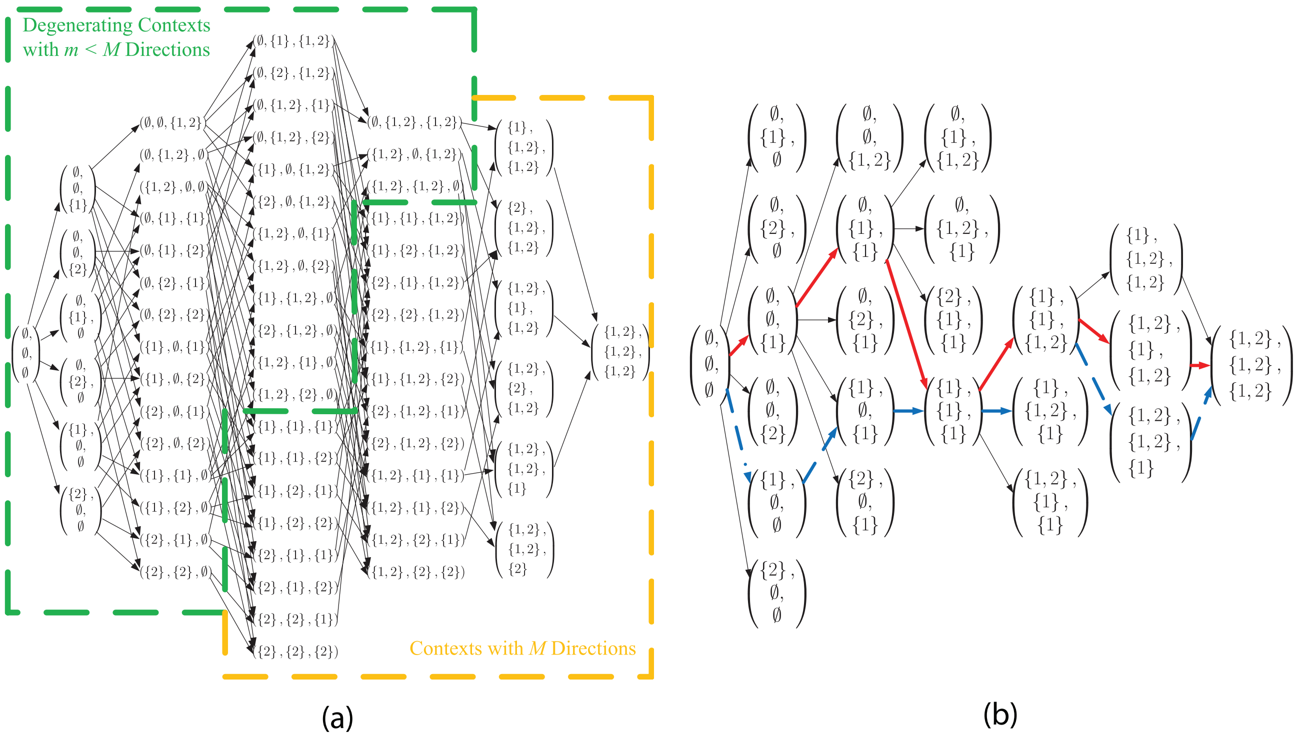


Fig. 1. An example of model graph with depth $D = 2$ and $M = 3$ directions. (a) The complete trellis-based graph with $2^{2\cdot 3} - 1 = 63$ nodes and $2 \cdot 3 + 1 = 7$ vertical slices. Each node is connected to its succeeding nodes with arrow lines; (b) Two selected paths (dashed and solid) and their branches in the model graph. The two paths share two common nodes.

like tree, which will affect the efficient construction of context models. In this section, model graph is developed to specify and regularize the structures of contexts. Consequently, the estimated probability for prediction is proposed to weight over all the valid context models.

A. Model Graph

For GCM with multi-directional structuring, model graph $\mathcal{G}$ is constructed to specify all the finite order combinations of predicted symbols as well as their restrictions in constructing a context model. Each of its nodes γ is defined to be a valid context structure, so that it can be represented with the corresponding index set $\mathcal{I}$. Therefore, $\mathcal{G}$ is defined to specify all the occurrences of contexts with its nodes based on Definition 4.

Definition 5 (Model Graph): Given depth D and the number of direction M, the model graph $\mathcal{G}$ is a trellis-like graph rooted from M-ary vector $(\emptyset, \ldots, \emptyset)$, where each node corresponds to an index set for finite order combination of predicted symbols. For arbitrary internal node $\gamma = \{\gamma^{(j)}\}_{j=1}^{M}$, its succeeding node γ' in $\mathcal{G}$ satisfies that

i) $\mathcal{I}(\gamma) \subset \mathcal{I}(\gamma')$ and $l(\gamma') = l(\gamma) + 1$,

ii) $i_{l(\gamma^{(j)})} < i_{l(\gamma'^{(j)})} \leq D$ for $\gamma^{(j)}$ with $l(\gamma^{(j)}) < D$.

In analogy, its preceding node γ'' satisfies that

i) $\mathcal{I}(\gamma'') \subset \mathcal{I}(\gamma)$ and $l(\gamma'') = l(\gamma) - 1$,

ii) $\mathcal{I}(\gamma''^{(j)}) = \emptyset$ or $i_{l(\gamma''^{(j)})} < i_{l(\gamma'^{(j)})}$ for non-empty $\gamma^{(j)}$.

Definition 5 shows all $2^{DM} - 1$ possible context structures for GCM with given M and D. In $\mathcal{G}$, they locate in $DM+1$ vertical slices with C_{MD}^{l} nodes for the l-th one. Fig. 1(a) provides an example for model graph with $D = 2$ and $M = 3$, where each node is connected to its succeeding nodes with arrow lines. Definition 5 restricts the selected contexts in the construction of a context model, as it implies that contexts from the same path in $\mathcal{G}$ would violate the disjoint property in Definition 4. Thus, for contexts with same values of symbols in corresponding positions, a valid context model can contain only one context from a path. Remarkably, contrary to tree-like structure, paths in $\mathcal{G}$ can share some nodes. As shown in Fig. 1(b), the solid and dashed paths have two common nodes. This fact implies that the proposed model graph can represent extensive contexts that are not splittable in previous literature.

Furthermore, degenerating contexts with empty components are adopted in Definition 5, which means that contexts in GCM can be composed of predicted symbols from any part of the M interlaced data stream. For example, nodes in the green box in Fig. 1(a) represent all the context with $j < 3$ directions. As a result, GCM is adaptive to variable numbers of directions that are not greater than M.

B. Estimated Probability for GCM

The estimated probability for generalized context modeling is obtained by weighting over the model class generalized with multi-directional extension and combinatorial structuring.

$$P_w\left(x_1^N\right) = \sum_{\mathcal{S}\in\mathcal{M}} w(\mathcal{S}) \prod_{s\in\mathcal{S}} Pr\left(x_1^N|\mathbf{s}\right), \tag{3}$$

where $Pr(x_1^N|\mathbf{s})$ is the estimated probability conditioned on context $\mathbf{s}$. For simplicity, binary alphabet $\mathcal{A} = \{0,1\}$ is considered, where $Pr(x_1^N|\mathbf{s}) = P_e(a_{\mathbf{s}}, b_{\mathbf{s}})$ is estimated by counts of zeroes $a_{\mathbf{s}}$ and ones $b_{\mathbf{s}}$ based on context $\mathbf{s}$. Commonly, Krichevski-Trofimov (KT) estimated probability [39] is adopted, which weights over all source parameters θ with a $(\frac{1}{2}, \frac{1}{2})$-Dirichlet distribution. Consequently, it can achieve minimum average redundancy for the worst-case source parameters [40], [41].

$$P_e(a_s, b_s) = \int_0^1 \frac{1}{\pi\sqrt{(1-\theta)\theta}}(1-\theta)^{a_s}\theta^{b_s}d\theta = \frac{\frac{1}{2}\cdots\left(a_s-\frac{1}{2}\right)\cdot\frac{1}{2}\cdots\left(b_s-\frac{1}{2}\right)}{(a_s+b_s+1)!}. \quad (4)$$

KT-estimator can be computed in a sequential manner.

$$P_e(a_\mathbf{s}+1, b_\mathbf{s}) = \frac{a+\frac{1}{2}}{a+b+1}\cdot P_e(a_\mathbf{s}, b_\mathbf{s}) \text{ and}$$
$$P_e(a_\mathbf{s}, b_\mathbf{s}+1) = \frac{b+\frac{1}{2}}{a+b+1}\cdot P_e(a_\mathbf{s}, b_\mathbf{s}). \quad (5)$$

Such that, the estimated probability for each symbol x_t is obtained by weighting all $P_e(x_t|\mathbf{s}) = P_e(a_\mathbf{s}^{(t)}, b_\mathbf{s}^{(t)})$ based on context $\mathbf{s}$.

$$P_w(x_t) = \sum_\mathbf{s} w(\mathbf{s})P_e(x_t|\mathbf{s}), \quad (6)$$

where $w(\mathbf{s})$ is the non-negative weight for $P_e(x_t|\mathbf{s})$. Without prior knowledge, all the $2^{DM}-1$ possible contexts with M directions and depth D are considered for x_t. Since the counts of symbols in alphabet $\mathcal{A}$ based on contexts with shorter length can be obtained by merging those based on longer-length contexts, their weights can be compensated with a positive constant η. As a result, the weight for the context with length $l(\mathbf{s}) = \sum_{i=1}^M l(s^{(i)})$ is derived.

$$w(\mathbf{s}) = \frac{C_{MD}^{l(\mathbf{s})}\eta^{l(\mathbf{s})}}{\sum_{k=1}^{MD} C_{MD}^k \eta^k} = \frac{C_{MD}^{l(\mathbf{s})}\eta^{l(\mathbf{s})}}{(1+\eta)^{MD}-1}. \quad (7)$$

IV. Model Selection With MDL Principle

A. Separable Multi-directional Context Modeling

Since the size of model class grows with M in a polynomial manner, it is complicated to obtain weighted probabilities for prediction, especially when M is large. Thus, separable context modeling in each direction is considered to make the size of model class grow linearly with M.

Given M-directional context $\mathbf{s} = (s^{(1)}, \ldots, s^{(M)})$, the estimated conditional probability $Pr(x_1^N|\mathbf{s})$ for x_1^N and its marginals $Pr(x_1^N|s^{(j)})$, $1 \le j \le M$, satisfy that

$$Pr\left(x_1^N|s^{(j)}\right) = \sum_{\mathbf{s}^{\backslash(j)}} Pr\left(x_1^N|\mathbf{s}\right) Pr\left(\mathbf{s}^{\backslash(j)}\right), \quad (8)$$

where $\mathbf{s}^{\backslash(j)}$ is the vector of $M-1$ remaining components of $\mathbf{s}$ after removing $s^{(j)}$, and $\{Pr(\mathbf{s}^{\backslash(j)})\}$ is their probability distribution. For clarity, we denote $R = (L+1)^D - 1$, $U = R \cdot M$, and $V = R^M$. Consequently, $\{Pr(x_1^N|s^{(j)})\}$ can be approximated with the linear combination of $\{Pr(x_1^N|\mathbf{s})\}$. For $\mathbf{s} \in (\mathcal{A}^*)^M$ and $1 \le j \le M$, denote $\mathbf{p}_{est} = \{Pr(x_1^N|\mathbf{s})\}_{1\times V}$ and $\mathbf{p}_{marg} = \{Pr(x_1^N|s^{(i)} \in \mathcal{A}^*)\}_{1\times U}$ the vector of all possible estimated conditional probability and their marginals, respectively. Equation (9) can be drawn from (8).

$$\mathbf{p}_{marg} = \mathcal{H}\cdot\mathbf{p}_{est}, \quad (9)$$

where $\mathcal{H}$ is a $U \times V$ coefficient matrix with its elements h_{uv} determined by (11). By solving (9), Proposition 1 shows that the probability distribution for contexts with multi-directional structuring can be approximated by a linear combination of its marginals.

Proposition 1 (Separability for GCM): In GCM, prediction based on contexts with multi-directional structuring can be made separately in each of its directions.

Proof: Please refer to Appendix A. ■

Proposition 1 implies that GCM can be achieved by combining the context models with combinatorial structuring in each of its directions. Thus, context-based prediction conditioned on M-directional context are separable. Actually, since $V > U$ for $M > 1$, (9) is an ill-posed problem due to underdetermined matrix $\mathcal{H}$. Notice that the rank of matrix $\mathcal{H}$ is $U-M$, it holds that

$$\mathcal{H}\mathcal{H}^T = \beta\begin{bmatrix}\Lambda_{U-M\times U-M} & \\ & \mathbf{0}_{M\times M}\end{bmatrix}\beta^T, \quad (10)$$

where Λ is a $(U-M)\times(U-M)$ diagonal matrix featuring $\mathcal{H}$'s eigenvalues. Equation (10) shows that only $U-M$ free parameters need to be considered when estimating $\mathbf{p}_{est}$. The elements of $\mathcal{H}$ can be obtained by comparing (8) and (9).

$$h_{uv} = \begin{cases} Pr(\mathbf{s}^{\backslash(j)} = \mathbf{c}^{\backslash(j)}) & u \bmod R = \lceil v/R \rceil \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where $j = \lceil u/R \rceil$ and $\mathbf{c}$ is the contexts corresponding to v-th elements of $\mathbf{p}_{est}$. The elements of coefficient matrix $\mathcal{H}$ are related to the distribution of contexts $\{Pr(\mathbf{s}^{\backslash(j)})\}$, which can be considered as the prior of M interlaced data streams. For practical coding, $\mathbf{p}_{est}$ is approximated by maximizing the conditional distribution with maximum entropy.

$$\begin{cases} \min \sum_v p_{est}^{(v)} \log p_{est}^{(v)} \\ \text{s.t. } \sum_v h_{uv} p_{est}^{(v)} = p_{marg}^{(u)} \quad \forall u \end{cases} \quad (12)$$

where $p_{est}^{(v)}$ and $p_{marg}^{(u)}$ are the v-th and u-th component of $\mathbf{p}_{est}$ and $\mathbf{p}_{marg}$, respectively.

B. Context Model Selection With MDL Principle

MDL proposes a generic description for model $\mathcal{S}$ in terms of codelength assignment function, where the best model is supposed to describe x_1^N with fewest number of bits.

$$\hat{\mathcal{S}} = \arg\min_{\mathcal{S}\in\mathcal{M}} \mathcal{L}(x_1^N, \mathcal{S}) = \arg\min_{\mathcal{S}\in\mathcal{M}} [-\log p(x_1^N|\mathcal{S}) - \log p(\mathcal{S})] \quad (13)$$

where $\mathcal{L}\left(x_1^N, \mathcal{S}\right)$ is the code length assignment function that uniquely describes x_1^N with model $\mathcal{S}$. It should be underlined that MDL considers compressibility as an indirect way of measuring the ability of a model to capture statistics of the data. To meet with practical coding, ideal Shannon code length [25] is commonly proposed for the code length assignment function. Since $p(x_1^N, \mathcal{S}) = p(x_1^N|\mathcal{S})p(\mathcal{S})$, $-\log p(\mathcal{S})$ in (13) indicates the model complexity. Normalized maximum likelihood is an elegant solution to model selection problem in MDL sense, which finds the optimal distribution that minimizes the maximized expected KL-divergence to the "worst case" distribution q[31].

$$p^* = \arg\min_p \max_q \mathbb{E}_q \left[\ln \frac{f\left(x_1^N, \hat{\theta}\left(x_1^N\right)\right)}{p\left(x_1^N\right)} \right],$$

where f is the likelihood function, q ranges over all distributions with finite KL-divergence to f, and $\hat{\theta}(x_1^N)$ is the ML estimation of parameters based on x_1^N. Therefore, the negative logarithm of the NML distribution can evaluate context models in terms of MDL principle.

As shown in Proposition 1, generalized context modeling is separable with regard to its directions. Therefore, the model selection problem can be considered independently in each of its directions. Without loss of generality, NML function for contexts in the m-th direction is considered. According to (5), the estimated probability for x_1^N can be decomposed as the product of its symbols. NML function for arbitrary symbol $x_t, 1 \le t \le N$ is developed.

$$f_{NML}(x_1^N, \mathcal{M}) = \frac{f(x_1^N|\theta(x_1^N))}{\int_{\theta(x_1^N)\in\Theta} f(x_1^N|\theta(x_1^N))dx_1^N}$$

where $f(\cdot)$ is the likelihood function for x_1^N and Θ is its parameter space.

C. Evaluation of Model Complexity

The log-NML function [42] can be approximated by

$$-\log f_{NML}\left(x_1^N|\mathcal{S}\right) = -\log f\left(x_1^N|\theta\left(x_1^N\right)\right) + \frac{d}{2}\log\frac{N}{2\pi} + \log\int_\Theta \sqrt{|\mathbf{J}(\theta)|}d\theta + o(1) \quad (14)$$

where d is the order of $\mathcal{S}$ and $\mathbf{J}(\cdot)$ is the per sample Fisher information matrix, whose elements are obtained by

$$\mathbf{J}_{ij}(\theta) = -\frac{1}{N}\mathbb{E}\left\{\frac{\partial^2 \log f\left(x_1^N|\theta\right)}{\partial\theta_i\partial\theta_j}\right\}. \quad (15)$$

Comparing (13) and (14), the model complexity $\mathcal{L}(\mathcal{S})$ is evaluated by

$$\mathcal{L}(\mathcal{S}) = -\log P(\mathcal{S}) = \frac{d}{2}\log\frac{N}{2\pi} + \log\int_\Theta \sqrt{|\mathbf{J}(\theta)|}d\theta. \quad (16)$$

This fact implies that the class of context models selected by NML is optimal in the sense of MDL principle [43]. Considering the model complexity in (16), the convergence of average model complexity $\mathcal{L}(\mathcal{S})/N$ depends solely on the Fisher information matrix $\mathbf{J}(\theta)$ for parameter space Θ as $\log N/N \to 0$.

Assuming that the heterogeneous data is composed of M interlaced autoregressive sources $\{\pi_j\}_{j=1}^M$ with order k_j^* and noise variance τ_j, the asymptotic per sample Fisher information matrix for source π_j is given by [44].

$$\mathbf{J}(\theta, \tau_j) = \lim_{N\to\infty}\left\{\frac{\mathbf{J}_N(\theta,\tau_j)}{N}\right\} = \begin{bmatrix}\Gamma_m(\theta) & 0 \\ 0 & \frac{1}{2\tau_j^2}\end{bmatrix} \quad (17)$$

where $\mathbf{J}_N(\theta, \tau_j)$ is the information matrix for coefficient space of sequence x_1^N and $\Gamma_j(\theta)$ is the $k \times k$ unit-variance process autocovariance matrix. Such that the evaluation of Fisher information matrix $\mathbf{J}(\cdot)$ in (16) depends on $\Gamma_j(\theta)$. Denote $\rho^{(j)} = \{\rho_k^{(j)}\}_{j=1}^d$ the autocorrelations of the d_j-order AR model coefficients θ. The eigenvalues of $\Gamma_j(\theta)$ are in the form of $\{1/(1-\rho_k^{(j)})^2\}_{k=1}^{d_j}$. Consequently, we can derive its determinant.

$$|\Gamma_j(\theta)| = \prod_{k=1}^{d_j}\frac{1}{\left(1-\rho_k^{(j)}\right)^2}.$$

Based on the eigenvalues $\{\rho_k^{(j)}\}$, $\Gamma_j(\theta)$ is proven to be constant with the growth of N.

Proposition 2: Given heterogeneous data sequence x_1^N generated from M interlaced d-th autoregressive sources with autocorrelations $\{\rho^{(j)}\}$, its NML estimation is almost surely constant with the growth of N.

$$\int_{\Theta(\eta)}\sqrt{|\Gamma_j(\theta)|}d\theta \le c\left(\rho^{(j)}\right) \quad (18)$$

where $\Theta(\xi)$ is the set of parameters constraining $\rho^{(j)}$ with $\|\rho^{(j)}\|_\infty \le \xi^{(j)}$ in the j-th direction and $c\left(\rho^{(j)}\right)$ is a constant depends solely on $\rho^{(j)}$.

Proof: Please refer to Appendix B. ■

Proposition 2 shows that the average model complexity vanishes asymptotically with the growth of N. Recalling (16), we can obtain that

$$\mathcal{L}(\mathcal{S})/N \le \frac{d}{2N}\log\frac{N}{2\pi} + \frac{c(\rho^{(j)})}{N} \to 0,\ N\to\infty.$$

This fact implies that the model selection method for GCM can achieve the optimal code assignment under sufficient samples, even though extensive context models are adopted by multi-directional structuring.

D. Approximation With Sequential NML

Since NML function is obtained via a normalization over all sequences of given length, it cannot derive random process for efficient computation. Thus, sequential NML (SNML [45]) is adopted for prediction in heterogeneous data compression with asymptotically equivalent model complexity. For each symbol x_t, its SNML function $f_{SNML}(x_t|x_1^{t-1})$ is obtained by

$$f_{SNML}(x_t|x_1^{t-1}) = \frac{f(x_1^t, \theta(x_1^t))}{\int_\Theta f(x_1^{t-1}, x, \theta(x_1^{t-1}, x))dx}. \quad (19)$$

The likelihood f for x_t is related with the estimated probability based on contexts with combinatorial structuring.

$$f_{SNML}\left(x_t|s^{(j)}\right) = \frac{Pr\left(x_t|s^{(j)}\right)}{\sum_{s^{(j)}\in\mathcal{A}^*} Pr\left(x_t|s^{(j)}\right)}. \quad (20)$$

Here, the estimated probability $Pr(x_t|s^{(j)})$ depends on the count of the most-occurrence symbol. For example, the estimated probability by KT-estimator for binary sources is obtained by comparing the probability of the emerging '0' and '1'.

$$Pr(x_t|s^{(j)}) = \max\left\{\frac{a_{s(j)}^{(t)} + \frac{1}{2}}{a_{s(j)}^{(t)} + b_{s(j)}^{(t)} + 1}, \frac{b_{s(j)}^{(t)} + \frac{1}{2}}{a_{s(j)}^{(t)} + b_{s(j)}^{(t)} + 1}\right\}. \quad (21)$$

Consequently, the negative logarithm of SNML function evaluates the context models in the sense of MDL principle.

$$\mathcal{L}_{SNML}\left(x_t|s^{(j)}\right) = -\log f_{SNML}\left(x_t|s^{(j)}\right) = -\log Pr\left(x_t|s^{(j)}\right) + \log \sum_{s^{(j)}\in\mathcal{A}^*} Pr\left(x_t|s^{(j)}\right). \quad (22)$$

In (22), the first term of the right-hand side is the code length led by context-based prediction and the second term defines the complexity of describing the contexts in GCM. Algorithm 1 describes SNML estimation for model selection and context-based prediction in GCM, where extensive contexts with multi-directional structuring are evaluated and selected for each symbol x_t on all the M directions. The optimal class of context models can be obtained by combining the selected contexts for each symbol x_t. Consequently, prediction based on such class of models $\hat{\mathcal{M}}$ minimizes the MDL evaluation $\mathcal{L}(x_1^N, \hat{\mathcal{M}})$ for x_1^N.

Algorithm 1: MDL-based Model Selection for GCM using Sequential NML Function

```
for t = 1···N do
    for j = 1···M do
        Generate context s^(j) in j-th direction from x_1^{t-1}.
        for s^(j) ∈ A* do
            Estimate Pr(x_t|s^(j)) for s^(j) with (21).
            Calculate log-SNML function L(x_t|s^(j)) for
            s^(j) with (20) and (22).
        end for
        Obtain ŝ^(j) with minimal log-SNML function.
    end for
    Obtain the context ŝ^(ĵ) with minimal log-SNML
    function.
    Predict x_t based on context ŝ^(ĵ).
end for
```

V. Model Redundancy for Generalized Context Modeling

Model redundancy is led by specifying the actual context model in the model class. In this section, we discuss the model redundancy for generalized context modeling with multi-directional structuring.

A. Model Redundancy for Combinatorial Structuring

In this subsection, model redundancy for combinatorial structuring is developed, where the number of directions M is set to one. Given the actual context model $\mathcal{S}_a$, the model redundancy is upper-bounded.

Proposition 3 (Model Redundancy for Combinatorial Structuring): For model class $\mathcal{M}$ with maximum order D, an upper bound of model redundancy led by combinatorial structuring is derived as

$$\bar{\rho}_{CS} = -\log \frac{P_w(x_1^N)}{\prod_{s\in\mathcal{S}_a} P_e(x_1^N|s)} \leq -L^D \log \frac{\eta^D}{(1+\eta)^D - 1}, \quad (23)$$

where L is the size of alphabet and η is the compensated weights for contexts with various lengths.

Proof: Please refer to Appendix C. ■

Equation (23) implies that the upper bound of model redundancy led by combinatorial structuring only depends on D, such that the per symbol model redundancy asymptotically vanishes with the growth of N. The upper bound $\bar{\rho}_{CS}$ can be also explained in the view of structure of $\mathcal{S}_a$. Since model redundancy is led by specifying the actual model $\mathcal{S}_a$, its upper bound indicates the cost for describing the actual model with maximal size. In (23), the upper bound is achieved for the context model constructed by L^D contexts with length D, which is the greatest subset of contexts for constructing a valid context model.

B. Model Redundancy for Multi-Directional Structuring

The upper bound of model redundancy can be generalized with multi-directional extension. Without loss of generality, we discuss the M-th directional case. Assuming that the maximum orders of contexts are D for all the directions. An upper bound of model redundancy for M-directional actual model $\mathcal{S}_a$ is developed.

Proposition 4 (Model Redundancy for Multi-Directional Extension): For model class $\mathcal{M}$ with number of directions M and maximum order D, the upper bound of model redundancy led by multi-directional extension is derived as

$$\bar{\rho}_{ME} = -\log \frac{P_w(x_1^N)}{\prod_{\mathbf{s}\in\mathcal{S}_a} P_e(x_1^N|\mathbf{s})} \leq -L^{MD} \log \frac{\eta^{MD}}{(1+\eta)^{MD} - 1}, \quad (24)$$

where L is the size of alphabet and η is the compensated weights for contexts with various lengths.

Proof: Please refer to Appendix D. ■

Equation (24) implies that the upper bound of model redundancy for GCM (with multi-directional structuring) is only related with the number of directions M and maximum order D. This fact means that the per symbol model redundancy is asymptotically forced to zero, as $\bar{\rho}_{ME}/N \to 0$ for $N \to \infty$. Analogically, $\bar{\rho}_{ME}$ upper bounds the cost for specifying the actual model with number of directions M and maximum order D. The upper bound is achieved when the actual model constructed by the largest valid subset of contexts. For M-directional case, L^D contexts with length D are required for each direction.

VI. Context Pruning for Model Selection

Table I and Fig. 2 show the configuration of generalized context models when the longest contexts available are 3 bytes, where d stands for the d-th symbols from current symbol and ✓ means that the symbol is chosen by the context models. For greater Ds, their tables can be analogically made according to Table I. The selected contexts for each symbol x_t are assigned a weight, such that the estimation is made by weighting over them. Also, the assigned weights are adjusted according to the evaluation with SNML, as shown in (20) and (14).

To determine the optimal combination of context structures, they are evaluated by comparing the performance obtained when they are included or excluded in the prediction. Algorithm 2 shows the process for excluding those redundant context structures. In each iteration, model classes based on various context structures are constructed for prediction and their performances are compared in terms of MDL evaluation for context pruning. Fig. 3 shows the pruning results for four files in Calgary corpus, which can be categorized into two typically

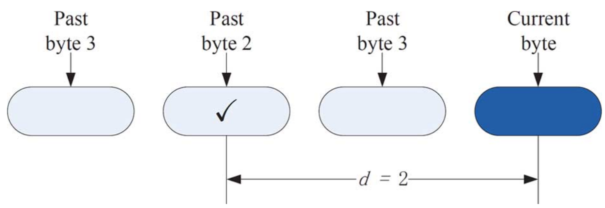


Fig. 2. An illustrative figure for model No. 6 in Table I, where d is the distance from current symbol. A symbol is included as a part of the context when it is checked.

TABLE I
GENERALIZED CONTEXT MODELS WITH DEPTH $D = 3$ BYTES

Symbol with Distance d \ Model No.	1	2	3	4	5	6	7
1	✓	✓	✓	✓			
2		✓	✓		✓	✓	
3			✓	✓	✓		✓

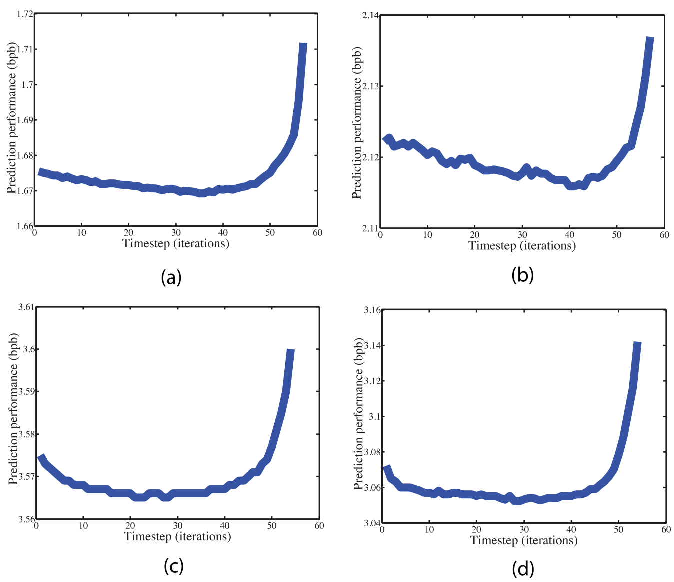


Fig. 3. Context model pruning for four files in Calgary corpus, (a) bib; (b) paper1; (c) geo; (d) obj1. For each file, context pruning is conducted by a greedy algorithm, which excludes the least effective context model at each timestep. The models are evaluated by the compression cost with regard to their estimated probabilities.

distinctive kinds of contexts. In each figure, context structure is iteratively excluded at each timestep, where the prediction performance after excluding each one is compared to determine the least effective context structure for prediction. In all the figures, the prediction performance increases at first, then holds in a period of time, and finally decreases sharply after a cut-off point. This fact implies that for heterogeneous data compression, a selected model class performs better in inference with less model complexity at the same time. In comparison to Figs. 3(a) and 3(b), the improvement of performance with a pruned model class is more obvious in Figs. 3(c) and 3(d). It means that the context pruning algorithm performs better for heterogeneous data. Moreover, the predictive performance after cutoff point degrades more sharply in Fig. 3(c) and 3(d), which means that heterogeneous data, e.g., *obj1* and *geo* , might be predicted with a class of model to describe the interlaced data streams. Consequently, generalized context modeling will improve the predictive performance for heterogeneous data compression in a more noticeable sense.

Algorithm 2: Context Pruning for Refining Model Class

Constructing the model class $\mathcal{M}$ of 2^D-1 context structures with their parameters and initializing $K=2^D-1$.
while $K > 0$ **do**
  Calculating cumulative weighted probability P_K based on all context models with K context structures as shown in Algorithm 1.
  Evaluating its cost in terms of MDL $\mathcal{L}_K$ by cumulating the cost for each symbol x_t as shown in Algorithm 1.
  Initializing $\mathcal{L}_{K-1} = \mathcal{L}_K$.
  for $k = 1 \cdots K$ **do**
    Excluding k-th context model from $\mathcal{M}$ and calculating cumulative weighted probability $P_{K-1}^{(k)}$ after exclusion.
    Evaluating its cost in terms of MDL $\mathcal{L}_{K-1}^{(k)}$.
    if $\mathcal{L}_{K-1} > \mathcal{L}_{K-1}^{(k)}$ **then**
      $\mathcal{L}_{K-1} = \mathcal{L}_{K-1}^{(k)}$.
      Storing k
    end if
  end for
  if $\mathcal{L}_K > \mathcal{L}_{K-1}$ **then**
    Removing k-th context model from $\mathcal{M}$.
    $K = K - 1$
  end if
end while

VII. APPLICATION INTO HETEROGENEOUS DATA COMPRESSION

A. Application Into Calgary Corpus

The comparisons between the proposed method and CTW are made by evaluating their compression performance for Calgary corpus [46], a collection of text and binary data files. The compression performance for Calgary corpus is shown when depth D is 3, 4, 5, and 6 bytes, respectively. In CTW, both results by zero redundancy (ZR) estimator and KT estimator are observed. Table II shows the detailed results. In the table, improvements of compression performance for text-like files (ASCII encoding) by the proposed method tend to decrease with the growth of depth D. In detail, the gap between the proposed method and the CTW estimators is about 4% to 6% for files, e.g., *bib*, *book1*, *book2*, *news*, *paper1 paper6*, *progc*, *progl*, and *progp* in Table II, but not more than 4% in Table II. However, it is not the case for non-ASCII files. The improvements over the CTW estimators are about 7%—9% for executable programs *obj1* and *obj2* and is up to 12% for the seismic data *geo*. Lite PAQ (LPAQ) improves the compression performance of data with homogeneous formats by mixing variable-order context models with an approximate matching model for long contexts. Table II shows that GCM outperforms LPAQ by 6%-10% for non-ASCII files. These results imply that the proposed models trivially improve the performance of text-like data, while they perform better in the non-ASCII files that contains complicated context structure for prediction.

TABLE II
COMPRESSION PERFORMANCE (BPB) FOR CALGARY CORPUS OBTAINED BY GCM, CTW WITH KT ESTIMATOR (CTW-KT), CTW WITH ZERO REDUNDANCY ESTIMATOR (CTW-ZR), PPMD, AND PAQ WITH DEPTH $D = 3$, 4, 5, AND 6 BYTES, RESPECTIVELY

	bib	book1	book2	geo	news	obj1	obj2	paper1	paper2
GCM	1.83	2.39	1.97	3.75	2.27	3.18	1.95	2.24	2.26
CTW-KT	2.15	2.46	2.28	4.54	2.70	3.84	2.76	2.54	2.46
CTW-ZR	2.11	2.45	2.25	4.55	2.67	3.74	2.68	2.46	2.41
PPMd	2.05	2.45	2.21	4.35	2.58	3.52	2.53	2.41	2.38
PAQ	1.84	2.40	1.97	4.26	2.29	3.40	2.07	2.24	2.27
LPAQ	1.87	2.23	1.88	4.04	2.30	3.51	2.24	2.21	2.18
	paper3	paper4	paper5	paper6	pic	progc	progl	progp	trans
GCM	2.52	2.76	2.84	2.26	0.70	2.19	1.43	1.41	1.22
CTW-KT	2.71	2.97	3.10	2.60	0.81	2.59	1.96	1.94	1.88
CTW-ZR	2.66	2.89	3.00	2.52	0.81	2.49	1.89	1.84	1.75
PPMd	2.62	2.85	2.95	2.46	0.79	2.42	1.81	1.77	1.68
PAQ	2.52	2.76	2.83	2.27	0.71	2.20	1.43	1.41	1.23
LPAQ	2.41	2.71	2.83	2.28	0.74	2.24	1.62	1.59	1.49

	bib	book1	book2	geo	news	obj1	obj2	paper1	paper2
GCM	1.73	2.23	1.82	3.59	2.16	3.13	1.83	2.15	2.16
CTW-KT	1.99	2.27	2.04	4.52	2.50	3.82	2.60	2.44	2.34
CTW-ZR	1.91	2.25	2.00	4.53	2.43	3.72	2.49	2.33	2.27
PPMd	1.83	2.25	1.96	4.33	2.32	3.50	2.35	2.25	2.23
PAQ	1.72	2.19	1.80	4.24	2.15	3.38	2.00	2.14	2.14
LPAQ	1.71	2.15	1.76	4.00	2.10	3.50	1.98	2.11	2.10
	paper3	paper4	paper5	paper6	pic	progc	progl	progp	trans
GCM	2.42	2.70	2.79	2.20	0.69	2.14	1.40	1.39	1.19
CTW-KT	2.62	2.94	3.07	2.52	0.80	2.50	1.84	1.87	1.71
CTW-ZR	2.53	2.82	2.94	2.41	0.81	2.38	1.74	1.74	1.55
PPMd	2.48	2.77	2.87	2.32	0.78	2.28	1.64	1.65	1.45
PAQ	2.41	2.70	2.77	2.19	0.70	2.12	1.39	1.37	1.18
LPAQ	2.35	2.67	2.78	2.17	0.68	2.11	1.37	1.38	1.17

	bib	book1	book2	geo	news	obj1	obj2	paper1	paper2
GCM	1.69	2.18	1.78	3.59	2.13	3.12	1.78	2.13	2.14
CTW-KT	1.94	2.22	1.97	4.52	2.46	3.83	2.55	2.42	2.32
CTW-ZR	1.85	2.20	1.92	4.53	2.37	3.72	2.42	2.30	2.24
PPMd	1.75	2.19	1.88	4.32	2.24	3.50	2.27	2.22	2.19
PAQ	1.71	2.17	1.78	4.24	2.14	3.38	1.99	2.13	2.13
LPAQ	1.69	2.14	1.74	3.99	2.08	3.50	1.97	2.11	2.10
	paper3	paper4	paper5	paper6	pic	progc	progl	progp	trans
GCM	2.40	2.69	2.78	2.18	0.68	2.12	1.39	1.37	1.17
CTW-KT	2.60	2.94	3.07	2.51	0.80	2.48	1.79	1.84	1.66
CTW-ZR	2.50	2.81	2.94	2.39	0.80	2.35	1.68	1.70	1.48
PPMd	2.45	2.76	2.85	2.29	0.76	2.25	1.57	1.60	1.36
PAQ	2.40	2.69	2.77	2.18	0.70	2.12	1.37	1.37	1.17
LPAQ	2.35	2.67	2.78	2.17	0.68	2.11	1.36	1.37	1.16

	bib	book1	book2	geo	news	obj1	obj2	paper1	paper2
GCM	1.68	2.16	1.77	3.58	2.12	3.08	1.76	2.13	2.13
CTW-KT	1.93	2.21	1.95	4.52	2.44	3.82	2.53	2.41	2.31
CTW-ZR	1.83	2.18	1.89	4.53	2.35	3.72	2.40	2.29	2.23
PPMd	1.74	2.18	1.86	4.32	2.35	3.50	2.23	2.20	2.18
PAQ	1.66	2.15	1.74	3.59	2.09	3.14	1.86	2.08	2.08
LPAQ	1.68	2.11	1.72	3.98	2.06	3.49	1.96	2.10	2.09
	paper3	paper4	paper5	paper6	pic	progc	progl	progp	trans
GCM	2.39	2.69	2.78	2.17	0.68	2.11	1.39	1.37	1.17
CTW-KT	2.60	2.94	3.07	2.51	0.80	2.47	1.77	1.82	1.63
CTW-ZR	2.50	2.82	2.93	2.37	0.80	2.33	1.65	1.68	1.44
PPMd	2.45	2.76	2.84	2.27	0.76	2.22	1.52	1.56	1.31
PAQ	2.34	2.62	2.72	2.12	0.49	2.08	1.35	1.34	1.17
LPAQ	2.35	2.66	2.77	2.16	0.68	2.10	1.34	1.36	1.15

B. Application Into Executable Compression

Fig. 4 sketches the compression performance under various depth D, including text files *bib* and *paper1*, seismic data *geo*, and executable program *obj1*. Figs. 4(a) and 4(b) show that the compression ratio of all the schemes increases with the growth of D. However, Figs. 4(c) and 4(d) imply that the classical methods like CTW estimators cannot exploit the correlations in the sources with complicated statistics, even though D is large. Moreover, Fig. 4 shows that the proposed model with a length of one or two bytes are both efficient in utilizing the correlation in complicated contexts at the cost of moderate complexity in the compression of Calgary corpus (each equivalent to a first order or second order model).

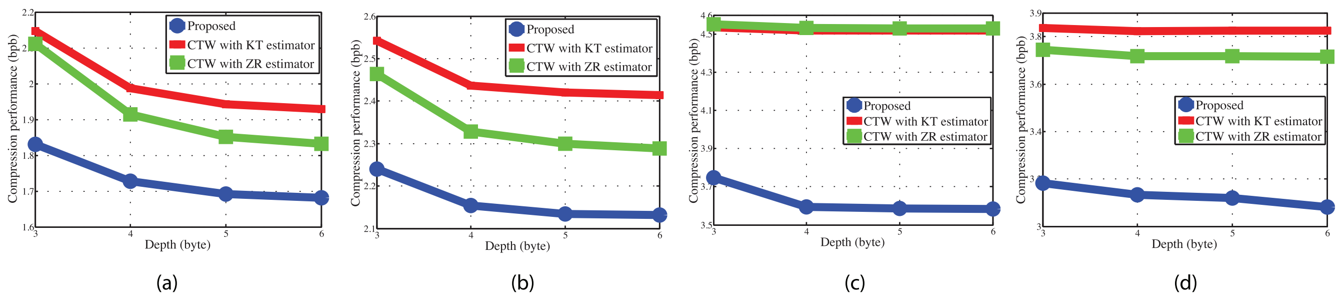


Fig. 4. Compression performance (bpb) comparison of the proposed method and CTW for four files in Calgary corpus, (a) bib; (b) paper1; (c) geo; (d) obj1. Compression performance is obtained under depth $D = 3, 4, 5, 6$.

TABLE III
GENERALIZED CONTEXT MODELS FOR EXECUTABLE FILE COMPRESSION

Model No. / Symbol with Distance d	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
2		✓	✓	✓	✓	✓					✓	✓		✓
3			✓	✓	✓	✓	✓				✓		✓	
4				✓	✓	✓		✓				✓	✓	
5					✓	✓			✓					
6						✓				✓				

The proposed method is further evaluated by applying it into compression of large executable files. Executable files are composed of interlaced data streams describing various data fields. Consequently, arbitrary structures of contexts are allowed to exploit the multi-directional correlations among them. In this application, M is set to 4 for correlations in the data fields of instruction opcodes, displacements, and immediate data and within the instructions, respectively. In each direction, depth D is set to 6 bytes and the candidate class of context models can refer to Table III. All the experiments are made under a 3.2GHz Intel core-i7 CPU with 40MB memory limitation. Table IV shows the compression performances for the proposed method and other benchmarks, e.g., WinRAR, PPMd, PPMonstr, CTW, LPAQ and PAQ [19]. It can be seen that compared with CTW estimators, the performance gain caused by GCM is about 12%-17%. Moreover, since we only select a combination of models to lower the complexity while maintaining compression efficiency, the time cost for GCM is close to the CTW estimators, though it involves additional models led by multi-directional structuring. It is noted that PPMd and PPMonstr are the improved version of PPM, where the former emphasizes on speed and the latter on modeling for non-stationary data sources (executable files are only one kind). Obviously, the complex context structure in the executables hampers the two PPM-based compressor in performance, while GCM is obviously better, exceeding PPMd by 10% and PPMonstr by 4% in general. As for the complexity in context weighting, the proposed scheme is about 3 times the time cost of PPMonstr. LPAQ provides a comparable results to PPMonstr with a lower complexity, but there is a gap of 0.3–0.5 bpb in coding performance between LPAQ and GCM. Considering that the speed is about 100 KB/s-110 KB/s, however, it is enough for many applications such as network transmission with the 1 Mbps bandwidth. PAQ8 combines probabilities with neural network, and thus, reaches a high compression ratio; however, it also leads to massive computation and memory

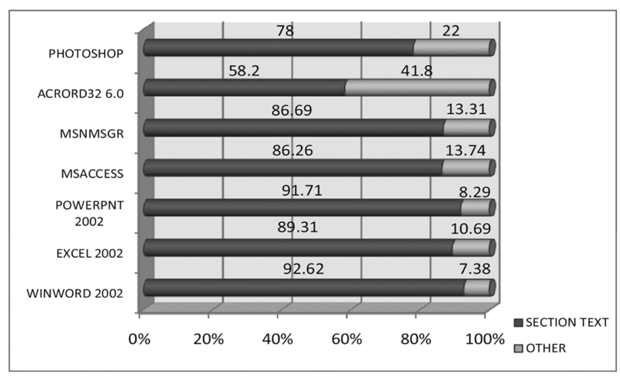


Fig. 5. Proportion of Section Text in executable files.

utilization. While the proposed scheme is comparable to PAQ8, its time cost is about one-third PAQ8. In particular, there are about 1.7% and 1.4% discrepancies for ACRORD32.EXE and PHOTOSHOP.EXE. It is derived from the reason that the size of Section Text in the two files is less than the other executables, the detailed proportion can be found in Fig. 5.

C. Application Into DNA Sequences Compression

In this subsection, GCM is employed on genomic data compression. DNA sequences are composed of repeated patters of four difference kinds of nucleotides, namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), with the exception of insertion, deletion and substitution. This fact means that genomic data can be divided into repeatable patterns of nucleotides and regular non-repeat regions. Thus, multi-directional structuring with $M = 2$ is adopted. Table V shows the results for the standard dataset in most DNA compression publications, where GCM is compared with Finite-Context Models (FCM [47]), CTW+LZ [11], CTW [9], GZIP [48] and PPMd. CTW-LZ improves CTW with Lempel-Ziv algorithm [49] to

TABLE IV
COMPRESSION PERFORMANCE FOR EXECUTABLE FILES IN BITS PER BYTE (BPB)

	WINWORD 2002	EXCEL 2002	POWERPNT 2002	MSACCES	MSNMSGR 8.0	ACRORD32 6.0	PHOTOSHOP 8.0
GCM	2.94	3.06	2.51	2.67	2.03	2.54	2.14
PAQ8	2.98	3.10	2.52	2.70	2.00	2.41	2.04
LPAQ	3.25	3.43	3.06	3.04	2.52	2.82	2.45
PPMd	3.69	3.79	3.37	3.37	2.78	3.26	2.77
PPMonstr	3.25	3.40	2.98	2.98	2.38	2.85	2.39
WinRAR	3.54	3.77	3.16	3.28	2.46	2.91	2.57
CTW-KT	4.02	4.12	3.68	3.66	3.24	3.95	3.50
CTW-ZR	4.11	4.14	3.69	3.67	3.24	4.04	3.50

TABLE V
COMPRESSION PERFORMANCE FOR DNA SEQUENCE IN BITS PER BYTE (BPB)

Sequence	Size (byte)	GCM	FCM	CTW+LZ	GZIP	PPMd	CTW
CHMPXX	121024	1.63	1.63	1.67	2.22	1.87	1.84
CHNTXX	155844	1.60	1.63	1.61	2.29	1.96	1.93
HEHCMVCG	229354	1.81	1.85	1.84	2.28	1.98	1.96
HUMDYSTROP	38770	1.91	1.93	1.92	2.38	1.98	1.92
HUMHBB	73308	1.82	1.87	1.81	2.23	1.96	1.89
MPOMTCG	186609	1.89	1.92	1.90	2.28	1.99	1.96
PANMTPACGA	100314	1.84	1.86	1.86	2.23	1.90	1.87
VACCG	191737	1.75	1.77	1.76	2.19	1.90	1.86
Total	1096960	1.78	1.81	1.79	2.25	1.93	1.91

fit the structure of approximate repeats in DNA sequence, and FCM rapidly captures variable-order statistical information along the DNA sequences. Table V shows that the proposed method performs better than all the other algorithms in average.

D. Computational Complexity

The computation complexity of GCM is based on M and D, which is proportional to the number of context models. To be concrete, its complexity is $O(2^{MD})$ for GCM with multi-directional structuring provided in Definition 4. It can be further reduced with separable context modeling, which makes the number of context models grows linearly with M. Thus, the complexity is $O(M \cdot 2^{D})$ in practical applications.

In practice, GCM operates on a PC with a 3.2 GHz Intel Core i7 processor and complied with $VC++$ 9.0 with "DEBUG" configuration. Table VI shows and compares the encoding time for Calgary corpus obtained by GCM, CTW, PPMd, LPAQ and PAQ. M is set to one for a fair comparison to the benchmarks, and the results are obtained under $D = 3$ and $D = 6$, respectively. It shows that GCM can achieve better compression performance at the cost of 2 to 24 times and 4–10 time the computational complexity in comparison to PPMd and LPAQ, respectively. When compared with PAQ, GCM can obtain competitive results with approximately 60% less complexity. These facts imply that GCM can make a proper tradeoff between compression performance and computational complexity.

VIII. CONCLUSION

In this paper, we propose the generalized context modeling for heterogeneous data compression, which adapts its structure and parameters to the specific sources. The context-based prediction is based on the subset of context models derived from the alleged model class with dynamical pruning in terms of MDL evaluation. The classical context modeling is generalized with multi-directional extension and combinatorial structuring, such that extensive context models are generated to exploit interlaced correlations in heterogeneous data. In order to derive the estimated probability for prediction, model graph is designed to constrain the adoption of contexts in GCM. For generalized context modeling, the model selection algorithm for GCM is developed to obtain the optimal class of models in MDL sense, especially for data with large sizes. For generalized context modeling, the model selection algorithm for GCM is developed to obtain the optimal class of models in MDL sense. We also develop the additional upper bound of model redundancy, which is proven to be related to the number of directions M and the maximum order D of Makrov sources. Moreover, the potential of separable prediction for GCM is demonstrated. Consequently, the divergence between the class of selected models by SNML and the actual distributions is proven to be independent of the size of sequence.

APPENDIX A
PROOF OF PROPOSITION 1

Proposition 1 holds if (9) has at least one solution. Denote $\bar{\mathcal{H}} = (\mathcal{H}\ \mathbf{p}_{marg})$ the augmented matrix derived from (9). The necessary and sufficient conditions that there exists at least one vector $\mathbf{p}_{est}$ satisfying (9) is

$$rank(\mathcal{H}) = rank(\bar{\mathcal{H}}),$$

where $rank(\cdot)$ is the rank of a matrix.

Firstly, we consider the $U \times V$ coefficient matrix $\mathcal{H}$. Since $V > U$ for $M > 1$, $rank(\mathcal{H})$ depends on the rank of its row vectors. Denote $\mathbf{h}_u^{(j)}$ the $1 \times V$ row vector corresponding to $Pr(x_1^N|s^{(j)} = c_u^{(j)})$ in vector $\mathbf{p}_{marg}$. For $\mathbf{h}_u^{(j)}$,

$$Pr\left(x_1^N|s^{(j)} = c_u^{(j)}\right) = \sum_{\mathbf{s}^{\backslash(j)}} Pr\left(x_1^N|\mathbf{s}^{\backslash(j)}, s^{(j)} = c_u^{(j)}\right) Pr\left(\mathbf{s}^{\backslash(j)}\right).$$

TABLE VI

ENCODING TIME T (SEC) AND RUN-TIME RATIO R OF GCM AND BENCHMARK METHODS CTW, PPMD, PAQ, AND LPAQ FOR THE CALGARY CORPUS UNDER $M = 1$ AND $D = 3$ AND 6, RESPECTIVELY. RUN-TIME RATIO R IS ASSESSED AS: $R = T_{\text{GCM}}/T_{\text{BENCHMARK}}$

		bib	book1	book2	geo	news	obj1	obj2	paper1	paper2
GCM	T	1.26/1.99	7.20/11.46	5.92/9.40	1.10/1.77	3.63/5.70	0.20/0.32	2.46/3.72	0.64/1.01	0.88/1.47
CTW-KT	T	0.20/0.40	1.40/3.30	1.10/2.40	0.20/0.30	0.70/1.50	0.10/0.10	0.50/0.80	0.10/0.20	0.20/0.30
	R	6.30/4.98	5.14/3.47	5.38/3.92	5.50/5.90	5.19/3.80	2.00/3.20	4.92/4.65	6.40/5.05	4.40/4.90
CTW-ZR	T	0.30/0.40	1.50/3.30	1.20/2.30	0.30/0.30	0.80/1.40	0.10/0.10	0.50/0.80	0.20/0.20	0.30/0.30
	R	4.20/4.98	4.80/3.47	4.93/4.09	3.67/5.90	4.54/4.07	2.00/3.20	4.92/4.65	3.20/5.05	2.93/4.90
PPMd	T	0.06/0.10	0.35/0.72	0.30/0.51	0.18/0.20	0.24/0.40	0.03/0.04	0.20/0.27	0.04/0.06	0.05/0.09
	R	21.00/19.90	20.57/15.92	19.73/18.43	6.11/8.85	15.13/14.25	6.67/8.00	12.30/13.78	16.00/16.83	17.60/16.33
PAQ	T	3.10/4.87	21.07/33.40	17.09/26.51	2.80/4.47	10.40/16.27	0.59/0.93	6.81/10.53	1.46/2.26	2.25/3.49
	R	0.41/0.41	0.34/0.34	0.35/0.36	0.39/0.40	0.35/0.35	0.34/0.34	0.36/0.35	0.44/0.45	0.39/0.42
LPAQ	T	0.13/0.21	0.94/1.22	0.85/0.98	0.15/0.20	0.53/0.59	0.05/0.07	0.39/0.43	0.08/0.11	0.13/0.17
	R	9.69/9.48	7.66/9.39	6.96/9.59	7.33/8.85	6.85/9.66	4.00/4.57	6.31/8.65	8.00/9.18	6.77/8.65
		paper3	paper4	paper5	paper6	pic	progc	progl	progp	trans
GCM	T	0.52/0.83	0.17/0.28	0.11/0.19	0.44/0.65	5.36/8.82	0.39/0.61	0.71/1.12	0.48/0.75	1.07/1.70
CTW-KT	T	0.10/0.20	0.10/0.10	0.10/0.10	0.10/0.10	0.70/1.20	0.10/0.10	0.10/0.20	0.10/0.20	0.20/0.30
	R	5.20/4.15	1.70/2.80	1.10/1.90	4.40/6.50	7.66/7.35	3.90/6.10	7.10/5.60	4.80/3.75	5.35/5.67
CTW-ZR	T	0.10/0.20	0.10/0.10	0.10/0.10	0.10/0.10	0.80/1.20	0.10/0.10	0.20/0.20	0.10/0.10	0.30/0.30
	R	5.20/4.15	1.70/2.80	1.10/1.90	4.40/6.50	6.70/7.35	3.90/6.10	3.55/5.60	4.80/7.50	3.57/5.67
PPMd	T	0.03/0.06	0.02/0.03	0.01/0.02	0.03/0.04	0.23/0.26	0.03/0.04	0.04/0.06	0.03/0.04	0.05/0.07
	R	17.33/13.83	8.50/9.33	11.00/9.50	14.67/16.25	23.30/33.92	13.00/15.25	17.75/18.67	16.00/18.75	21.40/24.29
PAQ	T	1.28/1.98	0.37/0.57	0.33/0.52	1.05/1.62	14.66/22.47	1.09/1.69	2.00/3.06	1.39/2.11	2.63/3.99
	R	0.41/0.42	0.46/0.49	0.33/0.37	0.42/0.40	0.37/0.39	0.36/0.36	0.36/0.37	0.35/0.36	0.41/0.43
LPAQ	T	0.08/0.11	0.04/0.04	0.02/0.05	0.07/0.10	0.63/0.72	0.07/0.11	0.10/0.15	0.07/0.13	0.13/0.18
	R	6.50/7.55	4.25/7.00	5.50/3.80	6.29/6.50	8.51/12.25	5.57/5.55	7.10/7.47	6.86/5.77	8.23/9.44

For each pair of numbers, the left one represents $D = 3$ case and the right one for $D = 6$ case.

Since M interlaced Markov sources are not correlated, the subset $\{\mathbf{h}_u^{(m)}\}$ for m-th direction satisfies

$$\sum_{u=1}^{R} Pr\left(s^{(j)} = c_u^{(j)}\right)\mathbf{h}_u^{(j)} = (Pr(\mathbf{s}))_{\mathbf{s}\in(\mathcal{A}^*)^M}, \tag{25}$$

where $\{c_u^{(j)} \in \mathcal{A}^*\}$ are the valid values for contexts in j-th direction. For any pair $j \neq j'$, it can be obtained from (25)

$$\sum_{u=1}^{R} Pr\left(s^{(j)} = c_u^{(j)}\right)\mathbf{h}_u^{(j)} = \sum_{u'=1}^{R} Pr\left(s^{(j')} = c_{u'}^{(j')}\right)\mathbf{h}_{u'}^{(j')}.$$

According to the construction of $\mathbf{p}_{est}$ and $\mathbf{p}_{marg}$, it holds $\mathbf{h}_u^{(j)} \neq \mathbf{h}_{u'}^{(j')}$ for any proper pairs $j \neq j'$ and $u \neq u'$. This fact means that, for any pair $j \neq j'$, $\mathbf{h}_u^{(j)}$ cannot be represented by the linear combination of $\{\mathbf{h}_u^{(j')}\}$. Consequently, the rank of coefficient matrix $\mathcal{H}$ is $U - M$.

On the other hand, denote $\bar{\mathbf{h}}_u^{(j)} = (\mathbf{h}_u^{(j)} \quad P(x_1^N|s^{(j)} = c_u^{(j)}))$ the corresponding vector in the augmented matrix $\bar{\mathcal{H}}$. For the given sequence x_1^N, its probability $P(x_1^N)$ is fixed. Consequently, it fulfills for arbitrary j

$$\sum_{c_u^{(j)}\in\mathcal{A}^*} Pr\left(x_1^N|s^{(j)}=c_u^{(j)}\right) Pr\left(s^{(j)}=c_u^{(j)}\right) = Pr(x_1^N) = C.$$

Such that, it still holds that

$$\sum_{u=1}^{R} Pr\left(s^{(j)} = c_u^{(j)}\right)\bar{\mathbf{h}}_u^{(j)} = \left((Pr\left(\mathbf{s}\right))_{\mathbf{s}\in(\mathcal{A}^*)^M} \quad C\right). \tag{26}$$

Since $rank(\mathcal{H}) \leq rank(\bar{\mathcal{H}})$, the rank of augmented matrix $\bar{\mathcal{H}}$ is also $U - M$. As a result, it draws that $rank(\mathcal{H}) = rank(\bar{\mathcal{H}})$, which comes to Proposition 1.

APPENDIX B
PROOF OF PROPOSITION 2

Without loss of generality, we consider $|\Gamma_j(\theta)|$ in the j-th direction. If the model order d is greater than the actual order d_j^* of source π_j, [50] shows that the ML estimator of $\rho^{(j)}$ follows that $\hat{\rho^{(j)}}(x_1^N) \rightarrow |\rho^{(j)}|_\infty$ almost surely as $N \rightarrow \infty$. Since $\|\rho^{(j)}\|_\infty \leq \xi^{(j)}$, it can obtain for ML estimator of η $\hat{\xi}(x_1^N) = \hat{\rho^{(j)}}(x_1^N) \rightarrow |\rho^{(j)}|_\infty$. For $d \leq d_j^*$, the ML estimator $\hat{\rho^{(j)}}(x_1^N)$ minimizes the KL divergence from the actual distribution with order d_j^*. Therefore, $\hat{\eta}(x_1^N)$ converges to $\|\arg\min_\rho\{\mathbb{D}_{KL}(p_\rho\|\pi_j)\}\|_\infty$. This means that the Fisher information matrix is constant with N in the j-th direction.

Moreover, it can be obtained from the convergence in the j-th direction,

$$\int_{\Theta(\eta)} \sqrt{|\Gamma_j(\theta)|}d\theta \leq c\left(\rho^{(j)}\right)'. \tag{27}$$

Consequently, for $1 \leq j \leq M$,

$$\int_{\Theta(\eta)} \sqrt{|\Gamma_j(\theta)|}d\theta \leq \max_j c\left(\rho^{(j)}\right)'. \tag{28}$$

As a result, we can draw the conclusion that the NML estimation of x_1^N is constant with N in GCM.

APPENDIX C
PROOF OF PROPOSITION 3

Consider that the actual context model $\mathcal{S}_a$ with number of contexts $|\mathcal{S}_a|$ and maximum order d_a is in the model class $\mathcal{M}$. The model redundancy led by combinatorial structuring is

$$\rho_{CS} = -\log\frac{\sum_{\mathcal{S}\in\mathcal{M}} w(\mathcal{S})\prod_{s\in\mathcal{S}} Pr(x_1^N|s)}{\prod_{s\in\mathcal{S}_a} Pr(x_1^N|s)}. \tag{29}$$

Combining (29) and (21), the weighted estimated probability in ρ_{CS} can be rewritten in a product form.

$$\rho_{CS} = -\log \frac{\sum_{\mathcal{S}\in\mathcal{M}} \prod_{s\in\mathcal{S}} w(s) Pr(x_1^N|s)}{\prod_{s\in\mathcal{S}_a} Pr(x_1^N|s)}. \quad (30)$$

According to (7), its upper bound $\bar{\rho}_{CS}$ is derived.

$$\begin{aligned}\bar{\rho}_{CS} &= \sup_{\mathcal{S}_a\in\mathcal{M}} \rho_{CS} \\ &\le \sup_{\mathcal{S}_a\in\mathcal{M}} \left[-\log \prod_{s\in\mathcal{S}_a} w(s)\right] = \sup_{\mathcal{S}_a\in\mathcal{M}} \left[-\sum_{s\in\mathcal{S}_a} \log w(s)\right].\end{aligned}$$

For $\mathcal{S}_a$ over $\mathcal{A} = \{a_1, \ldots, a_L\}$, its contexts are constrained.

$$\sum_{s\in\mathcal{S}_a} L^{d_a - l(s)} = L^{d_a}.$$

Consequently, a tight upper bound for $\bar{\rho}_{CS}$ can be obtained by solving

$$\begin{cases} \max_{\mathcal{S}_a\in\mathcal{M}} -\sum_{s\in\mathcal{S}_a} \log w(s) \\ s.t. \sum_{s\in\mathcal{S}_a} L^{d_a - l(s)} = L^{d_a}, \quad 1 \le d_a \le D. \end{cases} \quad (31)$$

Such that the upper bound for model redundancy is developed

$$\bar{\rho}_{CS} \le -L^D \log \frac{\eta^D}{(1+\eta)^D - 1}. \quad (32)$$

Only when $d_a = D$ and $l(s) = D, \forall s \in \mathcal{S}_a$, $\bar{\rho}_{CS}$ achieves the upper bound.

APPENDIX D
PROOF OF PROPOSITION 4

Denote $\mathcal{S}_a^{(j)}$ the actual model in the j-th direction with number of contexts $|\mathcal{S}_a^{(j)}|$ and maximum order $d_a^{(j)}$. The model redundancy led by multi-directional extension is

$$\rho_{ME} = -\log \frac{\sum_{\mathcal{S}\in\mathcal{M}} w(\mathcal{S}) \prod_{\mathbf{s}\in\mathcal{S}} Pr(x_1^N|\mathbf{s})}{\prod_{\mathbf{s}\in\mathcal{S}_a} Pr(x_1^N|\mathbf{s})}, \quad (33)$$

where $\mathcal{S}_a = \mathcal{S}_a^{(1)} \times \cdots \times \mathcal{S}_a^{(M)}$ is the multi-directional actual model. Similar to Proposition 3, (33) can be rewritten as

$$\rho_{ME} = -\log \frac{\sum_{\mathcal{S}\in\mathcal{M}} \prod_{\mathbf{s}\in\mathcal{S}} w(\mathbf{s}) Pr(x_1^N|\mathbf{s})}{\prod_{\mathbf{s}\in\mathcal{S}_a} Pr(x_1^N|\mathbf{s})}. \quad (34)$$

Consequently, the model redundancy led by multi-directional extension is upper-bounded.

$$\begin{aligned}\bar{\rho}_{ME} &= \sup_{\mathcal{S}_a\in\mathcal{M}} \rho_{ME} \\ &\le \sup_{\mathcal{S}_a\in\mathcal{M}} \left[-\log \prod_{\mathbf{s}\in\mathcal{S}_a} w(\mathbf{s})\right] = \sup_{\mathcal{S}_a\in\mathcal{M}} \left[-\sum_{\mathbf{s}\in\mathcal{S}_a} \log w(\mathbf{s})\right].\end{aligned}$$

Such that the tight upper bound for $\bar{\rho}_{ME}$ is obtained by solving

$$\begin{cases} \max_{\mathcal{S}_a\in\mathcal{M}} -\sum_{\mathbf{s}\in\mathcal{S}_a} \log w(\mathbf{s}) \\ \text{s.t.} \sum_{\mathbf{s}\in\mathcal{S}_a} \prod_{j=1}^{M} L^{d_a^{(j)} - l(s^{(j)})} = L^{\sum_{j=1}^{M} d_a^{(j)}}, \quad 1 \le d_a^{(j)} \le D \end{cases} \quad (35)$$

where $\mathbf{s} \in (\mathcal{A}^*)^M$ is the M-directional context, where $s^{(j)}$ is the context in the j-th direction with length $l(s^{(j)})$. The upper bound for model redundancy led by multi-directional extension is derived.

$$\bar{\rho}_{ME} \le -L^{MD} \log \frac{\eta^{MD}}{(1+\eta)^{MD} - 1}. \quad (36)$$

The upper bound is achieved when $d_a^{(j)} = D$ and $l(s^{(j)}) = D$ for arbitrary $s^{(j)} \in \mathcal{S}_a^{(j)}$ and $1 \le j \le M$.

REFERENCES

[1] M. J. Weinberger, J. Rissanen, and M. Freder, "A universal finite memory source," *IEEE Trans. Inf. Theory*, vol. 41, pp. 634–652, May 1995.

[2] J. Rissanen, "A universal data compression systme," *IEEE Trans. Inf. Theory*, vol. 29, pp. 656–664, Sep. 1983.

[3] H. Jégou and C. Guillemot, "Robust multiplexed codes for compression of heterogeneous data," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1393–1407, Apr. 2005.

[4] M. Drinic, D. Kirovski, and H. Vo, "PPMexe: Program compression," *ACM Trans. Program. Lang. Syst.*, vol. 29, no. 1 article, 3, Jan. 2007.

[5] T. M. Cover and A. Shenhar, "Compound Bayes predictors for sequences with apparent Markov structure," *IEEE Trans. Syst. Man, Cybern.*, vol. 7, no. 6, pp. 421–424, Jun. 1977.

[6] M. Freder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1258–1270, Jul. 1992.

[7] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Trans. Commun.*, vol. 32, no. 4, pp. 396–402, Apr. 1984.

[8] A. Moffat, "Implementing the PPM data compression scheme," *IEEE Trans. Commun.*, vol. 38, pp. 1917–1921, Nov. 1990.

[9] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, pp. 653–664, May 1995.

[10] M. Drinic and D. Kirovski, "PPMexe: PPM for compressing software," in *Proc. Data Compression Conf.*, Mar. 2002, pp. 192–201.

[11] T. Matsumoto, K. Sadakane, and H. Imai, "Biological sequence compression algorithms," *Genome Informatics*, vol. 11, pp. 43–52, Dec. 2000.

[12] Y. Zhang and D. A. Adjeroh, "Prediction by partial approximate matching for lossless image compression," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 924–935, Jun. 2008.

[13] J. Yu and S. Verdu, "Schemes for bidirectional modeling of discrete stationary sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4789–4807, Nov. 2006.

[14] O. Dekel, S. Shalev-Shwartz, and Y. Singer, "Individual sequence prediction using memory-efficient context trees," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5251–5262, Nov. 2009.

[15] E. Ordentlich, M. J. Weinberger, and C. Chang, "On multi-directional context sets," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6827–6836, Oct. 2011.

[16] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "Context weighting for general finite-context sources," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1514–1520, Sep. 1996.

[17] J. Veness, K. S. Ng, M. Huttler, and M. Bowling, "Context tree switching," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Apr. 2012, pp. 327–336.

[18] M. Bellemare, J. Veness, and E. Talvitie, "Skip context tree switching," in *Proc. 31st Int. Conf. Mach. Learn. (ICML'14)*, Beijing, China, Jun. 2014, pp. 1458–1466.

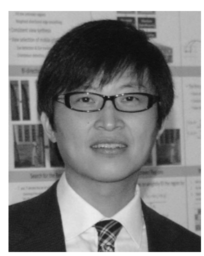
[19] M. V. Mahoney, Adaptive Weighing of Context Models for Lossless Data Compresson [Online]. Available: http://www.cs.fit.edu/mmahoney/compression

[20] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[21] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[22] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 629–636, Jul. 1984.
[23] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.
[24] A. Beirami and F. Fekri, "Results on the redundancy of universal compression for finite-length sequences," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, St. Petersburg, Russia, Jul. 2011, pp. 1504–1508.
[25] T. Cover and J. Thomas, *Elemets of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.
[26] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[27] Q. Ding and S. Kay, "Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1959–1969, May 2011.
[28] J. Rissanen, "MDL denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Jul. 2000.
[29] T. Roos, P. Myllymäaki, and J. Rissanen, "MDL denoising revisited," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3347–3360, Sep. 2009.
[30] D. F. Schmidt and E. Makalic, "The consistency of MDL for linear regression models with increasing signal-to-noise ratio," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1508–1510, Mar. 2012.
[31] J. I. Myung, D. J. Navarro, and M. A. Pitt, "Model selection by normalized maximum likelihood," *J. Math. Psychol.*, vol. 50, pp. 167–179, 2006.
[32] F. Haddadi, M. Malek-Mohammadi, M. M. Nayebi, and M. R. Aref, "Statistical performance analysis of MDL source enumeration in array processing," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 452–457, Jan. 2010.
[33] D. F. Schmidt and E. Makalic, "Estimating the order of an autoregressive model using normalized maximum likelihood," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 479–487, Feb. 2011.
[34] I. Ramirez and G. Sapiro, "An MDL framework for sparse coding and dictionary learning," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2913–2927, Jun. 2012.
[35] X. Wu, G. Zhai, X. Yang, and W. Zhang, "Adaptive sequential prediction of multidimensional signals with applications to lossless image coding," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 36–42, Jan. 2011.
[36] G. I. Shamir and D. J. Costello Jr., "Asymptotically optimal low-complexity sequential lossless coding for piecewise-stationary memoryless sources—Part I: The regular case," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2444–2467, Nov. 2000.
[37] G. I. Shamir and N. Merhav, "Low-complexity sequential lossless coding for piecewise-stationary memoryless sources," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1498–1519, Jul. 1999.
[38] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 714–722, May 1995.
[39] R. E. Krichevsky and V. K. Trofimov, "The performance of universal coding," *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 199–207, Mar. 1981.
[40] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
[41] Q. Xie and A. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, Mar. 1997.
[42] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.
[43] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1712–1717, Jul. 2001.
[44] B. Porat and B. Friedlander, "Computation of the exact information matrix of gaussian time series with stationary random components," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 118–130, 1986.
[45] T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," presented at the 1st Workshop Inf. Theoretic Methods Sci. Eng. (WITMSE-2008), Tampere, Finland, Aug. 2008.
[46] The Data Compression Resource on the Internet [Online]. Available: http://www.data-compression.info/Corpora/CalgaryCorpus/index.html
[47] A. J. Pinho, A. Neves, C. Bastos, and P. Ferreira, "DNA coding using finite-context models and arithmetic coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Apr. 2009, pp. 1693–1696.
[48] J. Gailly and M. Adler, GNU zip Jul. 2003 [Online]. Available: http://www.gzip.org
[49] J. Ziv, "A universal prediction lemma and applications to universal data compression and prediction," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1528–1532, May 2001.
[50] J. Rissanen and P. E. Caines, "The strong consistency of maximum likelihood estimators for ARMA processes," *Ann. Statist.*, vol. 7, no. 2, pp. 297–315, 1979.

Wenrui Dai (M'15) received B.S., M.S., and Ph.D. degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China in 2005, 2008, and 2014. He is currently a postdoctoral scholar with the Department of Biomedical Informatics, University of California, San Diego. His research interests include learning-based image/video coding, image/signal processing and predictive modeling.



Hongkai Xiong (M'01–SM'10) received the Ph.D. degree in communication and information system from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2003. Since then, he has been with the Department of Electronic Engineering, SJTU, where he is currently a Full Professor. From December 2007 to December 2008, he was with the Department of Electrical and Computer Engineering, Carnegie Mellon University (CMU), Pittsburgh, PA, USA, as a Research Scholar. From 2011 to 2012, he was a Scientist with the Division of Biomedical Informatics at the University of California (UCSD), San Diego, CA, USA.

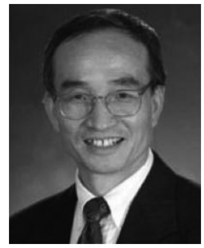
His research interests include source coding/network information theory, signal processing, computer vision and machine learning. He has published over 130 refereed journal/conference papers. He is the recipient of the Best Student Paper Award at the 2014 IEEE Visual Communication and Image Processing (IEEE VCIP'14), the Best Paper Award at the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (IEEE BMSB'13), and the Top 10% Paper Award at the 2011 IEEE International Workshop on Multimedia Signal Processing (IEEE MMSP'11).

In 2014, he was granted National Science Fund for Distinguished Young Scholar and Shanghai Youth Science and Technology Talent as well. In 2013, he was awarded a recipient of Shanghai Shu Guang Scholar. From 2012, he is a member of Innovative Research Groups of the National Natural Science. In 2011, he obtained the First Prize of the Shanghai Technological Innovation Award for "Network-oriented Video Processing and Dissemination: Theory and Technology". In 2010 and 2013, he obtained the SMC-A Excellent Young Faculty Award of Shanghai Jiao Tong University. In 2009, he was awarded a recipient of New Century Excellent Talents in University, Ministry of Education of China. He served as TPC members for prestigious conferences such as ACM Multimedia, ICIP, ICME, and ISCAS.

Jia Wang received the B.Sc. degree in electronic engineering, the M.S. degree in pattern recognition and intelligence control, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 1997, 1999, and 2002, respectively.

He is currently an Associate Professor of the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, and a member of Shanghai Key Laboratory of Digital Media Processing and Transmission. His research interests include multiuser information theory and its application in video coding.

Samuel Cheng (S'00–M'04–SM'12) received the B.S. degree in Electrical and Electronic Engineering from the University of Hong Kong, and the M.Phil. degree in Physics and the M.S. degree in Electrical Engineering from Hong Kong University of Science and Technology and the University of Hawaii, Honolulu, respectively. He received the Ph.D. degree in Electrical Engineering from Texas A&M University in 2004. He worked in Microsoft Asia, China, and Panasonic Technologies Company, New Jersey, in the areas of texture compression and digital watermarking during the summers of 2000 and 2001. In 2004, he joined Advanced Digital Imaging Research, a research company based near Houston, Texas, as a Research Engineer to perform biomedical imaging research and was promoted to Senior Research Engineer the next year. He is currently with the University of Oklahoma and the Tongji University in Shanghai. He has been awarded six US patents in miscellaneous areas of signal processing. His research interests include image/signal processing, pattern recognition, and information theory.



Yuan F. Zheng (F'97) received the MS and Ph.D. degrees in Electrical Engineering from The Ohio State University, in Columbus, Ohio in 1980 and 1984, respectively. His undergraduate education was received at Tsinghua University, Beijing, China in 1970. From 1984 to 1989, he was with the Department of Electrical and Computer Engineering at Clemson University, Clemson, South Carolina. Since August 1989, he has been with The Ohio State University, where he is currently Professor and was the Chairman of the Department of Electrical and Computer Engineering from 1993 to 2004. From 2004 to 2005, Professor Zheng spent sabbatical year at the Shanghai Jiao Tong University in Shanghai, China and continued to be involved as Dean of School of Electronic, Information and Electrical Engineering until 2008. Professor Zheng is an IEEE Fellow.

Professor Zheng's research interests include two aspects. One is in wavelet transform for image and video, and object classification and tracking, and the other is in robotics which includes robotics for life science applications, multiple robots coordination, legged walking robots, and service robots. Professor Zheng was and is on the editorial board of five international journals. Professor Zheng received the Presidential Young Investigator Award from Ronald Reagan in 1986, and the Research Awards from the College of Engineering of The Ohio State University in 1993, 1997, and 2007, respectively. Professor Zheng along with his students received the best conference and best student paper award a few times in 2000, 2002, and 2006, and received the Fred Diamond for Best Technical Paper Award from the Air Force Research Laboratory in Rome, New York in 2006. In 2004, Professor Zheng was appointed to the International Robotics Assessment Panel by the NSF, NASA, and NIH to assess the robotics technologies worldwide in 2004 and 2005.