Near-lossless Compression for Sparse Source Using Convolutional Low Density Generator Matrix Codes

Tingting Zhu and Xiao Ma
Sun Yat-sen University
Guangzhou, China
zhutt@mail2.sysu.edu.cn, maxiao@mail.sysu.edu.cn
Outline

• Problem Statement
• Related Research
• Convolutional LDGM Codes
• Numerical Results
• Conclusions and Future Work
Problem Statement

➢ Source & Entropy:

- A Bernoulli source, denoted as
  \[ U = U_0, U_1, \ldots, \] where \( U_t \in \mathbb{F}_2 \triangleq \{0, 1\} \) for \( t \geq 0 \),

  is independent and identically distributed (i.i.d.) according to \( P_U(1) = \theta \) and \( P_U(0) = 1 - \theta \).

- A sparse binary source \( \sim \) Bernoulli \( (\theta) \), \( 0 < \theta < \frac{1}{2} \).

- The entropy of the source is defined by

\[
H(U) \triangleq h(\theta) = -\theta \log \theta - (1 - \theta) \log(1 - \theta).
\]
Source Coding Theorem (Lossless Compression):

Let code rate $R > H(U)$. Then there exist fixed-length codes $(\phi_n, \psi_n)$ such that $R_n \leq R$ but $\text{BER} \to 0$. In the case when variable-length codes are allowed, we can make $\text{BER} = 0$.

Proof:
- This can be proved by at least three methods:
  - Typical Set
  - Method of Types
  - Random Binning

Problem Statement

Lossless Compression Algorithms

- Huffman coding
- Arithmetic coding
- LZ77 / LZ78 / LZW
- ...

Fixed-to-variable or
Variable-to-fixed or
Variable-to-variable length codes

Limitations

- Requirement: sufficiently long sequences
- Efficiency: delay and complexity
- Quality: the inherent error propagation
Fixed-to-fixed length Compression Scheme

- Let $G$ be a binary matrix of size $k \times n$, has the form

$$G = \begin{pmatrix}
G_{0,0} & G_{0,1} & \cdots & G_{0,n-1} \\
G_{1,0} & G_{1,1} & \cdots & G_{1,n-1} \\
\vdots & \vdots & \ddots & \vdots \\
G_{k-1,0} & G_{k-1,1} & \cdots & G_{k-1,n-1}
\end{pmatrix}$$

- The length of the sequence $u$ to be compressed is $k$
- The length of the compressed sequence $v$ is $n$
- The compression rate is defined as the code rate $R = n/k$.

- A linear block code of rate $R \triangleq \frac{n}{k}$:

  - **Encoder** $\varphi$ : $\mathbb{F}_2^k \rightarrow \mathbb{F}_2^n$, $v = uG$
  - **Decoder** $\psi$ : $\mathbb{F}_2^n \rightarrow \mathbb{F}_2^k$, find $\hat{u}$ such that $\hat{u}G = v$ and $P(\hat{u})$ is maximized.
Fixed-to-fixed length Compression Scheme

*Given that the elements of $G$ are independently and uniformly generated, the average decoding error probability

$$\Pr\{\psi(UG) \neq U\} \leq \varepsilon$$

where $\varepsilon$ is arbitrarily small, as long as $R > h(\theta)$ and $k \to \infty$.

Such a code ensemble is said to be universal.


Linear Block Codes

Definition:

A linear block code ensemble is called a **sparse random code ensemble** if the generator matrix has the form

\[ G = \begin{pmatrix}
G_{0,0} & G_{0,1} & \cdots & G_{0,n-1} \\
G_{1,0} & G_{1,1} & \cdots & G_{1,n-1} \\
\vdots & \vdots & \ddots & \vdots \\
G_{k-1,0} & G_{k-1,1} & \cdots & G_{k-1,n-1}
\end{pmatrix} \]

and \( G_{ij} (0 \leq i \leq k - 1, 0 \leq j \leq n - 1) \) is generated independently according to the Bernoulli distribution with success probability \( \Pr\{G_{ij} = 1\} = \rho < 1/2 \).
Lemma:

- Over the sparse code ensemble defined by $\rho < 1/2$, the codeword $v = uG$ with $W_H(u) = w$ is a Bernoulli sequence with success probability

$$\rho_w \triangleq \Pr\{V_j = 1 \mid W_H(U) = w\} = \frac{1-(1-2\rho)^w}{2}$$

Then we have $\rho_w \to \frac{1}{2}$ as $w \to \infty$.

- Furthermore, for any given positive integer $T \leq k$, $P_G(v_0^{n-1} \mid u) \triangleq \Pr\{V_0^{n-1} = v_0^{n-1} \mid U = u\} \leq P(0^n \mid u) \leq (1 - \rho_T)^n$,

for all $u \in \mathbb{F}_2^k$ with $W_H(u) \geq T$ and $v_0^{n-1} \in \mathbb{F}_2^n$. 
➢ Theorem:

- For any given positive \( \rho < 1/2 \), the code ensemble is \textit{universal} in terms of bit-error rate (BER) for sparse sources. That is, for any source with \( h(\theta) < R \), BER \( \rightarrow 0 \) as \( k \rightarrow \infty \).

◼ Proof:

- It can be proved by the method of typical set and the maximum-likelihood decoding algorithm.

\[
\text{BER}(u) = \frac{E[W_H(\hat{U} - u)]}{k} \\
= \sum \Pr\{\hat{u} \text{ is the most likely, } \hat{u}G = uG\} \frac{W_H(\hat{u} - u)}{k} \\
\leq \frac{T}{k} + \sum \Pr\{P(\hat{u}) \geq P(u), \hat{u}G = uG\} \\
\leq \frac{T}{k} + 2^{k(H + \epsilon)} (1 - \rho T)^n \\
= \frac{T}{k} + 2^{-k(R \log \frac{1}{1-\rho_T} - H - \epsilon)}.
\]

\[\text{converge} \quad k \rightarrow \infty \quad 0\]
Good channel codes can be leveraged for data compression.

- LDPC codes
- Turbo codes
- LDGM codes
- Spatial coupling
- Belief propagation
- Decimation
- Latency
- Complexity
- Tailored optimization algorithms
- Rate-distortion Limit
Related Research

Systematic Convolutional LDGM code

Universal & Flexible

A New Scheme for Near-lossless Data Compression Using Convolutional LDGM Codes

➢ The main advantage is that no complex optimization is required to construct good codes.


Convolutional LDGM Codes

Encoding

- **$L$ blocks** of data for compression, $u^{(0)}, u^{(1)}, \ldots, u^{(L-1)}$
- **Encoding memory** $m \geq 0$
- **Generator matrix** $G_i (0 \leq i \leq m)$: $m + 1$ matrices of size $k \times n$, with each column generated randomly and independently from all unit vectors.
- **Total code rate**
  \[
  R_L = \frac{n(L + m)}{(kL)} = \frac{n}{k} \cdot \frac{L + m}{L} \quad \text{for} \quad L \to \infty \quad \Rightarrow \quad \frac{n}{k}.
  \]

Figure: The framework of the proposed convolutional LDGM codes.


**Decoding**

- Iterative sliding window decoding algorithm
- Decompress the **receiving** $v^{(t)}$ to estimate the **original data** $u^{(t)}$

**Estimation of the Source Parameter**

- Estimate the parameter of the source by $\hat{\theta} = W_H (v^{(0)}) \big/ n$
- Initialize the assignment of $P (u^{(0)})$ in iterative decoding.

**Sliding Window Decoding**

- Initialization
- Iteration
  - ✓ Forward recursion
  - ✓ Backward recursion
  - ✓ Decision
- Cancelation
Decoding

- Normal graph: 3 types of constraint nodes
  - Node \( \equiv \): all the connecting variables take the same value;
  - Node \( + \): all the connecting variables sum to zero over \( \mathbb{F}_2 \);
  - Node \( G_i \): the \( i \)-th generator matrix.

Complexity Analysis

- Dominated by the operation at the node \( + \).
- In each iteration, the total decoding complexity is given by \( O(nm) \).

Figure: The normal graph of the proposed convolutional LDGM codes with memory \( m=2 \) and delay \( d=4 \).
Numerical Results

Universality of the proposed scheme

- The gap w.r.t. theoretical limit: $\leq 0.098$
- Increase with the parameter $\theta$
- In our simulations, $10^5$ blocks ($>10^8$ source digits) are simulated and no errors are found in most cases except that $\theta = 0.05$ (BER $\approx 10^{-5}$).

Figure: The compression rates with BER performance lower than $10^{-5}$. 

- Encoding memory $m = 16$
- Decoding delay $d = 32$
- Source block length $k = 1024$
Numerical Results

Decoding with / without the Knowledge about the Source Distribution

➢ The source distribution $\theta$: Given the fixed coding rate, BER decreases quickly with the source tending more sparse.

➢ Encoding memory $m$: BER can be lowered down by increasing $m$.

Figure: BER performance with and without $\theta$ ( $R=0.5$, simulated blocks=$10^6$ ).
Numerical Results

Table: Comparison of BER performance with and without $\theta$ (R=0.5, simulated blocks=$10^6$).

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$\theta$ Known / $\theta$ Unknown</th>
<th>BER performance with different memories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>m = 10</td>
</tr>
<tr>
<td>0.08</td>
<td>Known</td>
<td>3.9E-4</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>3.8E-4</td>
</tr>
<tr>
<td>0.09</td>
<td>Known</td>
<td>4.3E-4</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>4.2E-4</td>
</tr>
<tr>
<td>0.10</td>
<td>Known</td>
<td>4.7E-2</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>4.8E-2</td>
</tr>
<tr>
<td>0.11</td>
<td>Known</td>
<td>5.6E-2</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>5.6E-2</td>
</tr>
</tbody>
</table>

➢ No significant degradation in BER performance even if $\theta$ is unknown.
Conclusions and Future Work

A New Near-lossless Compression Scheme for Binary Sparse Source

✓ A fixed-to-fixed length encoding scheme
✓ Theoretical proof
✓ Practical scheme
✓ Estimate the source parameter

➢ Universal scheme for multiple sources
➢ Joint source-channel coding (JSCC)
Thank you for your attention!