中山大學
SUN YAT-SEN UNIVERSITY

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

$$\boldsymbol{U} = \boldsymbol{U_0}, \boldsymbol{U_1}, \ldots,$$ where $U_t \in \mathbb{F}_2 \triangleq \{\mathbf{0}, \mathbf{1}\}$ 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 ~ Bernoulli (θ), $0 < \theta < ½$.
- The entropy of the source is defined by

$$H(U) \triangleq h(\theta) = -\theta \log \theta - (1-\theta) \log(1-\theta).$$

Problem Statement

➢ Source Coding Theorem (Lossless Compression):

- Let code rate $R > H(U)$. Then there exist **fixed-length codes** (ϕ_n, ψ_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

#As proved in: T. M. Cover and J. A. Thomas, *Elements of Information Theory(Second edition),* John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

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

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

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

- A linear block code of rate $\boldsymbol{R} \triangleq \frac{\boldsymbol{n}}{\boldsymbol{k}}$:
  - **Encoder φ :** $\mathbb{F}_2^k \to \mathbb{F}_2^n$, $\boldsymbol{v} = \boldsymbol{u}\mathbf{G}$
  - **Decoder ψ :** $\mathbb{F}_2^n \to \mathbb{F}_2^k$, find $\hat{u}$ such that $\hat{\boldsymbol{u}}G = \boldsymbol{v}$ and $\mathrm{P}(\hat{u})$ is maximized.

Fixed-to-fixed length Compression Scheme

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

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

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

- #Such a code ensemble is said to be ***universal***.

*As proved in: I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems (Second edition),* Cambridge University Press, New York, 2011.

#As defined in: T. M. Cover and J. A. Thomas, *Elements of Information Theory(Second edition),* John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

Linear Block Codes

Definition:

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

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

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

➢ Lemma:

- Over the sparse code ensemble defined by $\rho < 1/2$, the codeword $v = u\mathbf{G}$ 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 $T \leq k$,

$$P_G\left(v_0^{n-1} \middle| u\right) \triangleq \Pr\{V_0^{n-1} = v_0^{n-1} | \, U = u\} \leq P(0^n|u) \leq (1-\rho_T)^n,$$

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 ***universal*** in terms of bit-error rate (BER) for sparse sources. That is, for any source with $h(\theta) < R$, $\text{BER} \to 0$ as $k \to \infty$.

Proof:

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

$$
\begin{aligned}
\text{BER}(\boldsymbol{u}) &= \frac{\mathbf{E}[W_H(\hat{\boldsymbol{U}} - \boldsymbol{u})]}{k} \\
&= \sum_{\hat{\boldsymbol{u}}} \Pr\{\hat{\boldsymbol{u}} \text{ is the most likely}, \hat{\boldsymbol{u}}\mathbf{G} = \boldsymbol{u}\mathbf{G}\} \frac{W_H(\hat{\boldsymbol{u}} - \boldsymbol{u})}{k} \\
&\leqslant \frac{T}{k} + \sum_{\hat{\boldsymbol{u}}: W_H(\hat{\boldsymbol{u}} - \boldsymbol{u}) \geqslant T} \Pr\{P(\hat{\boldsymbol{u}}) \geqslant P(\boldsymbol{u}), \hat{\boldsymbol{u}}\mathbf{G} = \boldsymbol{u}\mathbf{G}\} \\
&\leqslant \frac{T}{k} + 2^{k(H+\epsilon)}(1 - \rho_T)^n \\
&= \frac{T}{k} + 2^{-k(R \log \frac{1}{1-\rho_T} - H - \epsilon)}. \xrightarrow[k \to \infty]{\text{converge}} \mathbf{0}
\end{aligned}
$$

Related Research

➢ **Good channel codes can be leveraged for data compression.**

- LDPC codes
- Turbo codes
- LDGM codes
- …

Spatial coupling

Belief propagation

Decimation

→ Rate-distortion Limit

Latency

Complexity

Tailored optimization algorithms

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.**

*X. Ma, "Coding theorem for systematic low density generator matrix codes," in *Proc. IEEE 9th Int. Symp. on Turbo Codes and Iterative Inf. Processing(ISTC),* 2016, pp. 11–15.

#S. Cai, W. Lin, X. Yao, B. Wei, and X. Ma, "Systematic convolutional low density generator matrix code," [Online], 2019, Available: https://arxiv.org/abs/2001.02854.

Convolutional LDGM Codes

Encoding

- L **blocks of data for compression,** $\boldsymbol{u}^{(0)}, \boldsymbol{u}^{(1)}, \cdots, \boldsymbol{u}^{(L-1)}$
- **Encoding memory** $m \geq 0$
- **Generator matrix** $\boldsymbol{G}_i \ (0 \leq i \leq m)$ **:** $m+1$ **matrices of size** $k \times n$**, with each column generated randomly and independently from all unit vectors.**
- **Total code rate**

$$R_L = n(L+m)/(kL) = n/k \cdot (L+m)/L \xrightarrow{L \to \infty} n/k\,.$$

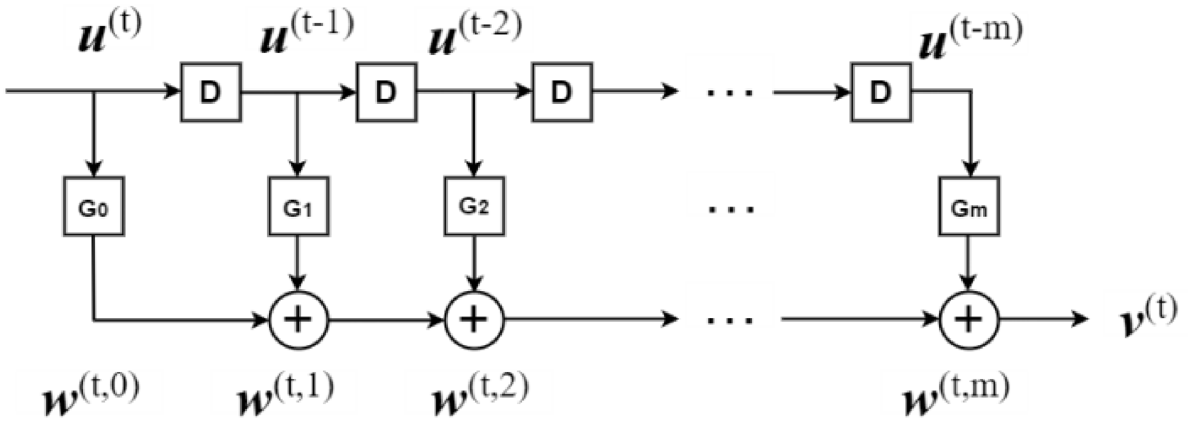

Figure: The framework of the proposed convolutional LDGM codes.

Decoding

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

Estimation of the Source Parameter → Sliding Window Decoding

- Estimate the parameter of the source by
  $$\widehat{\boldsymbol{\theta}} = \boldsymbol{W}_H\left(\boldsymbol{v}^{(0)}\right)/\boldsymbol{n}$$
- Initialize the assignment of $P\left(\boldsymbol{u}^{(0)}\right)$ in iterative decoding.

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

Decoding

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

Complexity Analysis

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

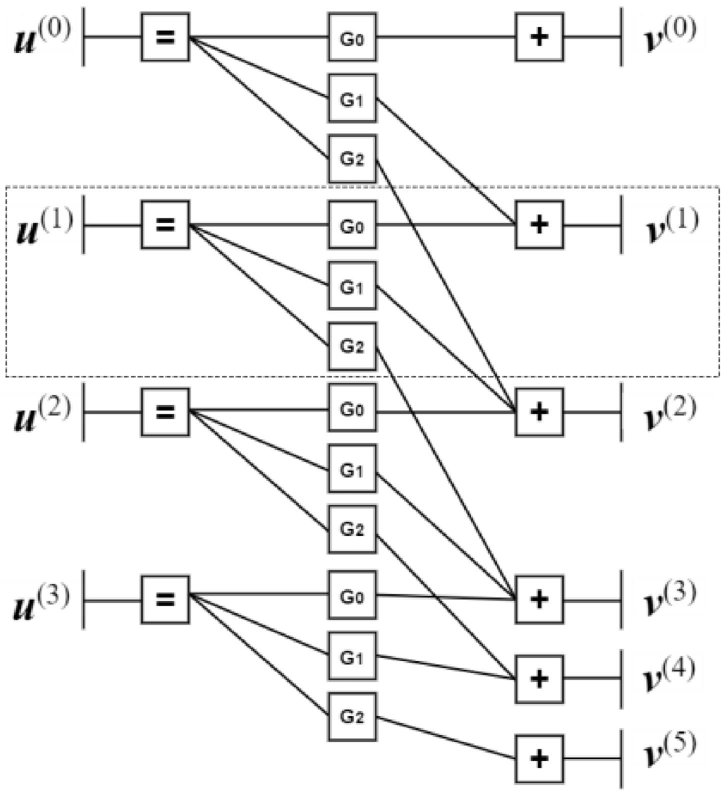

Figure: The normal graph of the proposed convolutional LDGM codes with memory m=2 and delay d=4.

Numerical Results

Universality of the proposed scheme

- The gap w.r.t. theoretical limit: ≤ 0.098
- Increase with the parameter θ
- 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 $\approx 10^{-5}$).

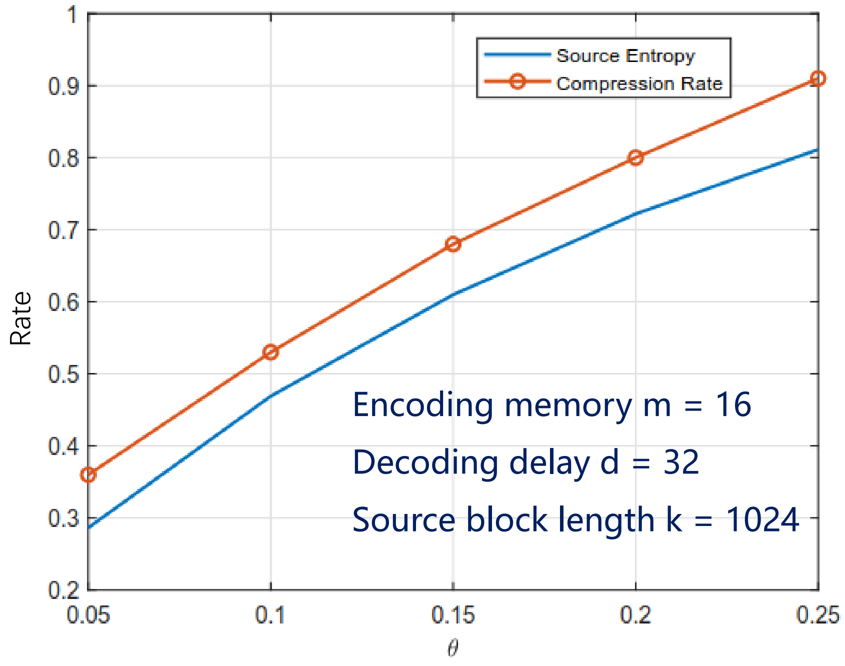

Figure: The compression rates with BER performance lower than 10^{-5}.

Numerical Results

Decoding with / without the Knowledge about the Source Distribution

- **The source distribution θ:**

  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.

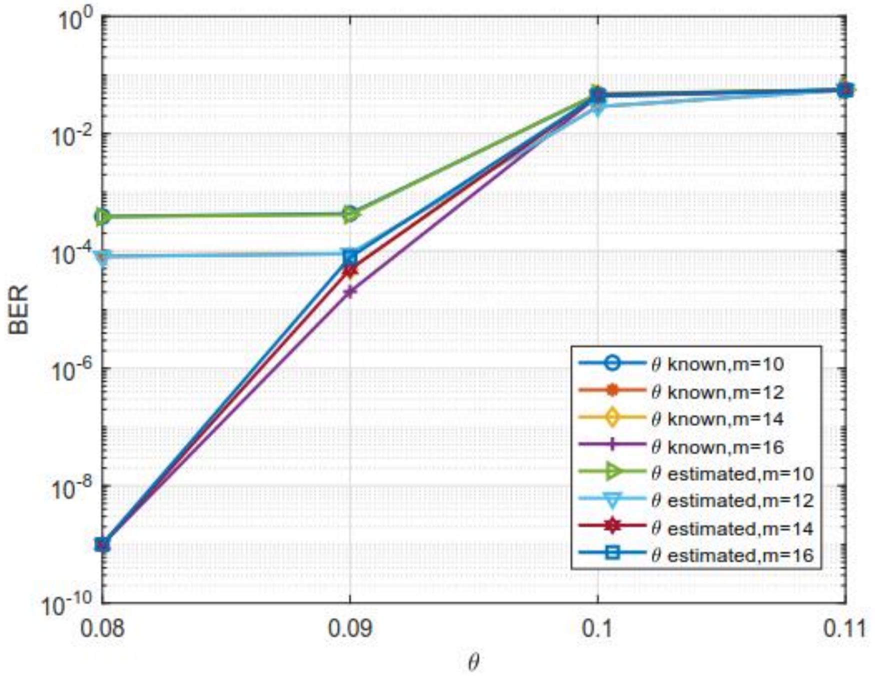

Figure: BER performance with and without θ (R=0.5, simulated blocks=10^6).

Numerical Results

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

θ	θ Known / θ Unknown	BER performance with different memories			
		m = 10	m = 12	m = 14	m = 16
0.08	Known	3.9E-4	8.1E-5	0	1.0E-9
	Unknown	3.8E-4	8.0E-5	1.0E-9	1.0E-9
0.09	Known	4.3E-4	9.0E-5	4.9E-5	2.0E-5
	Unknown	4.2E-4	9.0E-5	4.9E-5	7.8E-5
0.10	Known	4.7E-2	2.9E-2	4.6E-2	4.3E-2
	Unknown	4.8E-2	2.9E-2	4.6E-2	4.5E-2
0.11	Known	5.6E-2	5.5E-2	5.5E-2	5.5E-2
	Unknown	5.6E-2	5.6E-2	5.5E-2	5.5E-2

➢ **No significant degradation in BER performance even if θ 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)

中山大學
SUN YAT-SEN UNIVERSITY

Thank you for your attention!