SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Error Exponent for Concatenated Codes in DNA Data Storage under Substitution Errors

Yuxuan SHI[1], Shuo SHAO[1], Xiaohang ZHANG[2*], Yongjian WANG[2] & Yongpeng WU[3]

[1] *School of Cyber and Engineering, Shanghai Jiao Tong University, Shanghai* 200240*, China ;*
[2] *The National Computer Network Emergency Response Technical Team. Coordination Center of China, Beijing* 100029*, China;*
[3]*Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai* 200240*, China*

Appendix A Symbols Explanation

Symbols	Definition
$\boldsymbol{X}_i$	Input sequence
$\mathcal{X}^M$	Set of input sequences
$\boldsymbol{Y}_j$	Output sequence
$\mathcal{Y}^N$	Set of output sequences
I_j	Cluster label
Z_i	Cluster which contains the sequences from the input X_i
E_j	Error sequence
M	Cardinality of the input sequences/Outer code length
L	Length of the input sequences/Inner code length
N	Cardinality of the output sequences
c	Average times that each sequence has been drawn $c = \frac{N}{M}$
p	Crossover probability of BSC
$\hat{X}_i$	Output sequence after bit-wised majority decision
D_i	number of times the i-th input sequence has been drawn (realization d)
U_d	number of input sequences has been drawn same d times (realization u)
β	rate loss due toindex coding
C_d	Capacity of the sub-channel has been drawn d times
C	Capacity of the overall channel
R	Outer code rate
r	Inner code rate
$\mathcal{R}$	Overall rate
K	Dimension of outer code
k	Dimension of inner code
d_{out}	Minimum distance of outer code
d_{in}	Minimum distance of inner code
θ	Pre-decoding parameters
$E_G(p, r)$	Gallager error Exponent
η_i	Indicator of the sub-block
δ	The relative minimum distance ($\delta = \frac{d_{\text{out}}}{M}$)
$\mathcal{J}$	Set of the subscript after pre-decoding

Appendix B Important Definitions and Lemmas

We bound the error probability when both erasure and error probabilities converge to 0, in which the overall rate is shown as a function of inner code rate and coverage depth. We classify the decoding failures into erasures and errors using the specific decoder, where the error exponent is restricted by the Chernoff bound on independent indicators. The necessary definitions and lemmas are presented as follows.

* Corresponding author (email: zhangxiaohang@cert.org.cn)

i	$\boldsymbol{\eta}^M$	$\boldsymbol{d}^M$
1	η_1	1
2	η_2	3
3	η_3	3
4	η_4	4
5	η_5	3
6	η_6	4
7	η_7	2
8	η_8	1
9	η_9	0

$\Rightarrow$

d	$\sum \eta_i$	$\boldsymbol{u}^N$
0	η_9	1
1	$\eta_1+\eta_8$	2
2	η_7	1
3	$\eta_2+\eta_3+\eta_5$	3
4	$\eta_4+\eta_6$	2
5	0	0
6	0	0
$\vdots$	$\vdots$	$\vdots$
21	0	0

Table B1 Regroup of the binary indicators in terms of d for $M=9$ and $N=21$

Definition 1. Given the input strands $\{\boldsymbol{X}_i\}_1^M$ and the output $\{\boldsymbol{Y}_j\}_1^N$, random variable D_i refers to the numbers of the i-th strands has been drawn, with realization d. Moreover the number of input strands which have been drawn exactly d times, is defined as random variable U_d with realization u

$$D_i = |\{j\colon I_j = i, i \in [1,M]\}| \qquad \boldsymbol{D}^M = \{D_1, D_2, ..., D_M\}$$
$$U_d = |\{i\colon D_i = d, d \in [0,N]\}| \qquad \boldsymbol{U}^N = \{U_0, U_1, ..., U_N\}$$

Apparently $\sum_{i=1}^M D_i = N$ and $\sum_{d=0}^N U_d = M$.

Definition 2. Given $\hat{\boldsymbol{X}}_i$ and $\boldsymbol{X}_i$. Define such a indicator η_i implies whether the i-th sub-block is decoded successfully or not, i.e.

$$\eta_i = \begin{cases} 1 & \mathbb{D}_{\text{in}}(\hat{\boldsymbol{X}}_i) \neq \boldsymbol{X}_i \\ 0 & \mathbb{D}_{\text{in}}(\hat{\boldsymbol{X}}_i) = \boldsymbol{X}_i \end{cases}$$

where $\mathbb{D}_{\text{in}}$ refers to the inner decoder.

Lemma 1. With notation as above, fix $0 < \delta < 1$ and $\mathcal{J} = \{i : |\boldsymbol{Z}_i| \geqslant \theta\}$. Then the error probability of outer code could be bounded by

$$P_w = \sum_{\boldsymbol{u}^N} \mathbb{P}\left\{ \sum_{i\in\mathcal{J}} \eta_i \geqslant M\delta \,\middle|\, \boldsymbol{D}^M = \boldsymbol{d}_{\boldsymbol{u}}^M \right\} \cdot \mathbb{P}\{\boldsymbol{U}^N = \boldsymbol{u}^N\}, \tag{B1}$$

where $\boldsymbol{d}_{\boldsymbol{u}}^M \in \boldsymbol{\mathcal{D}}^M(\boldsymbol{u}^N) = \left\{\boldsymbol{d}^M : \Pi(\boldsymbol{d}^M) = \boldsymbol{u}^N\right\}$ and $\Pi(\cdot)$ is the function which is used to map vector $\boldsymbol{d}^M$ to $\boldsymbol{u}^N$. The function $\Pi(\cdot)$ and the regroup of the indicators $\boldsymbol{\eta}^M$ is illustrated in Table.B1

Proof. We abbreviate $\mathbb{P}\{\boldsymbol{D}^M = \boldsymbol{d}^M\} = \mathbb{P}\{\boldsymbol{d}^M\}$, the same as $\boldsymbol{u}^N$. The error probability can be reformulated as follows:

$$\begin{aligned}
P_w &= \sum_{\boldsymbol{d}^M} \mathbb{P}\left\{ \sum_{i\in\mathcal{J}} \eta_i \geqslant M\delta \,\middle|\, \boldsymbol{d}^M \right\} \mathbb{P}\left\{\boldsymbol{d}^M\right\} \\
&= \sum_{\boldsymbol{u}^N} \sum_{\boldsymbol{d}^M \in \boldsymbol{\mathcal{D}}^M} \underbrace{\mathbb{P}\left\{ \sum_{i\in\mathcal{J}} \eta_i \geqslant M\delta \,\middle|\, \boldsymbol{d}_{\boldsymbol{u}}^M \right\}}_{\mathcal{F}(\boldsymbol{u}^N)} \mathbb{P}\left\{\boldsymbol{d}^M\right\} \\
&\overset{(a)}{=} \sum_{\boldsymbol{u}^N} \mathcal{F}(\boldsymbol{u}^N) \sum_{\boldsymbol{d}^M \in \boldsymbol{\mathcal{D}}^M} \mathbb{P}\left\{\boldsymbol{d}^M\right\} \\
&= \sum_{\boldsymbol{u}^N} \mathcal{F}(\boldsymbol{u}^N) \mathbb{P}\{\boldsymbol{u}^N\}
\end{aligned}$$

(a) holds since $\mathbb{P}\left\{ \sum_{i\in\mathcal{J}} \eta_i \geqslant M\delta \middle| \boldsymbol{d}_{\boldsymbol{u}}^M \right\} = \mathcal{F}(\boldsymbol{u}^N)$ is a function of $\boldsymbol{u}^N$ rather than $\boldsymbol{d}^M$.

Remark 1. The specific set $\boldsymbol{\mathcal{D}}^M(\boldsymbol{u}^N)$, contains all $\boldsymbol{d}^M$ vectors which could be mapped to $\boldsymbol{u}^N$ vector. Given the condition that $\boldsymbol{D}^M$ is in this set, the error probability could be represented by a function depending on $\boldsymbol{u}^N$ rather than $\boldsymbol{d}^M$. This representation is attributed to that the conditional probability relies on the sum of binary indicators but not care about the permutation. It avoids the cumbersome distribution without losing the independence on these indicators.

Lemma 2 (Forney's Exponent [1])**.** There exists a concatenated code $\mathcal{C}_{\text{cont}}$ whose outer code is a MDS code, with overall length M_0 and overall rate $\mathcal{R}$. Consider parallel BSC sub-channels with the same flip rate p, its probability of decoding error $\mathrm{P}_{\text{err}}\left(\mathcal{C}_{\text{cont}}\right)$ is given by

$$\mathrm{P}_{\text{err}}\left(\mathcal{C}_{\text{cont}}\right) \leqslant \exp\left\{-M_0 E_C\left(p, \mathcal{R}\right)\right\} \quad \left(0 \leqslant \mathcal{R} < 1 - H_q(p)\right) \tag{B2}$$

where

$$E_F\left(p, \mathcal{R}\right) = \max_{0<r<1-H_q(p)} \left(1 - \frac{\mathcal{R}}{r}\right) E_G(p, r) \tag{B3}$$

which is called the concatenation exponent or Forney's exponent. $E_G(p, r)$ is the Gallager's random coding error exponent. The overall decoding complexity for the code $\mathcal{C}$ is given by at most

$$O\left(M_0^2 \log^2 M_0\right)$$

due to the fact that the outer code (which is always a Reed-Solomon code) used in [**?**] has a high decoding complexity.

Lemma 3 (Poissonization [2])**.** For any fixed $c > 0$ and the poisson distribution $p_c(d) = e^{-c}c^d/d!$ there exists a set of functions $f_d(M), d = 0, \ldots, N$ with $f_0(M) + f_1(M) + \cdots + f_N(M) = o(M)$ such that for $M \to \infty$

$$\mathbb{P}\left(|U_d - M p_c(d)| \leqslant f_d(M), d = 0, \ldots, N\right) \to 1 \tag{B4}$$

Remark 2. This lemma implies the deviation between a random variable U_d and the expectation $Mp_c(d)$ goes to $o(M)$ as $M \to \infty$. It means in this case $U_d \approx Mp_c(d)$ with large enough M. This effect is known as Poissonization although D_i are statistically dependent and each variable U_d could be viewed as no longer random but almost deterministic. Moreover, the bounding functions $f_d(M)$ are used to describe the "almost".

Appendix C Proof of Theorem 1

Proof. We start the proof by counting the number of erasures and errors during the transmission. Erasures come from the reject to the clusters of small sizes $C_d < r$, since they are not able to transfer reliable information from the perspective of decoder. Consider the unique decoding, erasures consume the same amount of parity check of outer code. Moreover, errors occur when the left redundancy cannot afford the error correction. These two random variables can be given as:

$$v = \sum_{d<\theta} U_d; \qquad w = \sum_{i \in \mathcal{J}} \eta_i$$

The error probability of the proposed concatenated codes $\mathrm{P}_{\text{err}}\left(\mathcal{C}_{\text{cont}}\right)$ can be bounded as

$$\begin{aligned}
\mathrm{P}_{\text{err}}\left(\mathcal{C}_{\text{cont}}\right) &= \mathbb{P}\left\{\boldsymbol{h} \neq \hat{\boldsymbol{h}}\right\} \\
&= \mathbb{P}\underbrace{\left\{v + 2w > M(1-R)\right\}}_{\mathcal{A}} \\
&= \sum_{\boldsymbol{d}^M} \mathbb{P}\left\{v + 2w > M(1-R) | \, \boldsymbol{d}^M\right\} \mathbb{P}\left\{\boldsymbol{d}^M\right\} \\
&\overset{(b)}{=} \sum_{\boldsymbol{u}^N \in \mathcal{T}} \mathbb{P}\left\{ v + 2w > M(1-R) | \, \boldsymbol{d}_{\boldsymbol{u}}^M\right\} \mathbb{P}\left\{\boldsymbol{u}^N\right\} + \sum_{\boldsymbol{u}^N \notin \mathcal{T}} \mathbb{P}\left\{ v + 2w > M(1-R) | \, \boldsymbol{d}_{\boldsymbol{u}}^M\right\} \mathbb{P}\left\{\boldsymbol{u}^N\right\} \\
&\overset{(c)}{\leqslant} \mathcal{P}(1-\epsilon) + \epsilon
\end{aligned}$$

where in (b) we use Lemma 1 and split $\mathrm{P}_{\text{err}}\left(\mathcal{C}_{\text{cont}}\right)$ into two parts, namely conditional probabilities given asymptotic and non-asymptotic distribution, where we define the asymptotic set $\mathcal{T} = \left\{\boldsymbol{u}^N : |U_d - Mp_c(d)| \leqslant f_d(M), d = 1, 2, ..., N\right\}$ and $f_d(M)$ is the deviation function dependent on M. From Lemma 3, given $M \to \infty$, we conclude the probability $\mathbb{P}\{\boldsymbol{u}^N \notin \mathcal{T}\} \to 0$, i.e. the marginals D_i approach Poisson distributions and the random variables $\frac{U_d}{M} \to p_c(d)$ can be viewed as asymptotically deterministic, in spite of the statistically dependence among D_i. In (c) we bound the asymptotic error probability by $\mathcal{P}$ the non-asymptotic one by 1.

The conditional probability with $\boldsymbol{u}^N \in \mathcal{T}$ can be stated as:

$$\begin{aligned}
&\mathbb{P}\left\{v + 2w > M(1-R) | \boldsymbol{d}_u^M\right\} \\
=&\mathbb{P}\left\{2w > M(1 - R - \sum_{d<\theta} p_c(d)) | \boldsymbol{d}_u^M\right\}
\end{aligned}$$

Next given the $\boldsymbol{d}_u^M$ sampling vector, $\{\eta_i\}_{i \in \mathcal{J}}$ are independent binomial random variables with different P_i respectively, where $P_i = \mathbb{P}\{d_{\mathrm{H}}(\boldsymbol{X}_i, \hat{\boldsymbol{X}}_i) \geqslant \frac{d_{\text{in}}}{2}\}$. Define $\bar{P} = \max_{1 \leqslant i \leqslant M} P_i$ we denote the maximum error probability of inner decoder. Then we bound this conditional error probability with Lemma 2

$$\mathbb{P}\left\{2w > M(1 - R - \sum_{d<\theta} p_c(d))\right\} \leqslant \sum_{i=M\tau}^{M} \binom{M}{i} \bar{P}^i (1-\bar{P})^{M-i}$$

$$\begin{aligned}
&\leqslant \sum_{i=\tau}^{M} \binom{M}{i} \bar{P}^i \leqslant \bar{P}^\tau \sum_{i=\tau}^{M} \binom{M}{i} \\
&\leqslant 2^M \cdot \bar{P}^{M\frac{\tau}{M}} \\
&\overset{(d)}{\leqslant} q^{-LM\left(E_g(p^\star,r)\frac{\tau}{M}-\epsilon\right)}
\end{aligned}$$

where

$$\tau = \frac{M}{2}(1 - R - \sum_{d<\theta} p_c(d))$$

$$p^\star = \sum_{i>\theta/2}^{\theta} \binom{\theta}{i} p^i (1-p)^{(\theta-i)}$$

In (d) we bound $\bar{P}^M$ with the Shannon coding theorem for q-ary symmetric channel, and $E_g(p^\star)$ is the Gallager error exponent [3]. The claim of Theorem 1 follows with $M \to \infty$, $\mathcal{P} = q^{-LM\left(E_q(p^\star,r)\frac{\tau}{M}-o(1)\right)}$, and the artificial small parameter ϵ.

References

1 S. Hirasawa and M. Kasahara, "Exponential error bounds and decoding complexity for block concatenated codes with tail biting trellis inner codes," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 9, no. 2, pp. 307–320, 2006.

2 A. Lenz, P. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the DNA storage channel," in *IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019.

3 R. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inf. Theory*, vol. 11, no. 1, pp. 3–18, 1965.