# The DNA Storage Channel: Capacity and Error Probability Bounds

Nir Weinberger and Neri Merhav

Technion - Israel Institute of Technology, Israel

# Outline

# DNA storage

- Why store information in DNA strands?

# DNA storage

- Why store information in DNA strands?
  - Enormous information density: 5 grams can store $8 \cdot 10^{21}$ bits

# DNA storage

- Why store information in DNA strands?
  - Enormous information density: 5 grams can store $8 \cdot 10^{21}$ bits
  - Extreme longevity: Messages from mammoths...

# DNA storage

- Why store information in DNA strands?
  - Enormous information density: 5 grams can store $8 \cdot 10^{21}$ bits
  - Extreme longevity: Messages from mammoths...
- Working prototypes starting from 2012, world record $\sim$ 200MB [Organick et al 2018]

# DNA storage

- Why store information in DNA strands?
    - Enormous information density: 5 grams can store $8 \cdot 10^{21}$ bits
    - Extreme longevity: Messages from mammoths...
- Working prototypes starting from 2012, world record $\sim 200$MB [Organick et al 2018]
- Still costly: $\sim \$500$ per 1MB of data

# DNA storage

- Why store information in DNA strands?
  - Enormous information density: 5 grams can store $8 \cdot 10^{21}$ bits
  - Extreme longevity: Messages from mammoths...
- Working prototypes starting from 2012, world record $\sim$ 200MB [Organick et al 2018]
- Still costly: $\sim$ \$500 per 1MB of data
- Check out: *"Information-Theoretic Foundations of DNA Data Storage"* [Shomorony and Heckel, FnT, 2022]

# The DNA storage channel model – writing/encoder

- Alphabet $\mathcal{X}$, in real-life $\mathcal{X} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$

# The DNA storage channel model – writing/encoder

- Alphabet $\mathcal{X}$, in real-life $\mathcal{X} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$
- A DNA molecule is a ***sequence*** $x^L \in \mathcal{X}^L$ (order matters)

# The DNA storage channel model – writing/encoder

- Alphabet $\mathcal{X}$, in real-life $\mathcal{X} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$
- A DNA molecule is a ***sequence*** $x^L \in \mathcal{X}^L$ (order matters)
- A codeword is a ***multiset*** of $M$ molecules (no order)

$$x^{LM} = (x_0^L, \dots x_{M-1}^L)$$

# The DNA storage channel model – writing/encoder

- Alphabet $\mathcal{X}$, in real-life $\mathcal{X} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$
- A DNA molecule is a ***sequence*** $x^L \in \mathcal{X}^L$ (order matters)
- A codeword is a ***multiset*** of $M$ molecules (no order)

$$x^{LM} = (x_0^L, \ldots x_{M-1}^L)$$

- A codebook is a set of different codewords $\mathcal{C} = \{x^{LM}(j)\}$

# The DNA storage channel model – reading

- Channel output is a ***multiset*** of $N$ molecules (order does not matter)

$$Y^{LN} = (Y_0^L, \ldots, Y_{N-1}^L)$$

# The DNA storage channel model – reading

- Channel output is a ***multiset*** of $N$ molecules (order does not matter)

$$Y^{LN} = (Y_0^L, \ldots, Y_{N-1}^L)$$

- Output molecule $Y_n^L$ is generated as:

# The DNA storage channel model – reading

- Channel output is a ***multiset*** of $N$ molecules (order does not matter)

$$Y^{LN} = (Y_0^L, \ldots, Y_{N-1}^L)$$

- Output molecule $Y_n^L$ is generated as:
  1. Sample one of the $M$ molecules of $x^{LM}$, independently, with replacement

# The DNA storage channel model – reading

- Channel output is a ***multiset*** of $N$ molecules (order does not matter)

$$Y^{LN} = (Y_0^L, \ldots, Y_{N-1}^L)$$

- Output molecule $Y_n^L$ is generated as:
  1. Sample one of the $M$ molecules of $x^{LM}$, independently, with replacement
  2. Sequencing $x^L$ to obtain $Y_n^L$ – Modeled as a DMC

$$W\left(y_n^L \mid x^L\right) = \prod_{i \in [L]} W(y_i \mid x_i)$$

# The DNA storage channel model – decoding

- The decoder is a mapping $(\mathcal{Y}^L)^N \to [|\mathcal{C}|]$

# The DNA storage channel model – decoding

- The decoder is a mapping $(\mathcal{Y}^L)^N \to [|\mathcal{C}|]$
- Equivalently, a set of the decision regions $\mathcal{D} = \{\mathcal{D}(j)\}_{j\in[|\mathcal{C}|]}$

# The DNA storage channel model – decoding

- The decoder is a mapping $(\mathcal{Y}^L)^N \to [|\mathcal{C}|]$
- Equivalently, a set of the decision regions $\mathcal{D} = \{\mathcal{D}(j)\}_{j \in [|\mathcal{C}|]}$
    - $\mathcal{D}(j)$ is the decision region of the $j$th codeword

$$\mathcal{D}(j) := \{y^{LN} : \mathcal{D}(y^{LN}) = j\}$$

# The DNA storage channel model – channel

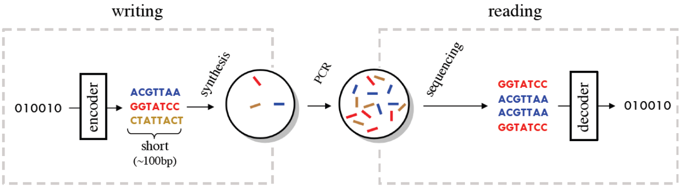


Figure: DNA storage model (Courtesy of Shomrony and Heckel)

# The DNA storage channel model – parameters

- $\mathsf{DNA} := (\alpha, \beta, W)$ is a sequence indexed by the number of molecules $M$

# The DNA storage channel model – parameters

- $\mathsf{DNA} := (\alpha, \beta, W)$ is a sequence indexed by the number of molecules $M$
- Coverage depth parameter $\alpha := \frac{N}{M}$

# The DNA storage channel model – parameters

- $\mathsf{DNA} := (\alpha, \beta, W)$ is a sequence indexed by the number of molecules $M$
- Coverage depth parameter $\alpha := \frac{N}{M}$
- Molecule length scaling: $\beta := \frac{L}{\log M} > 1$

# The DNA storage channel model – parameters

- $\mathsf{DNA} := (\alpha, \beta, W)$ is a sequence indexed by the number of molecules $M$
- Coverage depth parameter $\alpha := \frac{N}{M}$
- Molecule length scaling: $\beta := \frac{L}{\log M} > 1$
- DMC sequencing channel $W$

# The DNA storage channel model – parameters

- $\mathsf{DNA} := (\alpha, \beta, W)$ is a sequence indexed by the number of molecules $M$
- Coverage depth parameter $\alpha := \frac{N}{M}$
- Molecule length scaling: $\beta := \frac{L}{\log M} > 1$
- DMC sequencing channel $W$
- Coding rate

$$R = \frac{\log |\mathcal{C}|}{LM}$$

# The DNA storage channel model – parameters

- $\mathsf{DNA} := (\alpha, \beta, W)$ is a sequence indexed by the number of molecules $M$
- Coverage depth parameter $\alpha := \frac{N}{M}$
- Molecule length scaling: $\beta := \frac{L}{\log M} > 1$
- DMC sequencing channel $W$
- Coding rate

$$R = \frac{\log |\mathcal{C}|}{LM}$$

- Problem: What is the Shannon **capacity** of $\mathsf{DNA}$?

# Previous works

- Initial ideas: [MacKay, Sayer, and Goldman 2015], [Heckel, Shomorony, Ramchandran and Tse, 2017]

# Previous works

- Initial ideas: [MacKay, Sayer, and Goldman 2015], [Heckel, Shomorony, Ramchandran and Tse, 2017]
- Coding-theoretic papers: [Kovacevic and Tan 2018], [Lenz et al 2019], [Sima, Raviv and Bruck 2021], [Song, Cai, and Immink 2020] [Tang and Farnoud 2021]

# Previous works

- Initial ideas: [MacKay, Sayer, and Goldman 2015], [Heckel, Shomorony, Ramchandran and Tse, 2017]
- Coding-theoretic papers: [Kovacevic and Tan 2018], [Lenz et al 2019], [Sima, Raviv and Bruck 2021], [Song, Cai, and Immink 2020] [Tang and Farnoud 2021]
- The foundations of our work:

# Previous works

- Initial ideas: [MacKay, Sayer, and Goldman 2015], [Heckel, Shomorony, Ramchandran and Tse, 2017]
- Coding-theoretic papers: [Kovacevic and Tan 2018], [Lenz et al 2019], [Sima, Raviv and Bruck 2021], [Song, Cai, and Immink 2020] [Tang and Farnoud 2021]
- The foundations of our work:
  - The model, first capacity results, basic ideas [Shomorony and Heckel, 2021]

# Previous works

- Initial ideas: [MacKay, Sayer, and Goldman 2015], [Heckel, Shomorony, Ramchandran and Tse, 2017]
- Coding-theoretic papers: [Kovacevic and Tan 2018], [Lenz et al 2019], [Sima, Raviv and Bruck 2021], [Song, Cai, and Immink 2020] [Tang and Farnoud 2021]
- The foundations of our work:
  - The model, first capacity results, basic ideas [Shomorony and Heckel, 2021]
  - Refinement to a multinomial model [Lenz, Siegel, Wachter-Zeh, Yaakobi, 2019-2020]

# Previous works

- Initial ideas: [MacKay, Sayer, and Goldman 2015], [Heckel, Shomorony, Ramchandran and Tse, 2017]
- Coding-theoretic papers: [Kovacevic and Tan 2018], [Lenz et al 2019], [Sima, Raviv and Bruck 2021], [Song, Cai, and Immink 2020] [Tang and Farnoud 2021]
- The foundations of our work:
  - The model, first capacity results, basic ideas [Shomorony and Heckel, 2021]
  - Refinement to a multinomial model [Lenz, Siegel, Wachter-Zeh, Yaakobi, 2019-2020]
  - Both works only for $W = \mathsf{BSC}(w)$ (essentially)

# Outline

# Our strategy

- Error probability analysis of

# Our strategy

- Error probability analysis of
  1. Encoder: Standard random coding ensemble

# Our strategy

- Error probability analysis of
  1. Encoder: Standard random coding ensemble
  2. Decoder: High complexity, “optimal-like”

# Our strategy

- Error probability analysis of
  1. Encoder: Standard random coding ensemble
  2. Decoder: High complexity, "optimal-like"
- Result:

# Our strategy

- Error probability analysis of
  1. Encoder: Standard random coding ensemble
  2. Decoder: High complexity, "optimal-like"
- Result:
  1. A bound on the reliability function

# Our strategy

- Error probability analysis of
  1. Encoder: Standard random coding ensemble
  2. Decoder: High complexity, “optimal-like”
- Result:
  1. A bound on the reliability function
  2. Capacity bound is the vanishing point of the reliability function bound

# The binomial channel

- The $d$-order binomial extension of a DMC: $V : \mathcal{A} \to \mathcal{B}$ is the DMC
$$V^{\oplus d}[b^d \mid a] = \prod_{i=0}^{d-1} V(b_i \mid a)$$
for $a \in \mathcal{A}, b^d \in \mathcal{B}^d$

# The binomial channel

- The $d$-order binomial extension of a DMC: $V : \mathcal{A} \to \mathcal{B}$ is the DMC
$$V^{\oplus d}[b^d \mid a] = \prod_{i=0}^{d-1} V(b_i \mid a)$$
for $a \in \mathcal{A}, b^d \in \mathcal{B}^d$
- Interpretation: "$d$ independent observations on an input symbol $a \in \mathcal{A}$ over $V$"

# The binomial channel

- The $d$-order binomial extension of a DMC: $V : \mathcal{A} \to \mathcal{B}$ is the DMC
$$V^{\oplus d}[b^d \mid a] = \prod_{i=0}^{d-1} V(b_i \mid a)$$
for $a \in \mathcal{A}, b^d \in \mathcal{B}^d$
- Interpretation: "$d$ independent observations on an input symbol $a \in \mathcal{A}$ over $V$"
- Notation:

# The binomial channel

- The $d$-order binomial extension of a DMC: $V : \mathcal{A} \to \mathcal{B}$ is the DMC
$$V^{\oplus d}[b^d \mid a] = \prod_{i=0}^{d-1} V(b_i \mid a)$$
for $a \in \mathcal{A}, b^d \in \mathcal{B}^d$
- Interpretation: "$d$ independent observations on an input symbol $a \in \mathcal{A}$ over $V$"
- Notation:
  - $I(P_X, V)$ is the mutual information of a DMC $V$ with input distribution $P_X$

# The binomial channel

- The $d$-order binomial extension of a DMC: $V : \mathcal{A} \to \mathcal{B}$ is the DMC
$$V^{\oplus d}[b^d \mid a] = \prod_{i=0}^{d-1} V(b_i \mid a)$$
for $a \in \mathcal{A}, b^d \in \mathcal{B}^d$
- Interpretation: "$d$ independent observations on an input symbol $a \in \mathcal{A}$ over $V$"
- Notation:
  - $I(P_X, V)$ is the mutual information of a DMC $V$ with input distribution $P_X$
  - $\pi_\alpha(d)$ is the Poisson PMF with parameter $\alpha$

# Capacity lower bound (achievable)

Theorem

*The capacity of the DNA channel is lower bounded as*

$$C(\mathsf{DNA}) \geq \max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \pi_\alpha(0)).$$

# Capacity lower bound (achievable)

**Theorem**

*The capacity of the DNA channel is lower bounded as*

$$C(\mathsf{DNA}) \geq \max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \pi_\alpha(0)\right).$$

- Improves best known results: No constraints on $\alpha, \beta, W$!

# Interpretation

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \pi_\alpha(0))$$

- The relative number of molecules sampled $d$ times

# Interpretation

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \pi_\alpha(0))$$

- The relative number of molecules sampled $d$ times
  - The multinomial distribution is "Poissonized"

# Interpretation

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \pi_\alpha(0))$$

- The mutual information of a molecule sampled $d$ times

# Interpretation

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \pi_\alpha(0))$$

- The mutual information of a molecule sampled $d$ times
  - The MI is that of $d$-order binomial channel $W^{\oplus d}$

# Interpretation

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \pi_\alpha(0)\right)$$

- A loss term due to the lack of molecule order

# Interpretation

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \pi_\alpha(0))$$

- A loss term due to the lack of molecule order
  - The cost of (implicit) “indexing”

# Interpretation

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \pi_\alpha(0))$$

- Optimal input distribution should compromise all orders $W^{\oplus d}$

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V\colon \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V : \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)
- Notation: $V_{|\mathcal{B}_0}$ is a $|\mathcal{A}|$ rows, $|\mathcal{B}_0|$ columns submatrix

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V\colon \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)
- Notation: $V_{|\mathcal{B}_0}$ is a $|\mathcal{A}|$ rows, $|\mathcal{B}_0|$ columns submatrix
- Symmetric channels

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V\colon \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)
- Notation: $V_{|\mathcal{B}_0}$ is a $|\mathcal{A}|$ rows, $|\mathcal{B}_0|$ columns submatrix
- Symmetric channels
    - A DMC $V$ is **symmetric** if its rows are permutations of each other and so are the columns [Cover and Thomas]

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V: \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)
- Notation: $V_{|\mathcal{B}_0}$ is a $|\mathcal{A}|$ rows, $|\mathcal{B}_0|$ columns submatrix
- Symmetric channels
  - A DMC $V$ is **symmetric** if its rows are permutations of each other and so are the columns [Cover and Thomas]
    - For example: A modulo-additive channel $B = A \oplus C$

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V : \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)
- Notation: $V_{|\mathcal{B}_0}$ is a $|\mathcal{A}|$ rows, $|\mathcal{B}_0|$ columns submatrix
- Symmetric channels
  - A DMC $V$ is **symmetric** if its rows are permutations of each other and so are the columns [Cover and Thomas]
    - For example: A modulo-additive channel $B = A \oplus C$
  - A DMC $V$ is **weakly symmetric** if its rows are permutations and the columns have equal sums [Cover and Thomas]

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V\colon \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)
- Notation: $V_{|\mathcal{B}_0}$ is a $|\mathcal{A}|$ rows, $|\mathcal{B}_0|$ columns submatrix
- Symmetric channels
  - A DMC $V$ is **symmetric** if its rows are permutations of each other and so are the columns [Cover and Thomas]
    - For example: A modulo-additive channel $B = A \oplus C$
  - A DMC $V$ is **weakly symmetric** if its rows are permutations and the columns have equal sums [Cover and Thomas]
  - A DMC $V$ is **symmetric in Gallager's sense** if there exists a partition $\mathcal{B} = \bigcup_i \mathcal{B}_i$ such that $V_{|\mathcal{B}_i}$ is a symmetric DMC for all $i$

# A digression – symmetric channels

Motivation: When is the capacity lower bound achieving input distribution $P_X^*$ is uniform?

- Identify a DMC $V : \mathcal{A} \to \mathcal{B}$ with its probability transition matrix ($|\mathcal{A}|$ rows, $|\mathcal{B}|$ columns)
- Notation: $V_{|\mathcal{B}_0}$ is a $|\mathcal{A}|$ rows, $|\mathcal{B}_0|$ columns submatrix
- Symmetric channels
  - A DMC $V$ is **symmetric** if its rows are permutations of each other and so are the columns [Cover and Thomas]
    - For example: A modulo-additive channel $B = A \oplus C$
  - A DMC $V$ is **weakly symmetric** if its rows are permutations and the columns have equal sums [Cover and Thomas]
  - A DMC $V$ is **symmetric in Gallager's sense** if there exists a partition $\mathcal{B} = \bigcup_i \mathcal{B}_i$ such that $V_{|\mathcal{B}_i}$ is a symmetric DMC for all $i$
- In all these cases $P_X^*$ for $V$ is uniform

# A digression – binomial extension symmetric channels

- Back to the DNA channel: We need to maximize $\sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$

# A digression – binomial extension symmetric channels

- Back to the DNA channel: We need to maximize $\sum_{d\in\mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$
- Problem: If $W$ is symmetric then $W^{\oplus d}$ is not necessarily symmetric (not even in Gallager's sense)

# A digression – binomial extension symmetric channels

- Back to the DNA channel: We need to maximize $\sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$
- Problem: If $W$ is symmetric then $W^{\oplus d}$ is not necessarily symmetric (not even in Gallager's sense)
- Example:

$$W_1 = \frac{1}{15} \cdot \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 2 & 5 & 1 \\ 2 & 5 & 1 & 3 & 4 \\ 3 & 4 & 5 & 1 & 2 \\ 5 & 1 & 4 & 2 & 3 \end{bmatrix}$$

  but $W_1^{\oplus 2}$ is not symmetric

# A digression – binomial extension symmetric channels

- Back to the DNA channel: We need to maximize $\sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$
- Problem: If $W$ is symmetric then $W^{\oplus d}$ is not necessarily symmetric (not even in Gallager's sense)
- Example:

$$W_1 = \frac{1}{15} \cdot \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 2 & 5 & 1 \\ 2 & 5 & 1 & 3 & 4 \\ 3 & 4 & 5 & 1 & 2 \\ 5 & 1 & 4 & 2 & 3 \end{bmatrix}$$

  but $W_1^{\oplus 2}$ is not symmetric
  - The capacity achieving input distribution is *not* uniform

# A digression – binomial extension symmetric channels

- Back to the DNA channel: We need to maximize $\sum_{d\in\mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$
- Problem: If $W$ is symmetric then $W^{\oplus d}$ is not necessarily symmetric (not even in Gallager's sense)
- Example:

$$W_1 = \frac{1}{15} \cdot \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 2 & 5 & 1 \\ 2 & 5 & 1 & 3 & 4 \\ 3 & 4 & 5 & 1 & 2 \\ 5 & 1 & 4 & 2 & 3 \end{bmatrix}$$

  but $W_1^{\oplus 2}$ is not symmetric
  - The capacity achieving input distribution is *not* uniform

### Proposition

*Let $V : \mathcal{A} \to \mathcal{B}$ be a modulo-additive DMC. Then $V^{\oplus d}$ is symmetric in Gallager's sense for all $d \in \mathbb{N}^+$.*

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

### Proposition

*If $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$, and $W$ is a symmetric channel in Gallager's sense, then the lower bound on the capacity is achieved by the uniform input distribution.*

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

### Proposition

*If $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$, and $W$ is a symmetric channel in Gallager's sense, then the lower bound on the capacity is achieved by the uniform input distribution.*

- Proof:

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

### Proposition

*If $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$, and $W$ is a symmetric channel in Gallager's sense, then the lower bound on the capacity is achieved by the uniform input distribution.*

- Proof:
  - A detailed inspection of all possible channels of $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

## Proposition

*If $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$, and $W$ is a symmetric channel in Gallager's sense, then the lower bound on the capacity is achieved by the uniform input distribution.*

- Proof:
  - A detailed inspection of all possible channels of $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$
  - A taxonomy of small doubly-permutation atoms

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

## Proposition

*If $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$, and $W$ is a symmetric channel in Gallager's sense, then the lower bound on the capacity is achieved by the uniform input distribution.*

- Proof:
  - A detailed inspection of all possible channels of $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$
  - A taxonomy of small doubly-permutation atoms
- **Open questions:**

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

### Proposition

*If $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$, and $W$ is a symmetric channel in Gallager's sense, then the lower bound on the capacity is achieved by the uniform input distribution.*

- Proof:
  - A detailed inspection of all possible channels of $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$
  - A taxonomy of small doubly-permutation atoms
- **Open questions:**
  - When does operations such as binomial extension preserve symmetry?

# Capacity lower bound (achievable) – input distribution

- Also: The counterexample had $|\mathcal{A}| = |\mathcal{B}| = 5$, but $|\mathcal{X}| = 4$ for practical DNA channels

### Proposition

*If $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$, and $W$ is a symmetric channel in Gallager's sense, then the lower bound on the capacity is achieved by the uniform input distribution.*

- Proof:
  - A detailed inspection of all possible channels of $|\mathcal{X}| \leq 4$, $|\mathcal{Y}| \leq |\mathcal{X}|$
  - A taxonomy of small doubly-permutation atoms
- **Open questions:**
  - When does operations such as binomial extension preserve symmetry?
  - How can this systematically be proven?

# Error probability bound – basic definitions

- $\theta_d$ represents the fraction of molecules sampled $d \in \mathbb{N}_+$ times

$$\left\{ \theta_d \geq 0, \qquad \sum_{d \in \mathbb{N}} \theta_d = 1 \right\}$$

# Error probability bound – basic definitions

- $\theta_d$ represents the fraction of molecules sampled $d \in \mathbb{N}_+$ times

$$\left\{ \theta_d \geq 0, \qquad \sum_{d \in \mathbb{N}} \theta_d = 1 \right\}$$

- Denote

$$R(\{\theta_d\}) := \sum_{d \in \mathbb{N}} \theta_d \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \theta_0)$$

# Error probability bound – basic definitions

- $\theta_d$ represents the fraction of molecules sampled $d \in \mathbb{N}_+$ times

$$\left\{ \theta_d \geq 0, \qquad \sum_{d \in \mathbb{N}} \theta_d = 1 \right\}$$

- Denote

$$R(\{\theta_d\}) := \sum_{d \in \mathbb{N}} \theta_d \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \theta_0)$$

  - Can be interpreted as "instantaneous capacity"

# Error probability bound – basic definitions

- $\theta_d$ represents the fraction of molecules sampled $d \in \mathbb{N}_+$ times

$$\left\{ \theta_d \geq 0, \qquad \sum_{d \in \mathbb{N}} \theta_d = 1 \right\}$$

- Denote

$$R(\{\theta_d\}) := \sum_{d \in \mathbb{N}} \theta_d \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \theta_0)$$

  - Can be interpreted as "instantaneous capacity"
  - Note: This is a notation only

# Error probability bound – basic definitions

- $\theta_d$ represents the fraction of molecules sampled $d \in \mathbb{N}_+$ times

$$\left\{ \theta_d \geq 0, \qquad \sum_{d \in \mathbb{N}} \theta_d = 1 \right\}$$

- Denote

$$R(\{\theta_d\}) := \sum_{d \in \mathbb{N}} \theta_d \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}(1 - \theta_0)$$

  - Can be interpreted as "instantaneous capacity"
  - Note: This is a notation only

- $d_{\mathrm{KL}}(p \,||\, q)$ is the binary KL divergence

# Reliability function bound

Theorem

*It holds that*

$$\liminf_{M\to\infty} -\frac{1}{M}\log \overline{\mathsf{pe}}(\mathcal{C},\mathcal{D}) \geq$$

$$\max_{P_X\in\mathcal{P}(\mathcal{X})} \inf_{\{\theta_d\}_{d\in\mathbb{N}}} \sum_{d\in\mathbb{N}} \left(1-\sum_{i\in[d]}\theta_i\right)\cdot d_{KL}\left(\frac{\theta_d}{1-\sum_{i\in[d]}\theta_i} \,\middle\|\, \frac{\pi_\alpha(d)}{1-\sum_{i\in[d]}\pi_\alpha(i)}\right)$$

*where the infimum is subject to*

$$R(\{\theta_d\}) \leq R.$$

# Reliability function bound

**Theorem**

*It holds that*

$$\liminf_{M\to\infty} -\frac{1}{M}\log \overline{\mathsf{pe}}(\mathcal{C},\mathcal{D}) \geq$$

$$\max_{P_X\in\mathcal{P}(\mathcal{X})} \inf_{\{\theta_d\}_{d\in\mathbb{N}}} \sum_{d\in\mathbb{N}} \left(1-\sum_{i\in[d]}\theta_i\right)\cdot d_{KL}\left(\frac{\theta_d}{1-\sum_{i\in[d]}\theta_i} \,\middle\|\, \frac{\pi_\alpha(d)}{1-\sum_{i\in[d]}\pi_\alpha(i)}\right)$$

*where the infimum is subject to*

$$R(\{\theta_d\}) \leq R.$$

- The exponent vanishes when $\theta_d = \pi_\alpha(d)$ for all $d\in\mathbb{N}_+$

# Reliability function bound

**Theorem**

*It holds that*

$$\liminf_{M\to\infty} -\frac{1}{M}\log \overline{\mathsf{pe}}(\mathcal{C},\mathcal{D}) \geq$$

$$\max_{P_X\in\mathcal{P}(\mathcal{X})} \inf_{\{\theta_d\}_{d\in\mathbb{N}}} \sum_{d\in\mathbb{N}} \left(1-\sum_{i\in[d]}\theta_i\right)\cdot d_{KL}\left(\frac{\theta_d}{1-\sum_{i\in[d]}\theta_i}\,\middle\|\,\frac{\pi_\alpha(d)}{1-\sum_{i\in[d]}\pi_\alpha(i)}\right)$$

*where the infimum is subject to*

$$R(\{\theta_d\}) \leq R.$$

- The exponent vanishes when $\theta_d = \pi_\alpha(d)$ for all $d\in\mathbb{N}_+$
- $\Rightarrow$ Capacity lower bound follows as a corollary

$$C(\mathsf{DNA}) \geq R(\{\pi_\alpha(d)\}).$$

# Interpretation

- The exponent is dominated by **outage** $R(\{\theta_d\}) \leq R$

# Interpretation

- The exponent is dominated by **outage** $R(\{\theta_d\}) \leq R$
  - Outage is caused by under-sampled molecules

# Interpretation

- The exponent is dominated by **outage** $R(\{\theta_d\}) \leq R$
  - Outage is caused by under-sampled molecules
- Error probability decays as $e^{-\Theta(M)}$ and not $e^{-\Theta(ML)} = e^{-\Theta(M \log M)}$!

# Interpretation

- The exponent is dominated by **outage** $R(\{\theta_d\}) \leq R$
  - Outage is caused by under-sampled molecules
- Error probability decays as $e^{-\Theta(M)}$ and not $e^{-\Theta(ML)} = e^{-\Theta(M \log M)}$!
- Proof:

# Interpretation

- The exponent is dominated by **outage** $R(\{\theta_d\}) \leq R$
  - Outage is caused by under-sampled molecules
- Error probability decays as $e^{-\Theta(M)}$ and not $e^{-\Theta(ML)} = e^{-\Theta(M \log M)}$!
- Proof:
  1. Standard IID random coding ensemble

# Interpretation

- The exponent is dominated by **outage** $R(\{\theta_d\}) \leq R$
  - Outage is caused by under-sampled molecules
- Error probability decays as $e^{-\Theta(M)}$ and not $e^{-\Theta(ML)} = e^{-\Theta(M \log M)}$!
- Proof:
  1. Standard IID random coding ensemble
  2. High complexity, "optimal-like", decoder

# Comparison to previous schemes

- Very different!

# Comparison to previous schemes

- Very different!
- Encoder:

# Comparison to previous schemes

- Very different!
- Encoder:
  - No explicit indexing of molecules

# Comparison to previous schemes

- Very different!
- Encoder:
  - No explicit indexing of molecules
  - No inner/outer code

# Comparison to previous schemes

- Very different!
- Encoder:
  - No explicit indexing of molecules
  - No inner/outer code
- Decoder:

# Comparison to previous schemes

- Very different!
- Encoder:
  - No explicit indexing of molecules
  - No inner/outer code
- Decoder:
  - No greedy clustering as in [Lenz et al 2019]

# Comparison to previous schemes

- Very different!
- Encoder:
  - No explicit indexing of molecules
  - No inner/outer code
- Decoder:
  - No greedy clustering as in [Lenz et al 2019]
    - Clustering requires defining a metric – suitable to BSC/symmetric channels

# Comparison to previous schemes

- Very different!
- Encoder:
  - No explicit indexing of molecules
  - No inner/outer code
- Decoder:
  - No greedy clustering as in [Lenz et al 2019]
    - Clustering requires defining a metric – suitable to BSC/symmetric channels
    - Hard clustering is the source of limited regime of $(\alpha, \beta, W)$

# Comparison to previous schemes

- Very different!
- Encoder:
  - No explicit indexing of molecules
  - No inner/outer code
- Decoder:
  - No greedy clustering as in [Lenz et al 2019]
    - Clustering requires defining a metric – suitable to BSC/symmetric channels
    - Hard clustering is the source of limited regime of $(\alpha, \beta, W)$
  - Bonus: The decoder is universal w.r.t. $W$

# Proof snippets – sampling types

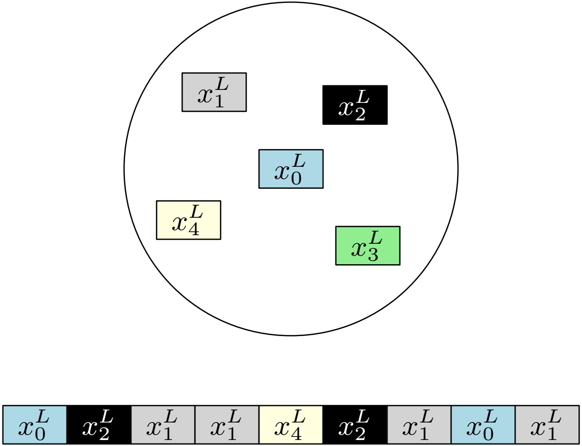


Figure: Sampling types

# Proof snippets – sampling types

- A new notion of *sampling types:*

# Proof snippets – sampling types

- A new notion of *sampling types:*
  - $U^N$ is the *molecule index vector* – $U_n$ is the molecule sampled at the $n$th draw

# Proof snippets – sampling types

- A new notion of *sampling types:*
  - $U^N$ is the *molecule index vector* – $U_n$ is the molecule sampled at the $n$th draw
  - $S^M$ is the *molecule duplicate vector* – $S_m$ is the number of times the $m$th molecule was sampled

# Proof snippets – sampling types

- A new notion of *sampling types:*
  - $U^N$ is the *molecule index vector* – $U_n$ is the molecule sampled at the $n$th draw
  - $S^M$ is the *molecule duplicate vector* – $S_m$ is the number of times the $m$th molecule was sampled
  - $Q^{N+1}$ is the *amplification vector* – $Q_d$ is the number of molecules sampled $d$ times

# Proof snippets – sampling types

- A new notion of *sampling types:*
  - $U^N$ is the *molecule index vector* – $U_n$ is the molecule sampled at the $n$th draw
  - $S^M$ is the *molecule duplicate vector* – $S_m$ is the number of times the $m$th molecule was sampled
  - $Q^{N+1}$ is the *amplification vector* – $Q_d$ is the number of molecules sampled $d$ times
- Related via the *empirical count operator* $\mathscr{N}$

$$Q^{N+1} = \mathscr{N}(S^M) = \mathscr{N}^{(2)}(U^N)$$

# Proof snippets – sampling types

- A new notion of *sampling types:*
  - $U^N$ is the *molecule index vector* – $U_n$ is the molecule sampled at the $n$th draw
  - $S^M$ is the *molecule duplicate vector* – $S_m$ is the number of times the $m$th molecule was sampled
  - $Q^{N+1}$ is the *amplification vector* – $Q_d$ is the number of molecules sampled $d$ times
- Related via the *empirical count operator* $\mathscr{N}$

$$Q^{N+1} = \mathscr{N}(S^M) = \mathscr{N}^{(2)}(U^N)$$

- The analysis requires estimating asymptotic sizes of sampling type classes, e.g.

$$\mathscr{T}^{(2)}_{q^{N+1}} = \left\{ u^N \in [M]^N : \mathscr{N}^{(2)}(u^N) = q^{N+1} \right\}$$

# Proof snippets – sampling types

- A new notion of *sampling types:*
  - $U^N$ is the *molecule index vector* – $U_n$ is the molecule sampled at the $n$th draw
  - $S^M$ is the *molecule duplicate vector* – $S_m$ is the number of times the $m$th molecule was sampled
  - $Q^{N+1}$ is the *amplification vector* – $Q_d$ is the number of molecules sampled $d$ times
- Related via the *empirical count operator* $\mathscr{N}$

$$Q^{N+1} = \mathscr{N}(S^M) = \mathscr{N}^{(2)}(U^N)$$

- The analysis requires estimating asymptotic sizes of sampling type classes, e.g.

$$\mathscr{T}^{(2)}_{q^{N+1}} = \left\{ u^N \in [M]^N \colon \mathscr{N}^{(2)}(u^N) = q^{N+1} \right\}$$

- Estimation via restricted partition numbers [Hardy and Ramanujan, Uspensky, and Rademacher]

# Proof snippets – channel model

- For all $U^N$ with a given $q^{N+1}$ *amplification vector*, the channel operation is "equivalent"

# Proof snippets – channel model

- For all $U^N$ with a given $q^{N+1}$ *amplification vector*, the channel operation is "equivalent"
- A mixture (over orders $d$) of binomial channels $W^{\oplus d}$, with mixing coefficients $\frac{q_d}{M}$

$$\mathcal{L}\left[y^{LN} \mid x^{LM}\right] = \sum_{q^{N+1} \in \mathscr{Q}(M,N)} \mathbb{P}\left[U^N \in \mathscr{T}^{(2)}_{q^{N+1}}\right]$$

$$\times \sum_{u^N \in \mathscr{T}^{(2)}_{q^{N+1}}} \frac{1}{\left|\mathscr{T}^{(2)}_{q^{N+1}}\right|} \prod_{d=0}^{N} W^{\oplus d}\left[b^d_{\mathcal{K}_d(u^N)} \mid a_{\mathcal{K}_d(u^N)}\right]$$

# Proof snippets – decoder

- The decoder is based on a metric (score)

$$\hat{x}^{ML} := \arg\max_{x^{LM} \in \mathcal{C}} \lambda(Y^{LN} \mid x^{LM})$$

# Proof snippets – decoder

- The decoder is based on a metric (score)

$$\hat{x}^{ML} := \arg\max_{x^{LM} \in \mathcal{C}} \lambda(Y^{LN} \mid x^{LM})$$

- Recall: Molecule $U_n$ was sampled at time $n \in [N]$. A sampling vector is $U^N \in [M]^N$

# Proof snippets – decoder

- The decoder is based on a metric (score)

$$\hat{x}^{ML} := \arg\max_{x^{LM} \in \mathcal{C}} \lambda(Y^{LN} \mid x^{LM})$$

- Recall: Molecule $U_n$ was sampled at time $n \in [N]$. A sampling vector is $U^N \in [M]^N$
- The metric is a maximization over all sampling events

$$\lambda(y^{LN} \mid x^{LM}) = \max_{u^N} \lambda(Y^{NL}, x^{ML}; u^N)$$

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \Big[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \Big]$$

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \Big[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \Big]$$

1. Based on the empirical mutual information (MMI)

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \left[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \right]$$

1. Based on the empirical mutual information (MMI)
   - Does not depend on the channel $W$ (universal)

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \Big[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \Big]$$

1. Based on the empirical mutual information (MMI)
   - Does not depend on the channel $W$ (universal)
2. Adapted for the IID ensemble using a KL divergence term

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d\in[N+1]} \theta_d L \cdot \Big[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \Big]$$

1. Based on the empirical mutual information (MMI)
   - Does not depend on the channel $W$ (universal)
2. Adapted for the IID ensemble using a KL divergence term
   - Fixed-composition codewords is problematic:

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \left[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \right]$$

1. Based on the empirical mutual information (MMI)
   - Does not depend on the channel $W$ (universal)
2. Adapted for the IID ensemble using a KL divergence term
   - Fixed-composition codewords is problematic:
     - A full codeword may have fixed composition, but not each molecule

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \Big[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \Big]$$

1. Based on the empirical mutual information (MMI)
   - Does not depend on the channel $W$ (universal)
2. Adapted for the IID ensemble using a KL divergence term
   - Fixed-composition codewords is problematic:
     - A full codeword may have fixed composition, but not each molecule
     - Fixed-composition molecules is too restrictive

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \Big[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \Big]$$

1. Based on the empirical mutual information (MMI)
   - Does not depend on the channel $W$ (universal)
2. Adapted for the IID ensemble using a KL divergence term
   - Fixed-composition codewords is problematic:
     - A full codeword may have fixed composition, but not each molecule
     - Fixed-composition molecules is too restrictive
3. A correction term: Not all sampling events have the same probability (“sampling types”)

# Proof snippets – decoder

- The conditional score:

$$\lambda(y^{LN} \mid x^{LM}; u^N) := -(1-\theta_0) M \log M$$
$$+ \sum_{d \in [N+1]} \theta_d L \cdot \left[ D(\hat{P}^d(x^{LM}; u^N) \,||\, P_X) + I_{\hat{P}^d(x^{LM}, y^{LN}; u^N)}(A; B^d) \right]$$

1. Based on the empirical mutual information (MMI)
   - Does not depend on the channel $W$ (universal)
2. Adapted for the IID ensemble using a KL divergence term
   - Fixed-composition codewords is problematic:
     - A full codeword may have fixed composition, but not each molecule
     - Fixed-composition molecules is too restrictive
3. A correction term: Not all sampling events have the same probability ("sampling types")
   - Inspired by the analysis of [Csiszár 1980] for joint source-channel coding

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder
  - Method of types (on steroids...)

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder
    - Method of types (on steroids...)
    - An obstacle:

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder
  - Method of types (on steroids...)
  - An obstacle:
    - The order $d \in [N+1]$ is unbounded as $N$ increases

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder
  - Method of types (on steroids...)
  - An obstacle:
    - The order $d \in [N+1]$ is unbounded as $N$ increases
    - The maximal output alphabet size of $\{V^{\oplus d}\}_{d\in[N]}$ increases with blocklength!

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder
  - Method of types (on steroids...)
  - An obstacle:
    - The order $d \in [N+1]$ is unbounded as $N$ increases
    - The maximal output alphabet size of $\{V^{\oplus d}\}_{d \in [N]}$ increases with blocklength!
  - Solution: A careful truncation argument

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder
  - Method of types (on steroids...)
  - An obstacle:
    - The order $d \in [N+1]$ is unbounded as $N$ increases
    - The maximal output alphabet size of $\{V^{\oplus d}\}_{d \in [N]}$ increases with blocklength!
  - Solution: A careful truncation argument
    - Assuming $q_d = 0$ for all $d \geq \overline{d}$

# Proof snippets – error probability analysis

- Condition on a given amplification vector $Q^{N+1} = q^{N+1}$
- Random coding analysis of the error probability of the universal decoder
  - Method of types (on steroids...)
  - An obstacle:
    - The order $d \in [N+1]$ is unbounded as $N$ increases
    - The maximal output alphabet size of $\{V^{\oplus d}\}_{d \in [N]}$ increases with blocklength!
  - Solution: A careful truncation argument
    - Assuming $q_d = 0$ for all $d \geq \overline{d}$
    - $\overline{d}$ is optimized later on

# Proof snippets – error probability analysis

- The conditional analysis shows that if

$$R \le \Gamma_{\overline{d}}(q^{N+1}) := \sum_{d \in [\overline{d}+1]} \frac{q_d}{M} \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \frac{q_0}{M}\right)$$

  then the conditional error probability decays as $e^{-\Theta(ML)}$

# Proof snippets – error probability analysis

- The conditional analysis shows that if

$$R \leq \Gamma_{\overline{d}}(q^{N+1}) := \sum_{d \in [\overline{d}+1]} \frac{q_d}{M} \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \frac{q_0}{M}\right)$$

  then the conditional error probability decays as $e^{-\Theta(ML)}$
- It holds

$$\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] = e^{-\Theta(M)}$$

# Proof snippets – error probability analysis

- The conditional analysis shows that if

$$R \le \Gamma_{\overline{d}}(q^{N+1}) := \sum_{d \in [\overline{d}+1]} \frac{q_d}{M} \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \frac{q_0}{M}\right)$$

  then the conditional error probability decays as $e^{-\Theta(ML)}$
- It holds

$$\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] = e^{-\Theta(M)}$$

- Hence

$$\overline{\mathsf{pe}}(\mathcal{C}, \mathcal{D}) \le \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) \ge R] \cdot e^{-\Theta(ML)} + \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] \cdot 1.$$

  **A "bad" sampling event dominates the error probability!**

# Proof snippets – error probability analysis

- The conditional analysis shows that if

$$R \leq \Gamma_{\overline{d}}(q^{N+1}) := \sum_{d \in [\overline{d}+1]} \frac{q_d}{M} \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \frac{q_0}{M}\right)$$

  then the conditional error probability decays as $e^{-\Theta(ML)}$
- It holds

$$\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] = e^{-\Theta(M)}$$

- Hence

$$\overline{\mathsf{pe}}(\mathcal{C}, \mathcal{D}) \leq \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) \geq R] \cdot e^{-\Theta(ML)} + \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] \cdot 1.$$

  **A "bad" sampling event dominates the error probability!**
- Evaluation of $\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R]$:

# Proof snippets – error probability analysis

- The conditional analysis shows that if

$$R \leq \Gamma_{\overline{d}}(q^{N+1}) := \sum_{d \in [\overline{d}+1]} \frac{q_d}{M} \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \frac{q_0}{M}\right)$$

  then the conditional error probability decays as $e^{-\Theta(ML)}$
- It holds

$$\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] = e^{-\Theta(M)}$$

- Hence

$$\overline{\mathsf{pe}}(\mathcal{C}, \mathcal{D}) \leq \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) \geq R] \cdot e^{-\Theta(ML)} + \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] \cdot 1.$$

  **A "bad" sampling event dominates the error probability!**
- Evaluation of $\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R]$:
  - The vector $Q^{N+1}$ is the empirical count of a *multinomial* $S^M$

# Proof snippets – error probability analysis

- The conditional analysis shows that if

$$R \le \Gamma_{\overline{d}}(q^{N+1}) := \sum_{d \in [\overline{d}+1]} \frac{q_d}{M} \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \frac{q_0}{M}\right)$$

  then the conditional error probability decays as $e^{-\Theta(ML)}$
- It holds

$$\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] = e^{-\Theta(M)}$$

- Hence

$$\overline{\mathsf{pe}}(\mathcal{C}, \mathcal{D}) \le \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) \ge R] \cdot e^{-\Theta(ML)} + \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] \cdot 1.$$

  **A "bad" sampling event dominates the error probability!**
- Evaluation of $\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R]$:
  - The vector $Q^{N+1}$ is the empirical count of a *multinomial* $S^M$
  - The multinomial distribution is *Poissonized*

# Proof snippets – error probability analysis

- The conditional analysis shows that if

$$R \leq \Gamma_{\overline{d}}(q^{N+1}) := \sum_{d \in [\overline{d}+1]} \frac{q_d}{M} \cdot I(P_X, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \frac{q_0}{M}\right)$$

  then the conditional error probability decays as $e^{-\Theta(ML)}$
- It holds

$$\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] = e^{-\Theta(M)}$$

- Hence

$$\overline{\mathsf{pe}}(\mathcal{C},\mathcal{D}) \leq \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) \geq R] \cdot e^{-\Theta(ML)} + \mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R] \cdot 1.$$

  **A "bad" sampling event dominates the error probability!**
- Evaluation of $\mathbb{P}[\Gamma_{\overline{d}}(Q^{N+1}) < R]$:
  - The vector $Q^{N+1}$ is the empirical count of a *multinomial* $S^M$
  - The multinomial distribution is *Poissonized*
    - Typically for expectations, here for tails

# Error exponent for fixed uniform sampling

- Suppose that each molecule is equally sampled
  $S_m = \alpha = \frac{N}{M}$ for all $m \in [M]$

# Error exponent for fixed uniform sampling

- Suppose that each molecule is equally sampled $S_m = \alpha = \frac{N}{M}$ for all $m \in [M]$

### Theorem

*Assume the ideal sampling of $S_m = \alpha$ for all $m \in [M]$ with probability 1. Then,*

$$\liminf_{M \to \infty} -\frac{1}{ML} \log \overline{\mathsf{pe}}(\mathcal{C}, \mathcal{D})$$
$$\geq \max_{P_X \in \mathcal{P}(\mathcal{X})} \min_{Q_{XY^\alpha} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}^\alpha)} D(Q_X \,||\, P_X) + D(Q_{Y^\alpha|X} \,||\, W^{\oplus\alpha} | Q_X)$$
$$+ \left[ D(Q_A \,||\, P_X) + I_Q(X; Y^\alpha) - \frac{1}{\beta} - R \right]_+ .$$

# Error exponent for fixed uniform sampling

- Suppose that each molecule is equally sampled $S_m = \alpha = \frac{N}{M}$ for all $m \in [M]$

**Theorem**

*Assume the ideal sampling of $S_m = \alpha$ for all $m \in [M]$ with probability 1. Then,*

$$\liminf_{M\to\infty} -\frac{1}{ML}\log \overline{\mathsf{pe}}(\mathcal{C},\mathcal{D}) \geq \max_{P_X\in\mathcal{P}(\mathcal{X})} \min_{Q_{XY^\alpha}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y}^\alpha)} D(Q_X \,||\, P_X) + D(Q_{Y^\alpha|X} \,||\, W^{\oplus\alpha}|Q_X) + \left[D(Q_A \,||\, P_X) + I_Q(X;Y^\alpha) - \frac{1}{\beta} - R\right]_+ .$$

- Despite loss of order, the error probability decays as $e^{-\Theta(ML)} = e^{-\Theta(M\log M)}$!

# Outline

# Capacity upper bound (converse) – definitions

- The *common-input MI deficit*

$$\mathsf{CID}(P_X, V) = 2 \cdot I(P_X, V) - I(P_X, V^{\oplus 2})$$

# Capacity upper bound (converse) – definitions

- The *common-input MI deficit*

$$\mathsf{CID}(P_X, V) = 2 \cdot I(P_X, V) - I(P_X, V^{\oplus 2})$$

  - Intuitively: The difference in mutual information for two independent inputs vs. identical inputs

# Capacity upper bound (converse) – definitions

- The *common-input MI deficit*

$$\mathsf{CID}(P_X, V) = 2 \cdot I(P_X, V) - I(P_X, V^{\oplus 2})$$

  - Intuitively: The difference in mutual information for two independent inputs vs. identical inputs

- The *d-order excess-rate* term by

$$\Omega_d(\beta, P_X, W) := \left[\min\left\{\frac{1}{\beta}, \frac{2}{\beta} - \mathsf{CID}(P_X, W^{\oplus d})\right\}\right]_+$$

# Capacity upper bound (converse)

Theorem

*Assume that* $\min_{x \in \mathcal{X},\, y \in \mathcal{Y}} W(y \mid x) > 0$. *Then, the capacity of the DNA channel is upper bounded as*

$$C(\mathsf{DNA}) \leq \max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot \left[ I(P_X, W^{\oplus d}) + \Omega_d(\beta, P_X, W) \right] - \frac{1}{\beta} \left(1 - \pi_\alpha(0)\right).$$

# Capacity upper bound (converse)

**Theorem**

*Assume that* $\min_{x\in\mathcal{X},\, y\in\mathcal{Y}} W(y\mid x) > 0$. *Then, the capacity of the DNA channel is upper bounded as*

$$C(\mathsf{DNA}) \le \max_{P_X\in\mathcal{P}(\mathcal{X})} \sum_{d\in\mathbb{N}^+} \pi_\alpha(d)\cdot\left[I(P_X, W^{\oplus d}) + \Omega_d(\beta, P_X, W)\right] - \frac{1}{\beta}\left(1-\pi_\alpha(0)\right).$$

- Similar to the lower bound, except for $\Omega_d(\cdot)$

# Tightness of the bound

Corollary

*Let*

$$P_X^*(\alpha,\beta,W) \in \arg\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot \left[ I(P_X, W^{\oplus d}) + \Omega_d(\beta, P_X, W) \right],$$

*and let*

$$\beta_{cr}(\alpha, W) := \min \left\{ \beta\colon \beta \geq \frac{2}{\mathsf{CID}(P_X^*(\alpha,\beta,W), W)} \right\}$$

*Then, for all* $\beta \geq \beta_{cr}(\alpha, W)$

$$C(\mathsf{DNA}) = \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I\left( P_X^*(\alpha, \beta_{cr}(\alpha,W), W), W^{\oplus d} \right) - \frac{1}{\beta}\left(1 - \pi_\alpha(0)\right).$$

# Proof idea

- Warning: The fulle proof is very complicated and long

# Proof idea

- Warning: The fulle proof is very complicated and long
  - Builds on the ideas of [Shomrony and Heckel 2021], [Lenz et al 2020]

# Proof idea

- Warning: The fulle proof is very complicated and long
  - Builds on the ideas of [Shomrony and Heckel 2021], [Lenz et al 2020]
- Goal: By Fano's inequality, bounding $I(X^{LM};Y^{LN})$ for *any* input distribution

# Proof idea

- Warning: The fulle proof is very complicated and long
  - Builds on the ideas of [Shomrony and Heckel 2021], [Lenz et al 2020]
- Goal: By Fano's inequality, bounding $I(X^{LM};Y^{LN})$ for *any* input distribution
- An *easy* converse

$$I(X^{LM};Y^{LN}) \leq \max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$$

# Proof idea

- Warning: The fulle proof is very complicated and long
  - Builds on the ideas of [Shomrony and Heckel 2021], [Lenz et al 2020]
- Goal: By Fano's inequality, bounding $I(X^{LM};Y^{LN})$ for *any* input distribution
- An *easy* converse

$$I(X^{LM};Y^{LN}) \leq \max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$$

- Problem: Missing the $-\frac{1}{\beta}(1-\pi_\alpha(0))$ term

# Proof idea

- Warning: The fulle proof is very complicated and long
  - Builds on the ideas of [Shomrony and Heckel 2021], [Lenz et al 2020]
- Goal: By Fano's inequality, bounding $I(X^{LM};Y^{LN})$ for *any* input distribution
- An *easy* converse

$$I(X^{LM};Y^{LN}) \leq \max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$$

- Problem: Missing the $-\frac{1}{\beta}(1-\pi_\alpha(0))$ term
  - Does a decoder of an optimal system must know which molecules have been sampled after correct decoding?

# Proof idea

- Warning: The fulle proof is very complicated and long
  - Builds on the ideas of [Shomrony and Heckel 2021], [Lenz et al 2020]
- Goal: By Fano's inequality, bounding $I(X^{LM}; Y^{LN})$ for *any* input distribution
- An *easy* converse

$$I(X^{LM}; Y^{LN}) \leq \max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X, W^{\oplus d})$$

- Problem: Missing the $-\frac{1}{\beta}(1 - \pi_\alpha(0))$ term
  - Does a decoder of an optimal system must know which molecules have been sampled after correct decoding?
  - Does a molecule must contain implicit information on its index?

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$
    - Low MI $I(P_X, W^{\oplus 2})$

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$
    - Low MI $I(P_X, W^{\oplus 2})$
    - Loss of order is immaterial

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$
    - Low MI $I(P_X, W^{\oplus 2})$
    - Loss of order is immaterial
  - Option 2: Independent molecules $X_1^L \overset{\text{IID}}{\sim} P_X \perp\!\!\!\perp X_2^L \overset{\text{IID}}{\sim} P_X$

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$
    - Low MI $I(P_X, W^{\oplus 2})$
    - Loss of order is immaterial
  - Option 2: Independent molecules $X_1^L \overset{\text{IID}}{\sim} P_X \perp\!\!\!\perp X_2^L \overset{\text{IID}}{\sim} P_X$
    - High MI $2I(P_X, W)$

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$
    - Low MI $I(P_X, W^{\oplus 2})$
    - Loss of order is immaterial
  - Option 2: Independent molecules $X_1^L \overset{\text{IID}}{\sim} P_X \perp\!\!\!\perp X_2^L \overset{\text{IID}}{\sim} P_X$
    - High MI $2I(P_X, W)$
    - Loss of order causes loss of $-\frac{1}{\beta}(1 - \pi_\alpha(0))$

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$
    - Low MI $I(P_X, W^{\oplus 2})$
    - Loss of order is immaterial
  - Option 2: Independent molecules $X_1^L \overset{\text{IID}}{\sim} P_X \perp\!\!\!\perp X_2^L \overset{\text{IID}}{\sim} P_X$
    - High MI $2I(P_X, W)$
    - Loss of order causes loss of $-\frac{1}{\beta}(1 - \pi_\alpha(0))$
- Optimal choice depends on $\beta$ and $W$

# Proof – main challenge

- Challenge: Characterizing optimal distribution on molecules. Why?
- Illustration: $M = 2$
  - Option 1: Identical molecules $X_1^L = X_2^L \overset{\text{IID}}{\sim} P_X$
    - Low MI $I(P_X, W^{\oplus 2})$
    - Loss of order is immaterial
  - Option 2: Independent molecules $X_1^L \overset{\text{IID}}{\sim} P_X \perp\!\!\!\perp X_2^L \overset{\text{IID}}{\sim} P_X$
    - High MI $2I(P_X, W)$
    - Loss of order causes loss of $-\frac{1}{\beta}(1 - \pi_\alpha(0))$
- Optimal choice depends on $\beta$ and $W$
- In the regime where capacity is known, *independent* molecules are optimal

# Proof outline – “far” and “close” molecules

- Challenge: What are “similar” and “independent” molecules?

# Proof outline – “far” and “close” molecules

- Challenge: What are “similar” and “independent” molecules?
- In [Lenz et al 2020] for a BSC with crossover $w$:

# Proof outline – “far” and “close” molecules

- Challenge: What are “similar” and “independent” molecules?
- In [Lenz et al 2020] for a BSC with crossover $w$:
  - Molecules are different if $d_H(x_i^L, x_j^L) \geq 4wL$

# Proof outline – “far” and “close” molecules

- Challenge: What are “similar” and “independent” molecules?
- In [Lenz et al 2020] for a BSC with crossover $w$:
  - Molecules are different if $d_H(x_i^L, x_j^L) \geq 4wL$
- In our work: Soft similarity measure, based on conditional typical sets

# Proof outline – “far” and “close” molecules

- Challenge: What are “similar” and “independent” molecules?
- In [Lenz et al 2020] for a BSC with crossover $w$:
  - Molecules are different if $d_H(x_i^L, x_j^L) \geq 4wL$
- In our work: Soft similarity measure, based on conditional typical sets
  - The conditional typical set $\mathcal{T}_L([W] \mid x_0^L) \subset \mathcal{Y}^L$ has *high* probability when $Y^L \sim W^L(\cdot \mid x_0^L)$

# Proof outline – “far” and “close” molecules

- Challenge: What are “similar” and “independent” molecules?
- In [Lenz et al 2020] for a BSC with crossover $w$:
  - Molecules are different if $d_H(x_i^L, x_j^L) \geq 4wL$
- In our work: Soft similarity measure, based on conditional typical sets
  - The conditional typical set $\mathcal{T}_L([W] \mid x_0^L) \subset \mathcal{Y}^L$ has *high* probability when $Y^L \sim W^L(\cdot \mid x_0^L)$
  - $x_1^L$ is “far” (“independent”) from $x_0^L$ if $\mathcal{T}_L([W] \mid x_0^L)$ has *low* probability when $Y^L \sim W^L(\cdot \mid x_1^L)$

# Proof outline – “far” and “close” molecules

- Challenge: What are “similar” and “independent” molecules?
- In [Lenz et al 2020] for a BSC with crossover $w$:
  - Molecules are different if $d_H(x_i^L, x_j^L) \geq 4wL$
- In our work: Soft similarity measure, based on conditional typical sets
  - The conditional typical set $\mathcal{T}_L([W] \mid x_0^L) \subset \mathcal{Y}^L$ has *high* probability when $Y^L \sim W^L(\cdot \mid x_0^L)$
  - $x_1^L$ is “far” (“independent”) from $x_0^L$ if $\mathcal{T}_L([W] \mid x_0^L)$ has *low* probability when $Y^L \sim W^L(\cdot \mid x_1^L)$
  - The required distance is **sub-linear** *in* $L$

# Proof outline – properties of “far” and “close” molecules

1. Let a *set* of $\Theta(M)$ pairwise “far” molecules be input to a permuting DNA channel

# Proof outline – properties of “far” and “close” molecules

1. Let a *set* of $\Theta(M)$ pairwise “far” molecules be input to a permuting DNA channel
   - Establish that observing the input and output molecules gain information on the channel permutation

# Proof outline – properties of “far” and “close” molecules

1. Let a *set* of $\Theta(M)$ pairwise “far” molecules be input to a permuting DNA channel
   - Establish that observing the input and output molecules gain information on the channel permutation
     - The equivocation given input and output is negligible compared to unconditional entropy

# Proof outline – properties of “far” and “close” molecules

1. Let a *set* of $\Theta(M)$ pairwise “far” molecules be input to a permuting DNA channel
   - Establish that observing the input and output molecules gain information on the channel permutation
     - The equivocation given input and output is negligible compared to unconditional entropy
2. Let a pair of “close” molecules be given

# Proof outline – properties of “far” and “close” molecules

1. Let a *set* of $\Theta(M)$ pairwise “far” molecules be input to a permuting DNA channel
    - Establish that observing the input and output molecules gain information on the channel permutation
        - The equivocation given input and output is negligible compared to unconditional entropy
2. Let a pair of “close” molecules be given
    - Establish that the mutual information is essentially as if they are identical $I(P_X, V^{\oplus 2})$

# Proof outline – bounding mutual information

- Upper bound the mutual information $I(X^{LM};Y^{LM})$ for Fano's argument

# Proof outline – bounding mutual information

- Upper bound the mutual information $I(X^{LM}; Y^{LM})$ for Fano's argument
  - Use a fixed composition codebook

# Proof outline – bounding mutual information

- Upper bound the mutual information $I(X^{LM};Y^{LM})$ for Fano's argument
  - Use a fixed composition codebook
  - A genie-aided decoder [Lenz et al 2019], that *cluster* output molecules to $\tilde{Y}^{LM}$
    $\Rightarrow$ Establish an upper bound on $I(X^{LM};\tilde{Y}^{LM})$

# Proof outline – bounding mutual information

- Upper bound the mutual information $I(X^{LM};Y^{LM})$ for Fano's argument
  - Use a fixed composition codebook
  - A genie-aided decoder [Lenz et al 2019], that *cluster* output molecules to $\tilde{Y}^{LM}$
    $\Rightarrow$ Establish an upper bound on $I(X^{LM};\tilde{Y}^{LM})$
  - Condition of $Q^{N+1}$: A subset of the input molecules is pairwise "far", the other subset is a "close" neighbor in the first set

# Proof outline – bounding mutual information

- Upper bound the mutual information $I(X^{LM};Y^{LM})$ for Fano's argument
  - Use a fixed composition codebook
  - A genie-aided decoder [Lenz et al 2019], that *cluster* output molecules to $\tilde{Y}^{LM}$
    $\Rightarrow$ Establish an upper bound on $I(X^{LM};\tilde{Y}^{LM})$
  - Condition of $Q^{N+1}$: A subset of the input molecules is pairwise "far", the other subset is a "close" neighbor in the first set
  - The tightest bound obtained for all pairwise "far" molecules

# Proof outline – a clustering decoder

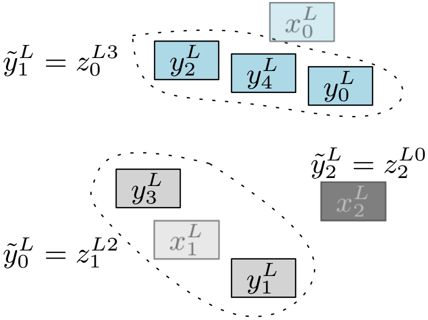


Figure: A clustering decoder

# Proof outline – bounding mutual information

- Bound the average MI over $Q^{N+1}$ via Poissonization

# Proof outline – bounding mutual information

- Bound the average MI over $Q^{N+1}$ via Poissonization
- "Single-letterization" is done in two stages (from $ML$ to $L$ and from $L$ to 1)

# Proof outline – bounding mutual information

- Bound the average MI over $Q^{N+1}$ via Poissonization
- "Single-letterization" is done in two stages (from $ML$ to $L$ and from $L$ to 1)
- Removing the fixed composition assumption

# Proof outline – bounding mutual information

- Bound the average MI over $Q^{N+1}$ via Poissonization
- "Single-letterization" is done in two stages (from $ML$ to $L$ and from $L$ to 1)
- Removing the fixed composition assumption
- Obtaining a bound in which $P_X$ is optimized once for all orders $d$

# A prospective refinement of the upper bound

- Recall: The CID is defined with a pair of molecules

$$\mathsf{CID}(P_X, V) = 2 \cdot I(P_X, V) - I(P_X, V^{\oplus 2})$$

# A prospective refinement of the upper bound

- Recall: The CID is defined with a pair of molecules

$$\mathsf{CID}(P_X, V) = 2 \cdot I(P_X, V) - I(P_X, V^{\oplus 2})$$

- Idea: Generalize to a scattering measure for *triplets* of molecules

# A prospective refinement of the upper bound

- Recall: The CID is defined with a pair of molecules

$$\mathsf{CID}(P_X, V) = 2 \cdot I(P_X, V) - I(P_X, V^{\oplus 2})$$

- Idea: Generalize to a scattering measure for *triplets* of molecules
- Why not quadruplets? quintuplets?

# A prospective refinement of the upper bound

- Recall: The CID is defined with a pair of molecules

$$\mathsf{CID}(P_X, V) = 2 \cdot I(P_X, V) - I(P_X, V^{\oplus 2})$$

- Idea: Generalize to a scattering measure for *triplets* of molecules
- Why not quadruplets? quintuplets?
- The game is *(most likely)* not worth the (*our*) candle

# The assumption on the channel

- Recall: $\min_{x \in \mathcal{X},\, y \in \mathcal{Y}} W(y \mid x) > 0$ is a qualifying condition for the converse

# The assumption on the channel

- Recall: $\min_{x \in \mathcal{X},\, y \in \mathcal{Y}} W(y \mid x) > 0$ is a qualifying condition for the converse
- Technically: Originates from the use of the *blowing-up lemma* in the proof

# The assumption on the channel

- Recall: $\min_{x\in\mathcal{X},\, y\in\mathcal{Y}} W(y \mid x) > 0$ is a qualifying condition for the converse
- Technically: Originates from the use of the *blowing-up lemma* in the proof
- Example: binary erasure sequencing channel

# The assumption on the channel

- Recall: $\min_{x \in \mathcal{X},\, y \in \mathcal{Y}} W(y \mid x) > 0$ is a qualifying condition for the converse
- Technically: Originates from the use of the *blowing-up lemma* in the proof
- Example: binary erasure sequencing channel
- Fundamentally: If $W(y \mid x) = 0$ then molecule ordering is easier

# The assumption on the channel

- Recall: $\min_{x\in\mathcal{X},\, y\in\mathcal{Y}} W(y \mid x) > 0$ is a qualifying condition for the converse
- Technically: Originates from the use of the *blowing-up lemma* in the proof
- Example: binary erasure sequencing channel
- Fundamentally: If $W(y \mid x) = 0$ then molecule ordering is easier
- Open problem: Not obvious if this is just a technical condition that can be removed

# A side result: MI for IID v.s. fixed composition inputs

**Lemma**

*Let $P_A \in \mathcal{P}_K(\mathcal{A})$ be a type for length $K$. Also let $A^K \sim P_A$ IID and $\tilde{A}^K \sim Uniform[\mathcal{T}_K(P_A)]$, and let $B^K$ and $\tilde{B}^K$ be their outputs over a DMC. Then*

$$0 \leq I(A^K; B^K) - I(\tilde{A}^K; \tilde{B}^K) = O(\sqrt{K} \cdot \log K).$$

# A side result: MI for IID v.s. fixed composition inputs

**Lemma**

*Let $P_A \in \mathcal{P}_K(\mathcal{A})$ be a type for length $K$. Also let $A^K \sim P_A$ IID and $\tilde{A}^K \sim Uniform[\mathcal{T}_K(P_A)]$, and let $B^K$ and $\tilde{B}^K$ be their outputs over a DMC. Then*

$$0 \leq I(A^K; B^K) - I(\tilde{A}^K; \tilde{B}^K) = O(\sqrt{K} \cdot \log K).$$

- Proof:

# A side result: MI for IID v.s. fixed composition inputs

**Lemma**

*Let $P_A \in \mathcal{P}_K(\mathcal{A})$ be a type for length $K$. Also let $A^K \sim P_A$ IID and $\tilde{A}^K \sim Uniform[\mathcal{T}_K(P_A)]$, and let $B^K$ and $\tilde{B}^K$ be their outputs over a DMC. Then*

$$0 \leq I(A^K; B^K) - I(\tilde{A}^K; \tilde{B}^K) = O(\sqrt{K} \cdot \log K).$$

- Proof:
  - Bounding entropy differences via Ornstein's $\overline{d}$-distance [Polyanskiy and Wu 2016]

# A side result: MI for IID v.s. fixed composition inputs

Lemma

*Let $P_A \in \mathcal{P}_K(\mathcal{A})$ be a type for length $K$. Also let $A^K \sim P_A$ IID and $\tilde{A}^K \sim Uniform[\mathcal{T}_K(P_A)]$, and let $B^K$ and $\tilde{B}^K$ be their outputs over a DMC. Then*

$$0 \leq I(A^K;B^K) - I(\tilde{A}^K;\tilde{B}^K) = O(\sqrt{K} \cdot \log K).$$

- Proof:
  - Bounding entropy differences via Ornstein's $\overline{d}$-distance [Polyanskiy and Wu 2016]
  - Bounding $\overline{d}$-distance by a KL divergence via Marton's transportation inequality [Marton 1996]

# A side result: MI for IID v.s. fixed composition inputs

### Lemma

*Let $P_A \in \mathcal{P}_K(\mathcal{A})$ be a type for length $K$. Also let $A^K \sim P_A$ IID and $\tilde{A}^K \sim Uniform[\mathcal{T}_K(P_A)]$, and let $B^K$ and $\tilde{B}^K$ be their outputs over a DMC. Then*

$$0 \le I(A^K; B^K) - I(\tilde{A}^K; \tilde{B}^K) = O(\sqrt{K} \cdot \log K).$$

- Proof:
  - Bounding entropy differences via Ornstein's $\overline{d}$-distance [Polyanskiy and Wu 2016]
  - Bounding $\overline{d}$-distance by a KL divergence via Marton's transportation inequality [Marton 1996]
- A refined bound appears in [Tang and Polyanskiy 2022]

# Outline

### Proposition

*Let $W$ be a modulo-additive channel, let $P_X^{(unif)}$ be the uniform distribution over $\mathcal{X}$. Then, for all*

$$\beta \geq \frac{2}{\mathsf{CID}(P_X^{(unif)}, W)}$$

*it holds that*

$$C(\mathsf{DNA}) = \sum_{d \in \mathbb{N}^+} \pi_\alpha(d) \cdot I(P_X^{(unif)}, W^{\oplus d}) - \frac{1}{\beta}\left(1 - \pi_\alpha(0)\right).$$

# Binary symmetric channels

- For a BSC with crossover probability $w$
$$\beta \geq \frac{2}{\log 2 - h_b(2w(1-w))}.$$

# Binary symmetric channels

- For a BSC with crossover probability $w$

$$\beta \geq \frac{2}{\log 2 - h_b(2w(1-w))}.$$

- [Lenz et al 2019-2020]: Only for $w < 1/8$

$$\beta > \overline{\beta}_{\text{cr}} := \frac{2}{\log 2 - h_b(4w)}.$$

# Binary symmetric channels – critical molecule length

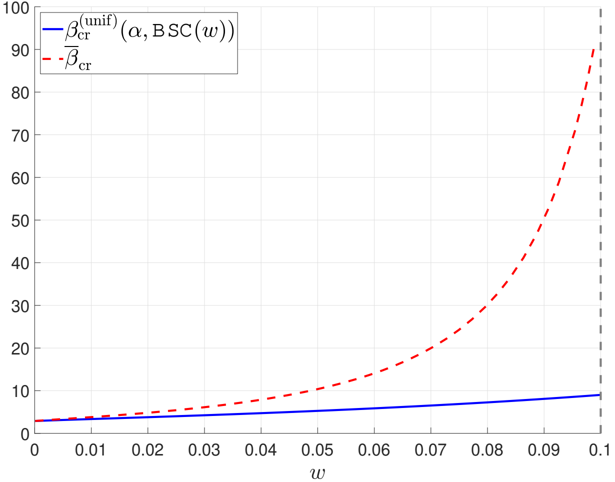


Figure: Comparison between [Lenz 2019] and our result

# Numerical computation

- Given input distribution $P_X$, all bounds can be accurately computed by convex optimization

# Numerical computation

- Given input distribution $P_X$, all bounds can be accurately computed by convex optimization
- Example: Asymmetric channel $|\mathcal{X}| = |\mathcal{Y}| = 4$

$$W_0(y \mid x) = \frac{1}{100} \cdot \begin{bmatrix} 94 & 2 & 2 & 2 \\ 2 & 70 & 25 & 3 \\ 3 & 2 & 85 & 10 \\ 10 & 5 & 5 & 80 \end{bmatrix}$$

# Numerical computation

- Given input distribution $P_X$, all bounds can be accurately computed by convex optimization
- Example: Asymmetric channel $|\mathcal{X}| = |\mathcal{Y}| = 4$

$$W_0(y \mid x) = \frac{1}{100} \cdot \begin{bmatrix} 94 & 2 & 2 & 2 \\ 2 & 70 & 25 & 3 \\ 3 & 2 & 85 & 10 \\ 10 & 5 & 5 & 80 \end{bmatrix}$$

- Uniform input distribution $P_X = (1/4, 1/4, 1/4, 1/4)$ (sub-optimal)

# A numerical example – capacity

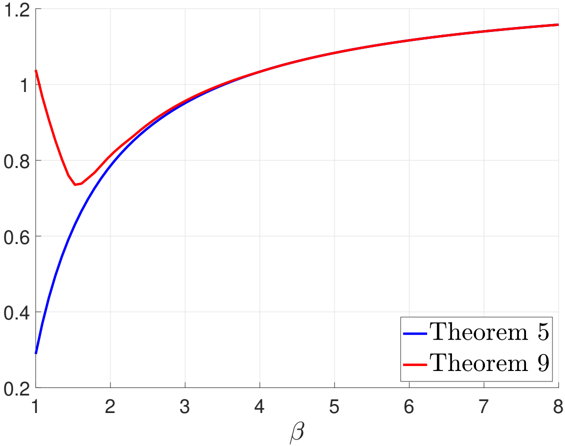


Figure: Upper and lower bounds on $C(\mathsf{DNA}(5, \beta, W_0))$ as a function of $\beta$ (in nats).

# A numerical example – reliability function

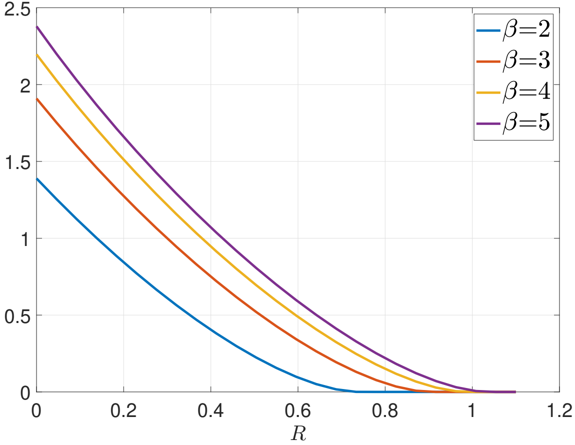


Figure: Right: Lower bound on the reliability function $E^*(R, \mathsf{DNA}(5, \beta, W_0), \{M\})$ as a function of $R$ (in nats).

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?
   - Finite blocklength analysis? slow decay rates $O(\frac{1}{\log M})$ to limits

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?
   - Finite blocklength analysis? slow decay rates $O(\frac{1}{\log M})$ to limits
2. Error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?
   - Finite blocklength analysis? slow decay rates $O(\frac{1}{\log M})$ to limits
2. Error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$
   - An outage behavior of the channel

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?
   - Finite blocklength analysis? slow decay rates $O(\frac{1}{\log M})$ to limits
2. Error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$
   - An outage behavior of the channel
   - Unlike for capacity, increasing $\alpha$ is not marginal in gain

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?
   - Finite blocklength analysis? slow decay rates $O(\frac{1}{\log M})$ to limits
2. Error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$
   - An outage behavior of the channel
   - Unlike for capacity, increasing $\alpha$ is not marginal in gain

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?
   - Finite blocklength analysis? slow decay rates $O(\frac{1}{\log M})$ to limits
2. Error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$
   - An outage behavior of the channel
   - Unlike for capacity, increasing $\alpha$ is not marginal in gain

N. Weinberger and N. Merhav,
"The DNA Storage Channel: Capacity and Error Probability Bounds,"
IT-T, May 2022

# Conclusion and open problems

1. Capacity is **settled** in the low-noise/high-$\beta$ regime for any DMC $W$
   - What is the capacity in the high-noise/low-$\beta$ regime?
   - Finite blocklength analysis? slow decay rates $O(\frac{1}{\log M})$ to limits
2. Error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$
   - An outage behavior of the channel
   - Unlike for capacity, increasing $\alpha$ is not marginal in gain

N. Weinberger and N. Merhav,
"The DNA Storage Channel: Capacity and Error Probability Bounds,"
IT-T, May 2022

# Outline

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
    - The computation of the metric of a single codeword has complexity $\Theta(M^N)$

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on
3. Codeword length is $ML$ but error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$ (DMC)

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on
3. Codeword length is $ML$ but error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$ (DMC)
   - Improve by increasing coverage depth scaling

$$\alpha = \frac{N}{M} = \Theta(1) \to \alpha = \alpha_M = \Omega(1)$$

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on
3. Codeword length is $ML$ but error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$ (DMC)
   - Improve by increasing coverage depth scaling

$$\alpha = \frac{N}{M} = \Theta(1) \to \alpha = \alpha_M = \Omega(1)$$

4. Capacity increase due to multi-draws can be marginal

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on
3. Codeword length is $ML$ but error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$ (DMC)
   - Improve by increasing coverage depth scaling

$$\alpha = \frac{N}{M} = \Theta(1) \to \alpha = \alpha_M = \Omega(1)$$

4. Capacity increase due to multi-draws can be marginal
   - Example: Let $W$ be a BSC with crossover probability $w = 0.01$

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on
3. Codeword length is $ML$ but error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$ (DMC)
   - Improve by increasing coverage depth scaling

$$\alpha = \frac{N}{M} = \Theta(1) \rightarrow \alpha = \alpha_M = \Omega(1)$$

4. Capacity increase due to multi-draws can be marginal
   - Example: Let $W$ be a BSC with crossover probability $w = 0.01$
   - $C(W^{\oplus d}) = (0.91, 0.97, 0.99)$ for $d = (1, 2, 3)$

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on
3. Codeword length is $ML$ but error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$ (DMC)
   - Improve by increasing coverage depth scaling

$$\alpha = \frac{N}{M} = \Theta(1) \to \alpha = \alpha_M = \Omega(1)$$

4. Capacity increase due to multi-draws can be marginal
   - Example: Let $W$ be a BSC with crossover probability $w = 0.01$
   - $C(W^{\oplus d}) = (0.91, 0.97, 0.99)$ for $d = (1, 2, 3)$

# Motivation

Previously:

1. The decoder has extremely high computational complexity ("optimal-like")
   - The computation of the metric of a single codeword has complexity $\Theta(M^N)$
2. The sequencing channel $W^{(L)}$ is a DMC
   - Practical channels also include deletions, insertions and so on
3. Codeword length is $ML$ but error probability decays as $e^{-\Theta(M)}$ rather than $e^{-\Theta(ML)}$ (DMC)
   - Improve by increasing coverage depth scaling

$$\alpha = \frac{N}{M} = \Theta(1) \to \alpha = \alpha_M = \Omega(1)$$

4. Capacity increase due to multi-draws can be marginal
   - Example: Let $W$ be a BSC with crossover probability $w = 0.01$
   - $C(W^{\oplus d}) = (0.91, 0.97, 0.99)$ for $d = (1, 2, 3)$

# Encoder and decoder

- Encoder: A *coded-index* scheme

# Encoder and decoder

- Encoder: A *coded-index* scheme
  - An inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ is partitioned to $M$ sub-codes $\{\mathcal{B}_m^{(L)}\}_{m \in [M]}$

# Encoder and decoder

- Encoder: A *coded-index* scheme
  - An inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ is partitioned to $M$ sub-codes $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$
  - An outer-code encodes a message to a sequence $x^{LM}$ where $x_m^L \in \mathcal{B}_m^{(L)}$

# Encoder and decoder

- Encoder: A *coded-index* scheme
  - An inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ is partitioned to $M$ sub-codes $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$
  - An outer-code encodes a message to a sequence $x^{LM}$ where $x_m^L \in \mathcal{B}_m^{(L)}$
- Decoder:

# Encoder and decoder

- Encoder: A *coded-index* scheme
  - An inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ is partitioned to $M$ sub-codes $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$
  - An outer-code encodes a message to a sequence $x^{LM}$ where $x_m^L \in \mathcal{B}_m^{(L)}$
- Decoder:
  - Inner code decoding: Each output molecule $y_n^L$ is decoded to a codeword in $\mathcal{B}^{(L)}$

# Encoder and decoder

- Encoder: A *coded-index* scheme
  - An inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ is partitioned to $M$ sub-codes $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$
  - An outer-code encodes a message to a sequence $x^{LM}$ where $x_m^L \in \mathcal{B}_m^{(L)}$
- Decoder:
  - Inner code decoding: Each output molecule $y_n^L$ is decoded to a codeword in $\mathcal{B}^{(L)}$
  - An erasure is declared if there is no consensus on $x_m^L$

# Encoder and decoder

- Encoder: A *coded-index* scheme
  - An inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ is partitioned to $M$ sub-codes $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$
  - An outer-code encodes a message to a sequence $x^{LM}$ where $x_m^L \in \mathcal{B}_m^{(L)}$
- Decoder:
  - Inner code decoding: Each output molecule $y_n^L$ is decoded to a codeword in $\mathcal{B}^{(L)}$
  - An erasure is declared if there is no consensus on $x_m^L$
  - Note: No (substantial) gain from multi-draws

# Encoder and decoder

- Encoder: A *coded-index* scheme
  - An inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ is partitioned to $M$ sub-codes $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$
  - An outer-code encodes a message to a sequence $x^{LM}$ where $x_m^L \in \mathcal{B}_m^{(L)}$
- Decoder:
  - Inner code decoding: Each output molecule $y_n^L$ is decoded to a codeword in $\mathcal{B}^{(L)}$
  - An erasure is declared if there is no consensus on $x_m^L$
  - Note: No (substantial) gain from multi-draws
- Outer code decoding

# Assumptions on the code

- A "black-box" inner code

# Assumptions on the code

- A "black-box" inner code
  1. Inner code rate: $R_b = \frac{1}{L}\log|\mathcal{B}^{(L)}| > 1/\beta$.

# Assumptions on the code

- A "black-box" inner code
  1. Inner code rate: $R_b = \frac{1}{L} \log |\mathcal{B}^{(L)}| > 1/\beta$.
  2. Vanishing inner code error probability: $\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-\Theta(L^{\zeta})}$ where $\zeta > 0$.

# Assumptions on the code

- A "black-box" inner code
  1. Inner code rate: $R_b = \frac{1}{L} \log |\mathcal{B}^{(L)}| > 1/\beta$.
  2. Vanishing inner code error probability: $\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-\Theta(L^{\zeta})}$ where $\zeta > 0$.
- A random outer code

Theorem

# Main result – random coding and expurgated bounds

Theorem

- If $N/M = \Theta(1)$ then

$$\liminf_{M\to\infty} -\frac{1}{M}\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq d_b\left(1 - \frac{R}{R_b - 1/\beta}\,\middle\|\, e^{-\frac{N}{M}}\right)$$

for any $R < (R_b - 1/\beta)(1 - e^{-\frac{N}{M}})$.

# Main result – random coding and expurgated bounds

## Theorem

- If $N/M = \Theta(1)$ then
$$\liminf_{M\to\infty} -\frac{1}{M}\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq d_b\left(1 - \frac{R}{R_b - 1/\beta} \,\middle\|\, e^{-\frac{N}{M}}\right)$$
for any $R < (R_b - 1/\beta)(1 - e^{-\frac{N}{M}})$.
- If $N/M = \omega(1)$ then
$$\liminf_{N\to\infty} -\frac{1}{N}\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M)$$
$$\geq \begin{cases} \frac{1}{2}\left[1 - \frac{R}{R_b - 1/\beta}\right], & \frac{N}{ML} < 2(R_b - 1/\beta) \\ \frac{ML}{N}\left[R_b - 1/\beta - R\right], & 2(R_b - 1/\beta) \leq \frac{N}{ML} < 4(R_b - 1/\beta) \\ \frac{1}{4}\left[1 - \frac{R}{R_b - 1/\beta}\right], & \frac{N}{ML} > 4(R_b - 1/\beta) \end{cases}$$
for any $R < R_b - 1/\beta$.

# Main result – random coding and expurgated bounds

Theorem

- If $N/M = \Theta(1)$ then
$$\liminf_{M\to\infty} -\frac{1}{M}\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq d_b\left(1 - \frac{R}{R_b - 1/\beta}\,\middle\|\, e^{-\frac{N}{M}}\right)$$
for any $R < (R_b - 1/\beta)(1 - e^{-\frac{N}{M}})$.
- If $N/M = \omega(1)$ then
$$\liminf_{N\to\infty} -\frac{1}{N}\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M)$$
$$\geq \begin{cases} \frac{1}{2}\left[1 - \frac{R}{R_b - 1/\beta}\right], & \frac{N}{ML} < 2(R_b - 1/\beta) \\ \frac{ML}{N}\left[R_b - 1/\beta - R\right], & 2(R_b - 1/\beta) \leq \frac{N}{ML} < 4(R_b - 1/\beta) \\ \frac{1}{4}\left[1 - \frac{R}{R_b - 1/\beta}\right], & \frac{N}{ML} > 4(R_b - 1/\beta) \end{cases}$$
for any $R < R_b - 1/\beta$.
- Based on random coding and expurgated analysis

# Main result – discussion

1. A phase transition between $N = \Theta(M)$ and $N = \omega(M)$

# Main result – discussion

1. A phase transition between $N = \Theta(M)$ and $N = \omega(M)$
2. Expurgation improves in the regime $\frac{N}{ML} > 4(R_b - 1/\beta)$

# Main result – discussion

1. A phase transition between $N = \Theta(M)$ and $N = \omega(M)$
2. Expurgation improves in the regime $\frac{N}{ML} > 4(R_b - 1/\beta)$
3. A slow decrease $O(\frac{1}{\log M})$ to the asymptotic scaling

# Main result – discussion

1. A phase transition between $N = \Theta(M)$ and $N = \omega(M)$
2. Expurgation improves in the regime $\frac{N}{ML} > 4(R_b - 1/\beta)$
3. A slow decrease $O(\frac{1}{\log M})$ to the asymptotic scaling
4. Establishing tightness of the bound seems challenging

# Main result – discussion

1. A phase transition between $N = \Theta(M)$ and $N = \omega(M)$
2. Expurgation improves in the regime $\frac{N}{ML} > 4(R_b - 1/\beta)$
3. A slow decrease $O(\frac{1}{\log M})$ to the asymptotic scaling
4. Establishing tightness of the bound seems challenging
   - Poissonization is used in the proof – tight for expectations but not for tails

# Conclusion

- A simplified analysis of a DNA storage scheme

# Conclusion

- A simplified analysis of a DNA storage scheme
  - Towards practical encoding/decoding methods

# Conclusion

- A simplified analysis of a DNA storage scheme
  - Towards practical encoding/decoding methods
- Error probability decays as $e^{-\Theta(N)}$

# Conclusion

- A simplified analysis of a DNA storage scheme
  - Towards practical encoding/decoding methods
- Error probability decays as $e^{-\Theta(N)}$
  - Increasing coverage depth is vital for improving error probability scaling

# Conclusion

- A simplified analysis of a DNA storage scheme
  - Towards practical encoding/decoding methods
- Error probability decays as $e^{-\Theta(N)}$
  - Increasing coverage depth is vital for improving error probability scaling

# Conclusion

- A simplified analysis of a DNA storage scheme
  - Towards practical encoding/decoding methods
- Error probability decays as $e^{-\Theta(N)}$
  - Increasing coverage depth is vital for improving error probability scaling

N. Weinberger
"Error Probability Bounds for Coded-Index DNA Storage Systems,"
IT-T, November 2022

# Conclusion

- A simplified analysis of a DNA storage scheme
  - Towards practical encoding/decoding methods
- Error probability decays as $e^{-\Theta(N)}$
  - Increasing coverage depth is vital for improving error probability scaling

N. Weinberger
"Error Probability Bounds for Coded-Index DNA Storage Systems,"
IT-T, November 2022

Thank You !