CAPACITY OF SHOT GUN SEQUENCING CHANNEL

Aditya Narayan Ravi

Alireza Vahid

Ilan Shormony

INFORMATION AND COMMUNICATION
COURSE PROJECT

G. ANANYA VARMA
20220102064

KRIPI SINGLA
2022102063

CAPACITY OF A CHANNEL

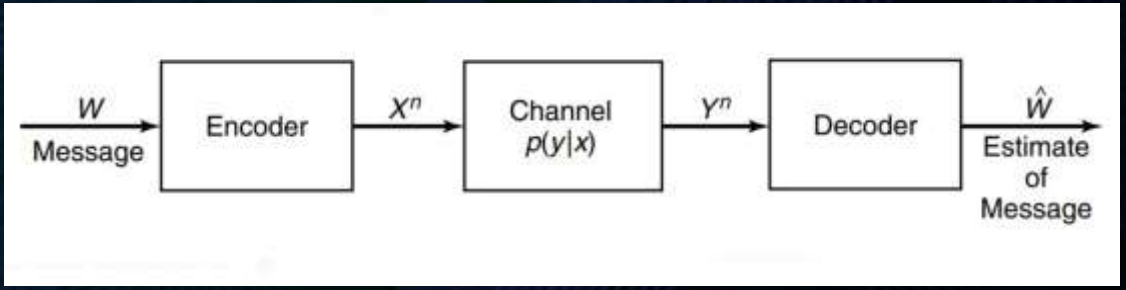
CAPACITY :

It is the rate at which the message is transmitted with lowest possible probability.

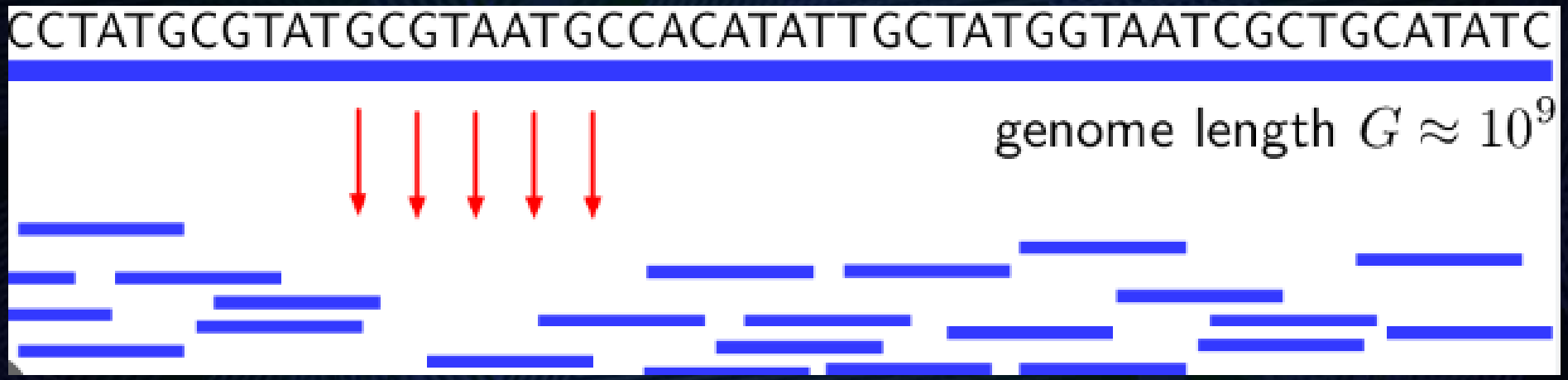
$$C = \sup_{p_X(x)} I(X;Y)$$

Where X:input message

Y:output code



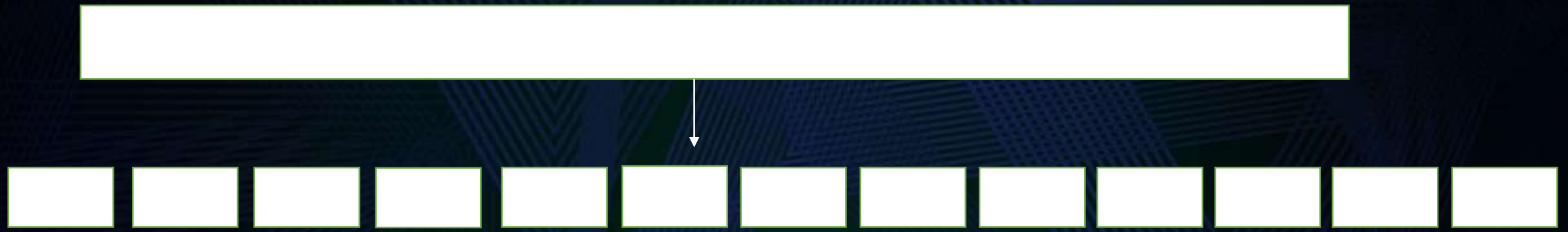
SHOTGUN SEQUENCING CHANNEL



DNA sequencing is basically the process of determining the precise order of nucleotides within a DNA molecule.

What is shotgun sequencing:

1. It includes breaking of the DNA strands into smaller fragments without any specification

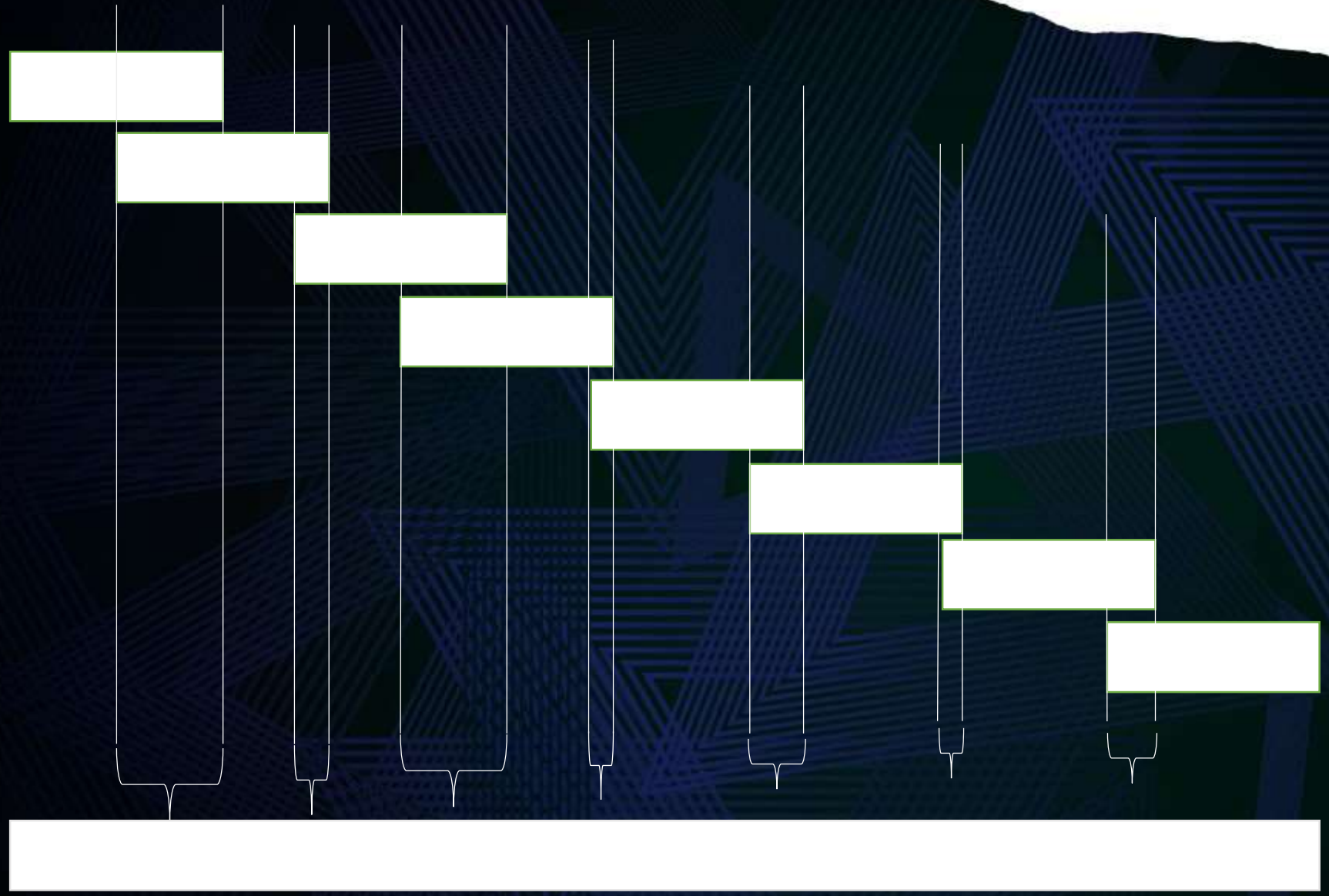


These fragments generated are known as contigs .

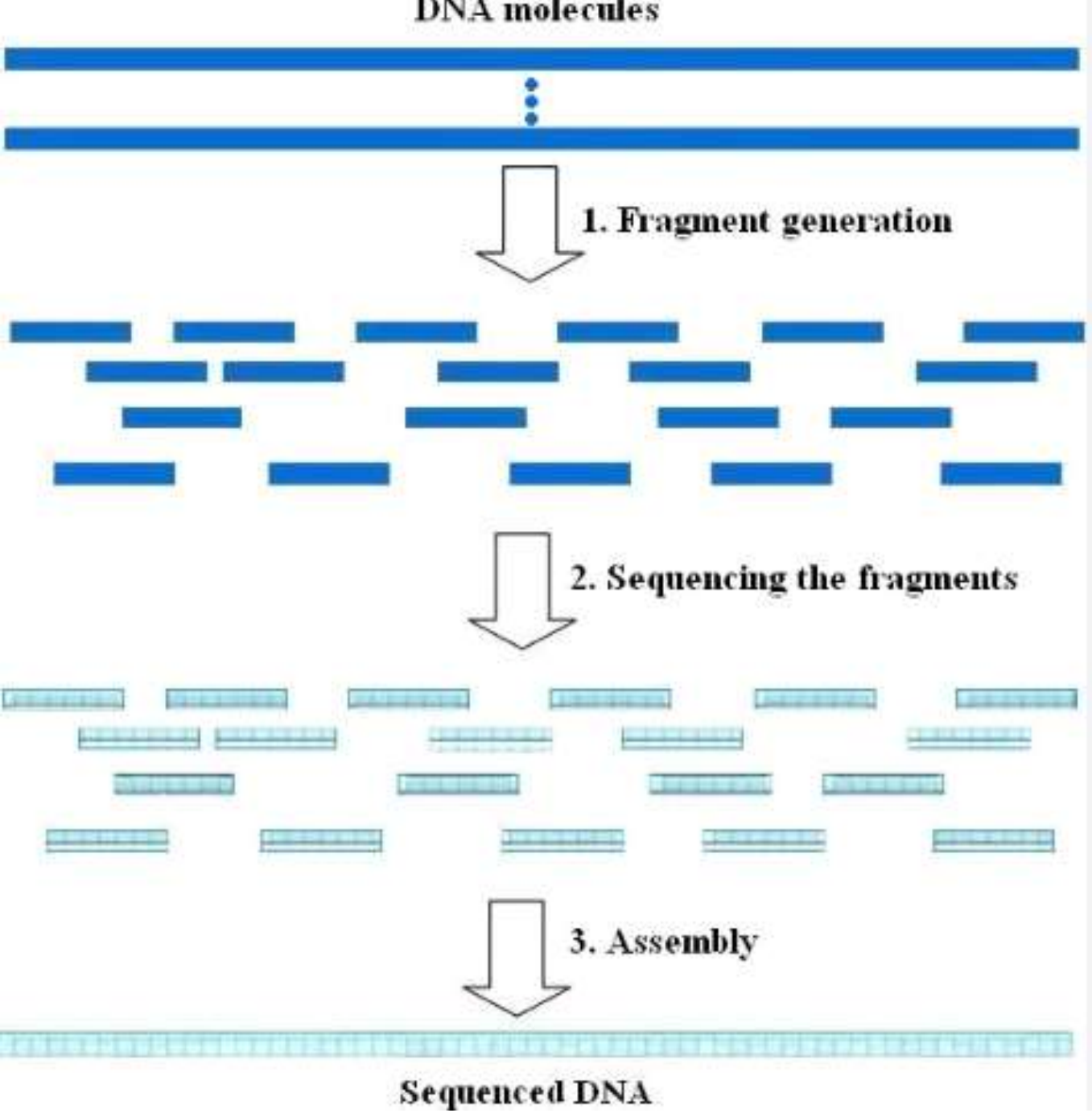
The whole process takes place int the following steps:

Extracting a large number of short reads from random locations of the target DNA sequence (e.g., the genome of an organism) in a massively parallel fashion.

Sequencing must then be followed by an assembly step, where the reads are merged based on regions of overlap with the intention of reconstructing the original DNA sequence.
Hence reliable construction takes place.



These are the overlapping regions and hence on merging we will be able to obtain the original sequence.



Shotgun sequencing is a DNA sequencing technique used to determine the order of nucleotides (A, T, C, and G) in a DNA molecule. It is named "shotgun" sequencing because it involves randomly breaking the DNA molecule into many small fragments, sequencing these fragments, and then reconstructing the original DNA sequence from the overlapping sequences.

1.Fragmentation: The DNA is randomly fragmented into smaller pieces. This can be done using physical or enzymatic methods.
2.Sequencing: Each DNA fragment is individually sequenced using a DNA sequencing technology, such as Sanger sequencing or next-generation sequencing (NGS) methods like Illumina sequencing.
3.Read Alignment: The resulting sequence reads are aligned to a reference genome or assembled based on overlapping regions to reconstruct the original DNA sequence.

Before completely getting into the topic ,

Let's get familiar with some terminology :

They are:

1. Read length
2. Coverage depth

Read length:
It is the length of the DNA fragments that are sequenced during the shotgun sequencing process

Coverage depth:
It is the measure of how many times a specific position in a genome is covered by sequenced reads.

PROBLEM STATEMENT

To figure out mathematical expressions for :

1. Read length
2. Coverage depth
3. Capacity of the channel

In a scenario where the DNA molecule is being sequenced as a code word from predefined codebook.

SHOTGUN SEQUENCING CHANNEL

DECODER

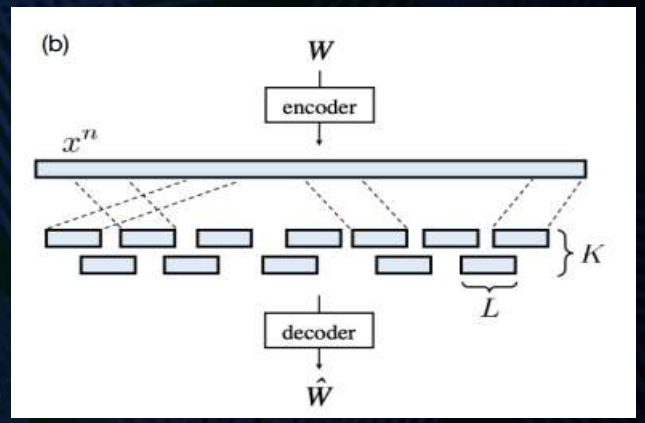
THE OUTPUT CODE

SETTING UP SHOTGUN SEQUENCING CHANNEL

Now , suppose we observe K random reads (i.e., substrings) of length L from an unknown length-n sequence x^n .
Consider the asymptotic regime where n → ∞ and the read length L scales as L = L^- log n, for a constant L^-.
They also defined c = KL/n to be the coverage depth; i.e., the average number of times each symbol in x^n is sequenced .

so ,

no of random reads : K
length of substring : L
total length of the DNA sequence: n
and coverage depth is KL/n



If x^n is an i.i.d (independent and identically distributed) Ber(1/2) sequence,

Then if L^-<2 , reconstruction is impossible for any c and if L^->2 then the reconstruction is possible .
i.e.

$$c_{LW} = \ln(n/\epsilon)$$

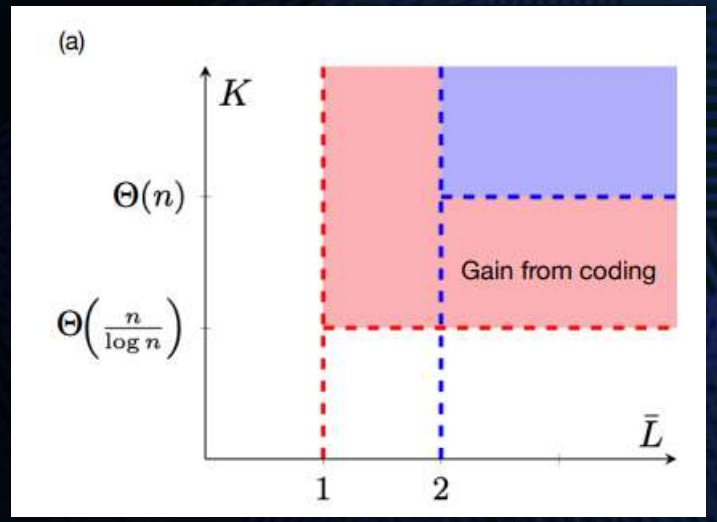
minimum coverage needed to guarantee that all symbols in x^n are sequenced at least once with probability 1- €.
This coverage length is known as Lander-Waterman Coverage.

FEASIBILITY REGION

Previously the value of the source code was chosen from the nature .

Now, with the advent of DNA based storage systems ,
x^n is now chosen from the predefined codebook.

Feasibility region helps us to study what changes occur when x^n is taken from predefined codebook.



As shown in this figure , we can observe that
Uncoded string is represented by blue
Coded string is represented by red
It is evident that the coded string results in gain and increase in the area of the feasibility region.

MATHEMATICAL FORMULATION OF SHOTGUN SEQUENCING CHANNEL

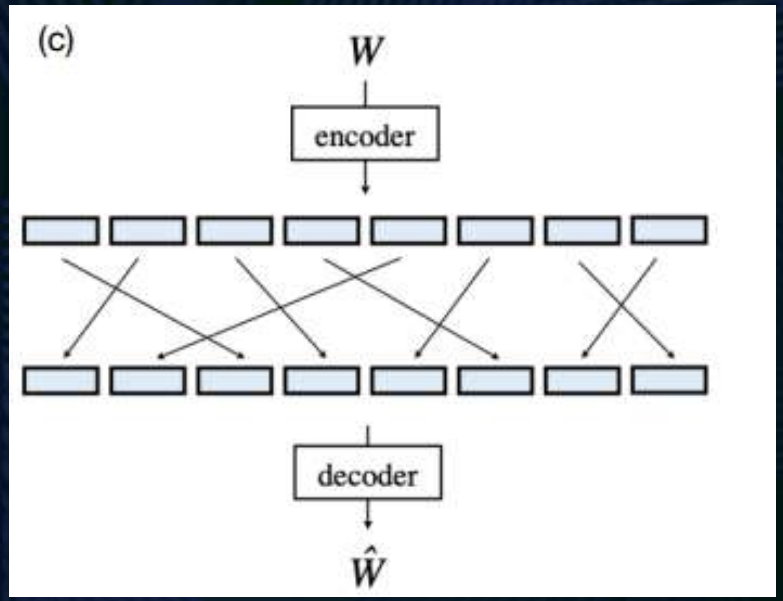
Here

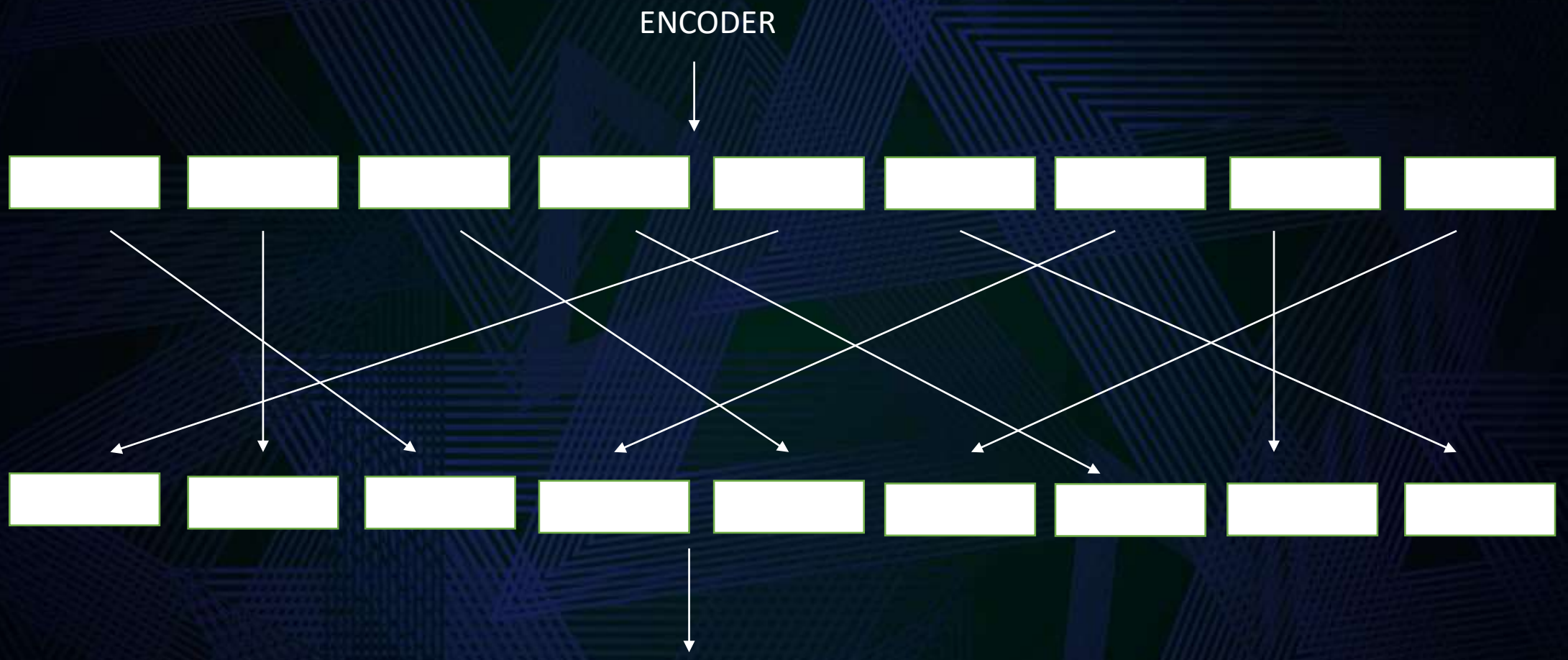
Channel input: binary length –n sequence channel x^n

Channel output: K random reads of length L from channel x^n

In order to build intuition:

Let’s have further look at the shuffling sample channel:




ENCODER
DECODER

SHUFFLING SAMPLING CHANNEL

In this case the input are M strings of length L, and the output are K strings, each chosen uniformly at random from the set of input strings. If we define the coverage depth for this setting as

$$c = KL/ML = K/M$$

which implies that, for $\bar{L} > 1$,

The capacity of this channel is :

$$C_{\text{shuf}} = \left(1 - e^{-c}\right)\left(1 - 1/\bar{L}\right),$$

For $\bar{L} < 1$, capacity becomes zero.

The term $\left(1 - e^{-c}\right)$ captures the loss due to unseen input strings and $(1 - 1/\bar{L})$ captures the loss due to the unordered nature of the output strings .

CAPACITY OF SHOT GUN SEQUENCING CHANNEL

From the setup of the shuffling sampling channel , we get the intuition that the capacity of the shot gun sequencing channel must depend on C and L^- in a similar way .

So hence the capacity of the shot gun sequencing channel for $L^- > 1$ is,

$$C_{\mathrm{SSC}} = 1 - e^{-c\left(1-\frac{1}{L}\right)}.$$

Notice that the dependence on L^- appears as the term $(1-1/L^-)$ in the exponent and, as $c \to \infty$, CSSC $\to 1$ for any $L >^- 1$. This contrasts with the shuffling-sampling channel, where $C_{shuf} \to 1 - 1/L^-$ as we increase the coverage depth c to infinity. Therefore, even in the high coverage depth regime, if $L^- \approx 1$, $C_{shuf} \approx 0$.

Hence $C_{shuf} < C_{ssc}$ for any C and L^-.

MAIN RESULTS

Uncoded case:
O(n) reads of length greater than 2log(n) are needed for reliable reconstruction.

Coded case:
Only O(n/log(n)) reads of length greater than log(n) are needed for capacity to be arbitrary close to 1 .

CONTRIBUTIONS:

READ THE RESEARCH PAPER TOGETHER AND MADE THE PRESENTATION DIVIDING THE WHOLE PAPER INTO TWO EQUAL HALVES.

THANK YOU