

**Departamento de Engenharia de Eletrónica e Telecomunicações e de Computadores**

**Comunicação Digital – Módulo 2**

|  |  |  |
| --- | --- | --- |
| Autores: | 50536 | Fábio Silva |
|  | 50553 | Bruno Raposo |

Relatório para a Unidade Curricular de Comunicação Digital da Licenciatura em Engenharia Informática e de Computadores

Docente: Artur Ferreira

03 – 05 – 2024

<< Esta página foi intencionalmente deixada em branco >

Índice

[Exercício 1 2](#_Toc167726709)

[**Códigos de controlo de erros** 2](#_Toc167726710)

[**Ausência de código de controlo de erros (Alínea i)** 3](#_Toc167726711)

[**Código de repetição (Alínea ii)** 5](#_Toc167726712)

[**Código de Hamming (Alínea iii)** 7](#_Toc167726713)

[Exercício 2 10](#_Toc167726714)

[Função de Progressão Aritmética 10](#_Toc167726715)

[Função Fatorial 11](#_Toc167726716)

[Função Mínimo Múltiplo Comum (MMC) 12](#_Toc167726717)

[Função de Números Primos 13](#_Toc167726718)

[Função de Frequência de Símbolos em Arquivos 14](#_Toc167726719)

[Exercício 3 14](#_Toc167726720)

[Análise à informação própria e à entropia e recolha do histograma 15](#_Toc167726721)

[Estimativas de ocorrências de símbolos e pares de símbolos 17](#_Toc167726722)

[Exercício 4 17](#_Toc167726723)

[Implementação Genérica de Fonte de Símbolos 18](#_Toc167726724)

[Geradores de Símbolos Específicos 19](#_Toc167726725)

[Códigos PIN 19](#_Toc167726726)

[Chaves do Euromilhões 20](#_Toc167726727)

[Palavras-passe robustas 20](#_Toc167726728)

[Compressão de Dados e Taxa de Compressão 21](#_Toc167726729)

[Exercício 5 22](#_Toc167726730)

[Sistemas Criptográficos 22](#_Toc167726731)

[Cifra de Vernam 22](#_Toc167726732)

[Realização do Exercício 23](#_Toc167726733)

[Exercício 6 24](#_Toc167726734)

[Binary Symmetric Channel (BSC) 24](#_Toc167726735)

[Outros Exemplos de Transmissão de Sequências de Bits 25](#_Toc167726736)

[Transmissões de ficheiros 26](#_Toc167726737)

[Bibliografia 27](#_Toc167726738)

# Exercício 1

Neste exercício, é proposta a **simulação da transmissão de ficheiros sobre o BSC implementado** no último exercício do módulo anterior. Adicionalmente, os testes realizados ao BSC recorrerão a **3 configurações distintas**, sendo estas:

1. Ausência de códigos de controlo de erros.
2. Código de repetição (3,1), em modo correção.
3. Código de Hamming (7,4), em modo correção.

Para cada configuração serão apresentados **4 exemplos com valores de *p* crescentes**, onde, para cada exemplo, serão apresentados os **valores BER e BER’**, tal como o **número de palavras diferentes** entre o ficheiro recebido e o ficheiro original.

**Códigos de controlo de erros**

Os códigos de controlo de erros têm como função a deteção e, por vezes, a correção de erros que possam surgir em uma mensagem resultante de uma transmissão sujeita a erros. Esta deteção e correção são obtidas através da introdução de redundância na mensagem original, isto é, na adição de bits adicionais aos bits originais da mensagem.

Os códigos de controlo de erros lecionados em aula são: o código de repetição, o código de bit paridade par, o código de Hamming e o CRC (Cyclic Redundant Check). Destes códigos apresentados, recorreremos ao código de repetição e ao código de Hamming na resolução deste exercício, ambos em modo de correção.

Todos os códigos apresentados acima, com exceção do CRC, pertencem a um conjunto de códigos de controlo de erros designado de códigos lineares de bloco. Alguns conceitos importantes sobre este conjunto são:

* Cada *bloco*, isto é, cada conjunto de bits aos quais vão ser adicionados bits redundantes, de *k* bits de mensagem origina uma palavra de código com *n* bits (*k* bits da mensagem + *q* bits redundantes);
* O vetor nulo pertence ao código;
* A soma modular de quaisquer duas palavras do código resulta em uma outra palavra do código;
* Uma propriedade destes códigos é a distância mínima (dmin) que consiste no número de bits distintos entre palavras de código diferentes;
* O limite de bits errados que a mensagem pode ter até o código não ser capaz de indicar que ocorreu um erro na transmissão pode ser determinado através da seguinte fórmula:

Uma imagem com Tipo de letra, tipografia, texto, branco

Descrição gerada automaticamente

Figura 1 – Fórmula de determinação do limite máximo de bits errados para existir deteção de erros na transmissão (l = nº de bits errados; dmin = distância mínima)

* Por fim, o limite de bits errados até o código não ser capaz de corrigir qualquer erro quando detetado pode ser determinado através da seguinte fórmula:

Uma imagem com Tipo de letra, file, diagrama, branco

Descrição gerada automaticamente

Figura 2 – Fórmula de determinação do limite máximo de bits errados para existir correção de erros na transmissão (t = nº de bits errados; dmin = distância mínima)

**Ausência de código de controlo de erros (Alínea i)**

Esta configuração do BSC tem como objetivo simular a transmissão de um ficheiro com a obra “Alice e o País das Maravilhas”, sujeita a erros, **sem recorrer a nenhum código de controlo de erros**.

Para a realização desta simulação recorreu-se a implementações realizadas no módulo anterior, nomeadamente, o BSC e a função de contagem de erros entre a mensagem recebida e a mensagem original.

A seguinte figura ilustra os resultados obtidos de 4 exemplos de simulação distintos com valores de *p* iguais a 0,001; 0,01; 0,1 e 0,5, respetivamente.

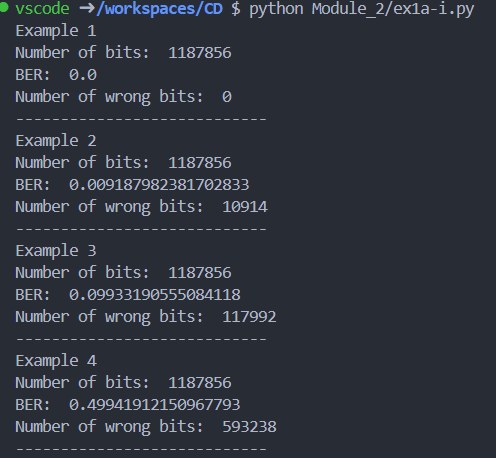


Figura 3 - Resultados da simulação sem código de controlo de erros

Uma vez que estes exemplos de simulação não recorrem a nenhum código de controlo de erros, não existe valor de BER’ a ser apresentado.

Ao observarmos a imagem, é possível compreender que quanto maior for o valor de *p*, ou seja, o BER, maior será o número de erros observados no ficheiro.

As seguintes figuras mostram um excerto dos ficheiros resultantes dos exemplos de transmissão do ficheiro com a obra “Alice e o País das Maravilhas”, pelo nosso BSC.

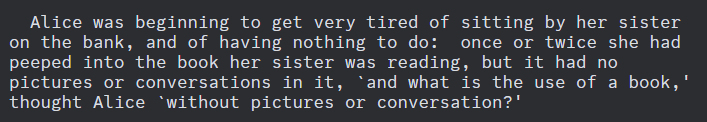


Figura 4 - Excerto do ficheiro resultante de uma transmissão ausente de controlo de erros com BER = 0,001

Uma imagem com texto, Tipo de letra, captura de ecrã, informação

Descrição gerada automaticamente

Figura 5 - Excerto do ficheiro resultante de uma transmissão ausente de controlo de erros com BER = 0,01

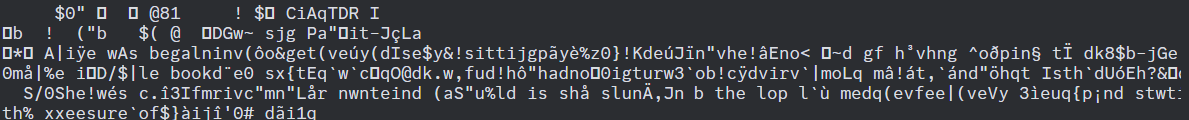


Figura 6 – Excerto do ficheiro resultante de uma transmissão ausente de controlo de erros com BER = 0,1

Uma imagem com texto, captura de ecrã, Tipo de letra, informação

Descrição gerada automaticamente

Figura 7 - Excerto do ficheiro resultante de uma transmissão ausente de controlo de erros com BER = 0,5

**Código de repetição (Alínea ii)**

O objetivo desta configuração do BSC consiste em, novamente, simular a transmissão da obra “Alice e o País das Maravilhas”, **recorrendo ao código de repetição (3,1) em modo de correção**, como forma de deteção e correção de erros durante a transmissão.

O código de repetição (3,1) consiste em um **código de controlo de erros** que, de forma a conseguir detetar uma possível troca de bits na transmissão, **adiciona a cada bit da mensagem duas repetições do mesmo**, antes desta ser transmitida. Com esta codificação, uma mensagem do tipo “0 1 1 0” resultará no código “000 111 111 000”.

Após a transmissão da mensagem codificada pelo BSC, o descodificador irá verificar, para cada **conjunto de 3 bits**, se algum destes foi modificado, indicando que houve um erro de transmissão quando **todos os bits do conjunto não são iguais**. Desta forma, este código de controlo de erros é capaz de **detetar erros de 1 e 2 bits na mensagem codificada**. Contudo, a sua **capacidade de correção de erros já se encontra limitada a apenas 1 bit**. Isto porque, após a transmissão da mensagem codificada, o descodificador vai verificar os 3 bits de cada conjunto e descodificar, os mesmos, no bit que se encontra em maioria, descodificando de forma errada, conjuntos de bits que tenham sofrido alterações em 2 ou 3 dos seus bits. Com esta descodificação, um código sujeito a erros do tipo “001 101 001 000” resultará na mensagem “0 1 0 0”.

Para a realização desta simulação recorreu-se, novamente, a implementações realizadas no módulo anterior, nomeadamente, o BSC e a função de contagem de erros entre a mensagem recebida e a mensagem original.

A seguinte figura apresenta os resultados obtidos de 4 exemplos de simulação distintos com valores de *p* iguais a 0,001; 0,01; 0,1 e 0,5, respetivamente.

Uma imagem com texto, menu, captura de ecrã, Tipo de letra

Descrição gerada automaticamenteUma imagem com texto, menu, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura 8 – Resultados da simulação com código de repetição (3,1)

Com as imagens apresentadas acima conseguimos perceber que:

* A implementação com o código de repetição irá resultar em um **menor número de erros na mensagem** recebida;
* Quanto maior o valor de *p*:
* **Maior será o número de erros na mensagem recebida**, quer a transmissão aplique o código de repetição, quer não;
* **Menor será a diferença entre o número de bits errados da transmissão com código de repetição e o número de bits errados da transmissão sem código de controlo de erros**. Isto porque, quanto maior for o BER da transmissão, maior será a probabilidade de cada bloco de bits apresentar erros em 2 ou 3 bits, impossibilitando a correção dos mesmos.

As figuras seguintes demonstram um excerto dos ficheiros resultantes dos exemplos de transmissão com código de repetição (3,1), da obra “Alice e o País das Maravilhas”, pelo nosso BSC.

Uma imagem com texto, captura de ecrã, Tipo de letra, informação

Descrição gerada automaticamente

Figura 9 - Excerto do ficheiro resultante de uma transmissão com código de repetição (3,1) e BER = 0,001

Uma imagem com texto, Tipo de letra, captura de ecrã, informação

Descrição gerada automaticamente

Figura 10 - Excerto do ficheiro resultante de uma transmissão com código de repetição (3,1) e BER = 0,01

**Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente**

Figura 11 - Excerto do ficheiro resultante de uma transmissão com código de repetição (3,1) e BER = 0,1

**Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente**

Figura 12 - Excerto do ficheiro resultante de uma transmissão com código de repetição (3,1) e BER = 0,5

**Código de Hamming (Alínea iii)**

Nesta configuração do BSC pretende-se simular, de novo, a transmissão da obra “Alice e o País das Maravilhas”, **recorrendo ao código de Hamming (7,4) em modo de correção**, como forma de deteção e correção de erros durante a transmissão.

O código de Hamming (7,4) consiste em um outro **código de controlo de erros** que, de forma a conseguir detetar erros na transmissão, **realiza 3 equações de forma a gerar os 3 bits de paridade do código**, antes desta ser transmitida. Estas equações são:

Uma imagem com Tipo de letra, texto, branco, tipografia

Descrição gerada automaticamente

Figura 13 - Equações de paridade

Uma vez que o cálculo dos bits de paridade é fixo, este código de controlo de erros apresenta apenas 16 palavras de código possíveis:

* 0000 000
* 0001 111
* 0010 101
* 0011 010
* 0100 110
* 0101 001
* 0110 011
* 0111 100
* 1000 011
* 1001 100
* 1010 110
* 1011 001
* 1100 101
* 1101 010
* 1110 000
* 1111 111

O código de Hamming é **capaz de detetar erros até 2 bits e corrigir erros de 1 bit**, isto porque **todos os seus códigos apresentam uma distância mínima igual a 3**.

Após a transmissão da mensagem codificada pelo BSC, o descodificador vai, para **cada** **conjunto de 7 bits**, aplicar, novamente, as equações de paridade para os bits da mensagem, comparando estes com os bits de paridade da mensagem resultante da transmissão. Caso esta comparação conclua que os bits são diferentes, então ocorreu um erro na transmissão da mensagem.

De forma a corrigir quaisquer erros de 1 bit na mensagem recebida, o descodificador vai recorrer a uma tabela de síndromas com o objetivo de detetar o bit errado e corrigir o mesmo. Esta tabela de síndromas, específica para este código de controlo de erros, consiste em uma tabela que relaciona uma codificação de 3 bits (a síndroma) com o número do bit da mensagem que se encontra errado, sendo que se a codificação for igual ao vetor nulo, então a mensagem não apresenta nenhum erro. A síndroma poderá ser determinada através da seguinte operação:

**(e0 XOR r0) +(e1 XOR r1) + (e2 XOR r2)**,

sendo ***e*** os bits resultantes da aplicação das equações de paridade nos bits de mensagem do código recebido e ***r*** os bits de paridade do código recebido.

A tabela de síndromas mencionada é ilustrada na seguinte figura:

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamente

Figura 14 – Tabela de síndromas específica para o código de Hamming (7,4)

Para a realização desta simulação recorreu-se, novamente, a implementações realizadas no módulo anterior, nomeadamente, o BSC e a função de contagem de erros entre a mensagem recebida e a mensagem original.

A seguinte figura apresenta os resultados obtidos de 4 exemplos de simulação distintos com valores de *p* iguais a 0,001; 0,01; 0,1 e 0,5, respetivamente.

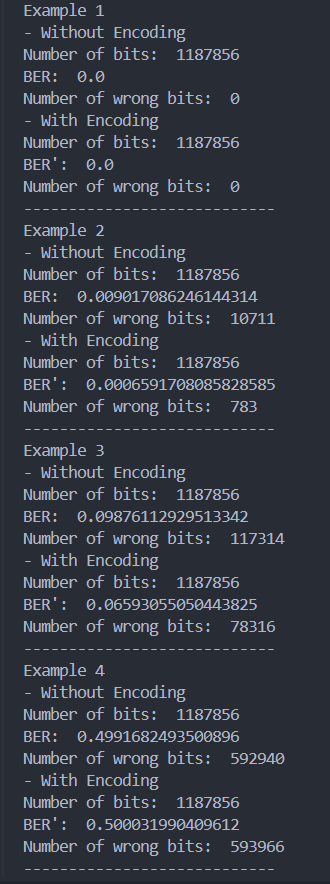
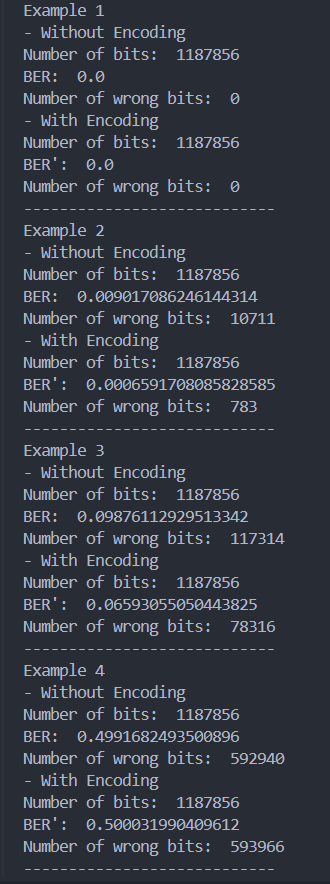


Figura 15 - Resultados da simulação com código de Hamming (7,4)

Com as imagens apresentadas acima conseguimos perceber que:

* A implementação com o código de Hamming irá resultar em um **menor número de erros na mensagem** recebida;
* Quanto maior o valor de *p*:
* **Maior será o número de erros na mensagem recebida**, quer a transmissão aplique o código de Hamming, quer não;
* **Menor será a diferença entre o número de bits errados da transmissão com código de Hamming e o número de bits errados da transmissão sem código de controlo de erros**. Isto porque, quanto maior for o BER da transmissão, maior será a probabilidade de cada bloco de bits apresentar erros em mais do que 1 bit, impossibilitando a correção dos mesmos.

As figuras seguintes demonstram um excerto dos ficheiros resultantes dos exemplos de transmissão com código de Hamming (7,4), da obra “Alice e o País das Maravilhas”, pelo nosso BSC.

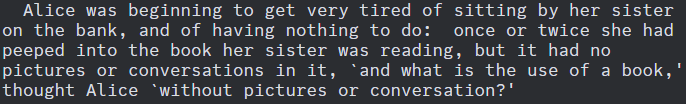


Figura 16 - Excerto do ficheiro resultante de uma transmissão com código de Hamming (7,4) e BER = 0,001

Uma imagem com texto, Tipo de letra, captura de ecrã, informação

Descrição gerada automaticamente

Figura 17 - Excerto do ficheiro resultante de uma transmissão com código de Hamming (7,4) e BER = 0,01

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura 18 - Excerto do ficheiro resultante de uma transmissão com código de Hamming (7,4) e BER = 0,1

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura 19 - Excerto do ficheiro resultante de uma transmissão com código de Hamming (7,4) e BER = 0,5

# Exercício 2

Neste exercício, exploraremos os conceitos de **erros em rajada** e **deteção de erros com Cyclic Redundancy Check (CRC)**. Primeiro, abordaremos como os erros em rajada afetam a integridade dos dados transmitidos. Em seguida, discutiremos o uso do CRC para detetar erros em blocos de dados.

Erros em rajada (Burst) – Alínea (a)

Erros em rajada ocorrem quando vários bits consecutivos em uma unidade de dados são corrompidos durante a transmissão. Os **erros em rajada** são frequentes em sistemas de comunicação e armazenamento. Diferentemente dos erros de bit individuais, que afetam apenas um bit, os erros em rajada podem impactar sequências maiores de bits. Esses erros podem ser causados por interferências no sinal, ruídos impulsivos ou outras perturbações. Exemplos de situações reais:

1. **QR Codes:**

* Os **QR codes** são amplamente utilizados para armazenar informações, como URLs ou dados de contato.
* Durante a leitura de QR codes danificados ou mal impressos, erros em rajada são uma possibilidade.

1. **Transmissão de Vídeo Digital:**

* Em sistemas de transmissão de vídeo digital, como a TV digital, os erros em rajada podem ocorrer devido a interferências no sinal, provocando por vezes distorção parcial do vídeo que está a ser transmitido.

Dentro deste contexto, existem vários códigos projetados para deteção e correção. No exercício 2, focarmo-nos num deles: o **Cyclic Redundancy Check (CRC)**. Discutiremos mais à frente como o CRC é utilizado e de que forma é capaz de detetar os erros em rajada.

Na alínea (**a**), desenvolvemos uma função em **Python** que recebe uma sequência de bits, um valor (**L**) que representa o tamanho dos erros em rajada e um valor (**p**) que corresponde à probabilidade de inserção de erros em rajada na sequência inicial.

No diagrama de blocos representado pela **Figura 20**, apresentamos os blocos pelos quais o arquivo de texto passará. A única diferença em relação ao exercício 1 é que substituiremos o **canal BSC** por um **canal Burst**, onde os erros em rajada serão introduzidos.

Uma imagem com texto, diagrama, Tipo de letra, file

Descrição gerada automaticamente

Figura 20 - Diagrama de blocos do canal

Para ilustrar a passagem por este canal de Burst, geramos quatro arquivos diferentes com base no mesmo texto. Em seguida, analisaremos como diferentes valores de (**L**) e (**p**) influenciam o conteúdo desses arquivos.

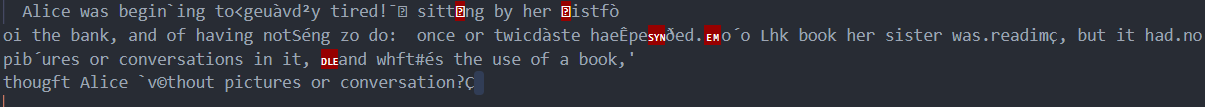


Figura 21 - Ficheiro gerado pela passagem pelo Burst com valores L=3 e p=0.01

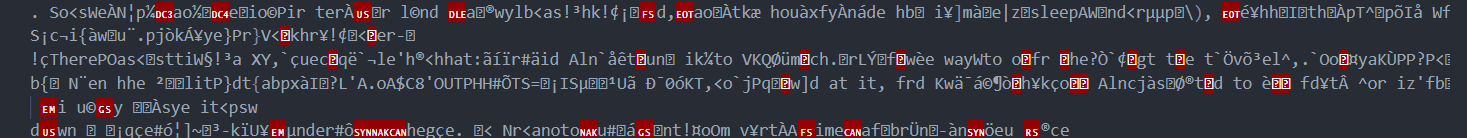


Figura 22 - Ficheiro gerado pela passagem pelo Burst com valores L=3 e p=0.1

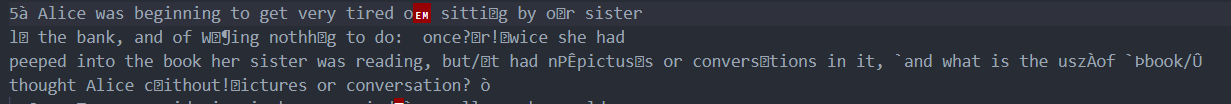


Figura 23 - Ficheiro gerado pela passagem pelo Burst com valores L=8 e p=0.01

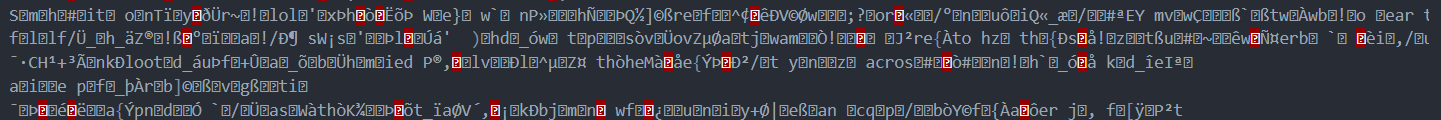


Figura 24 - Ficheiro gerado pela passagem pelo Burst com valores L=8 e p=0.1

É notável o impacto que os **erros em rajada** têm nos arquivos gerados. Mesmo com valores mais baixos de (**L**) e (**p**), o conteúdo inicial do texto foi significativamente afetado, como evidenciado na **Figura 21**. Além disso, na **Figura 24**, a leitura do texto ficou completamente ilegível.

É importante ressaltar que, no mundo real, a probabilidade de ocorrência de erros em rajada não é tão frequente quanto a que utilizámos para fins de teste. No entanto, estes exemplos permitem-nos compreender como erros desta dimensão influenciam as transmissões de dados.

## Deteção de Erros em Rajada com CRC – Alínea (b)

Na alínea anterior, implementámos o modelo de canal que introduz **erros em rajada** com dimensão (**L**). Agora, vamos aprofundar como o **Cyclic Redundancy Check (CRC)** pode ser aplicado a esse modelo para detetar esses erros.

O **Cyclic Redundancy Check (CRC)** é uma técnica amplamente utilizada para deteção de erros em transmissões de dados. Os principais conceitos relacionados ao CRC são: [1]

1. **Síndroma (s(X))**: [1]
   * O descodificador, em modo de deteção, calcula o síndroma (s(X)).
   * Dado que (c(X) = m(X)g(X)), qualquer palavra de código é um fator do polinómio gerador.
2. **Palavra Recebida (y(X))**: [1]
   * Seja (y(X) = c(X) + e(X)), onde (e(X)) é o padrão de erro.
   * Se (e(X)) for nulo, o síndroma também será nulo.
   * Caso contrário, o síndroma será não nulo e dependerá do valor de (e(X)).
3. **Divisão de Polinómios**: [1]
   * Na descodificação, realizamos a divisão de polinómios.
   * O síndroma é calculado para verificar se há erros.
   * Se o síndroma for nulo, não há erros detetados.
   * Caso contrário, erros são identificados e dependem do valor de (e(X)).
4. **Capacidade de Deteção de Erros**: [1]
   * O CRC tem alta capacidade de deteção de erros, especialmente de **burst de erros** (rajadas de erros).
   * Identifica erros em sequências contíguas de bits, mesmo quando esses erros ocorrem em blocos maiores.
5. **Limitações**: [1]
   * O CRC não corrige erros; ele apenas os deteta.
   * Se o síndroma for não nulo, o recetor sabe que ocorreram erros, mas não sabe exatamente quais bits estão errados.

No âmbito da alínea (**b**), desenvolvemos duas funções para deteção de erros em rajada usando o CRC. Uma função para **codificação** e outra para **decodificação**. O objetivo era calcular os bits de paridade da mensagem transmitida usando o polinômio gerador **CRC16**. Na decodificação, separamos a mensagem dos bits de paridade, recalculamos os bits de paridade e os comparamos com os bits de paridade recebidos para verificar se eram iguais e se não houve troca de bits durante a transmissão. Utilizamos a biblioteca **Python** chamada **CRCCHECK** para realizar o cálculo do CRC.

Para testar a deteção de erros usando o CRC, geramos três sequências de 1024 bits. Passámos cada sequência pelo canal de Burst conforme o diagrama de blocos da Figura 20. Em seguida, comparámos os bits de paridade recebidos com os bits de paridade calculados durante a decodificação. Vale realçar que utilizámos uma taxa de erro de bit (BER) de 0,002 e um valor de L igual a 16 o que nos permitiu atingir os resultados representados nas figuras abaixo.

Uma imagem com texto, captura de ecrã, Tipo de letra, informação

Descrição gerada automaticamente

Figura 25 - Exemplo 1 da passagem pelo canal de burst e deteção de erros feito pelo CRC

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura 26 - Exemplo 2 da passagem pelo canal de burst e deteção de erros feito pelo CRC

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura 27 - Exemplo 3 da passagem pelo canal de burst e deteção de erros feito pelo CRC

## Identificação de Erros Indetetáveis pelo CRC – Alínea (c)

Conforme mencionado anterior, uma das características notáveis do **Cyclic Redundancy Check (CRC)** é sua habilidade de deteção de **rajadas de erros**. No entanto, existem 3 situações onde não é possível detetar erros, se o padrão de erro for múltiplo do polinómio gerador: [2]

1. **Situação 1**:
   * Para rajadas de erros com comprimento menor ou igual ao grau ((n - k)) do polinómio gerador (g(x)), não podem corresponder aos múltiplos do polinómio gerador e todos os padrões de erros são detetados. [2]
2. **Situação 2**:
   * Para rajadas de erros com comprimento ((n - k + 1)), existe apenas um padrão de erro múltiplo de (g(x)) que coincide exatamente com (g(x)).
   * A relação de rajadas de erros não detetadas é inversamente proporcional ao número de padrões de erro possíveis. [2]
3. **Situação 3**:
   * Para rajadas de erros com comprimento ((n - k + 2)), o número de padrões de erro diferentes aumenta, mas também aumenta o número de múltiplos de (g(x)).
   * A relação de rajadas de erros não detetadas permanece constante. [2]

Para demonstrar a ocorrência de um erro em rajada que não é detetado, iniciámos o processo por consultar a documentação da biblioteca CRCCHECK. O objetivo era identificar o polinómio gerador utilizado pelo CRC16 desta biblioteca. Identificámos que polinómio gerador era .

Com o polinómio gerador(**4129**) identificado, procedemos à seleção de uma sequência de bits que fosse múltipla do polinómio e escolhemos “00010000001000010”(**8158**). Para provar que o erro não seria detetado, introduzimos bits errados na mensagem recebida para que também fosse um múltiplo do polinómio gerador. Assim, gerámos “000100000010000100000”(**33032**).

Confirmámos o que era o objetivo e o resultado que obtivemos está visível na figura 28. Podemos ver que foram gerados os mesmos bits de paridade para ambas as mensagens, o que corresponde exatamente ao que era esperado. Por este motivo, o CRC não irá detetar que houve diferença nas mensagens enviadas e recebidas. Este exemplo demonstra uma das situações em que um erro em rajada não é detetado pelo CRC.

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Figura 28 - Exemplo da não deteção de erros pelo CRC

### Exercício 3

Neste exercício, é proposta a implementação de funções capazes de analisar fontes de símbolos. A análise realizada deverá retornar o valor da informação própria de cada símbolo da fonte, o valor da entropia, o histograma e estimativas de ocorrências de símbolos e pares de símbolos.

## Análise à informação própria e à entropia e recolha do histograma

Para a realização das funções responsáveis pela recolha destes dados, será necessário explorar alguns conceitos básicos antes.

A cada ocorrência de um símbolo está associada a **informação própria** ou auto-informação. Esta é definada por:

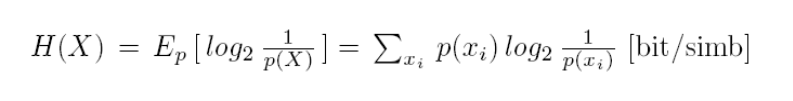
Uma imagem com Tipo de letra, tipografia, escrita à mão, caligrafia

Descrição gerada automaticamente

Sendo ***p(xi)*** correspondente à probabilidade de o símbolo ocorrer no ficheiro, que pode ser determinada por: nº de ocorrências do símbolo / nº de símbolos no ficheiro.

A **informação própria** de um símboloé inversamente proporcional à probabilidade de ocorrência do mesmo, isto é, quanto maior for ***p(xi)*** menor será ***I(xi)***.

A **entropia** de um certo ficheiro consiste no **valor esperado (médio)** da informação própria de cada símbolo, sendo determinada através do seguinte cálculo:



O valor desta, deverá ser **sempre** maior ou igual que 0 e menor ou igual que *log2(M)*, sendo *M* o número de símbolos presentes no ficheiro. No caso do valor da entropia ser zero, significa que a ocorrência dos símbolos é o mais previsível possível. Um bom exemplo será um ficheiro que apresenta apenas um único símbolo. No caso da entropia ser igual a *log2(M)*, significa que a ocorrência dos símbolos é o mais imprevisível possível. Este caso é alcançado quando todos os símbolos apresentam uma probabilidade de ocorrência e, consequentemente, um valor de informação própria iguais, entre si.

Por fim, tal como mencionado no exercício 1 deste módulo, um histograma de símbolos consiste numa representação gráfica que demonstra a frequência de ocorrência de todos os símbolos presentes em um ficheiro.

Para concretizar esta análise, foram criadas funções para determinar a probabilidade de ocorrência de cada símbolo do ficheiro, determinar a informação própria de cada símbolo com base na sua probabilidade de ocorrência previamente determinada, determinar o valor da entropia do ficheiro em análise recorrendo aos dois atributos previamente determinados e, por fim, estabelecer o histograma dos símbolos do ficheiro.

As seguintes figuras ilustram os valores dos atributos, mencionados, de uma pequena porção de símbolos do ficheiro fibonacci.kt.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura 8 - Porção do histograma de símbolos do ficheiro fibonacci.kt

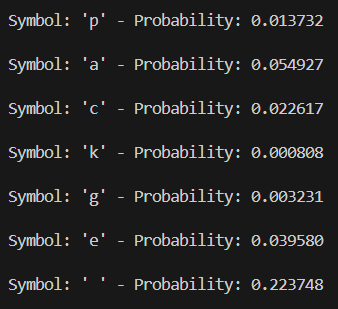
****

Figura 9 - Probabilidades de ocorrência de uma porção de símbolos do ficheiro fibonacci.kt

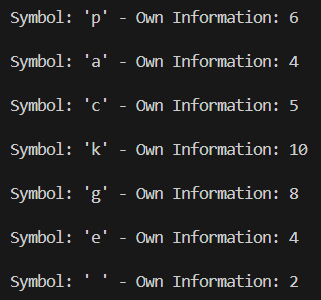
****

Figura 10 - Informação própria de uma porção de símbolos do ficheiro fibonacci.kt

****

Figura 11 - Valor da entropia do ficheiro fibonacci.kt

## Estimativas de ocorrências de símbolos e pares de símbolos

De forma a determinar, em percentagem, a estimativa de ocorrência dos símbolos ou pares de símbolos de um ficheiro, é essencial obter dois histogramas. Um histograma de símbolos e um histograma de pares de símbolos. Com estes, é realizada uma iteração por cada valor seu, tornando o mesmo em percentagem, recolhendo, por fim, os cinco símbolos e pares de símbolos com maior percentagem de ocorrência.

As **Figuras 17 e 18** ilustram os cinco símbolos e pares de símbolos com maior percentagem de ocorrência no ficheiro **ListaPalavrasPT.txt**. Dado que estamos a lidar com um ficheiro extenso e as funções implementadas terão de percorrer o mesmo ficheiro duas vezes, é expectável que os resultados demorem algum tempo a ser apresentados.

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Figura 12 - Top 5 símbolos com maior percentagem de ocorrência em ListaPalavrasPT.txt

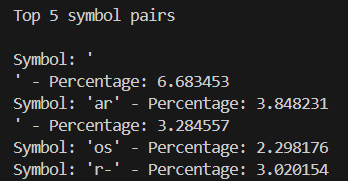


Figura 13 - Top 5 pares de símbolos com maior percentagem de ocorrência em ListaPalavrasPT.txt

# Bibliografia

|  |  |
| --- | --- |
| [1] | A. Ferreira, *10. Codificação de Canal (Códigos de Controlo de Erros).* |
| [2] | C. M. Ribeiro, Sistemas de Comunicação Digital, Lisboa: Instituto Politécnico de Lisboa, 2023. |