Resolução explicada dos exercícios 6 e 7 da folha 1 (tratados nas aulas PL dos dias 27, 28, 29 e 30 de outubro

exercício 6.a) O seguinte código faz o que é pedido

```
% exercício 6.a da folha 1
k=1;
while 1+2^-k>1
    k=k+1;
end
k=k-1
```

Guardado num ficheiro executável do Matlab, por exemplo **epsilon.m**, tem-se

```
>> epsilon

k =

    52
```

<u>Explicação:</u> no formato duplo da norma IEEE 754, a representação normalizada de um número é a seguinte

$$\pm (1.b_{-1}b_{-2}\cdots b_{-52})_2 \times 2^e$$

onde $b_i = 0$ ou $b_i = 1$, para cada $i = 1, \cdots, 52$, e $-1022 \leq e \leq 1023$. Denotamos por $\mathcal{F}$ o conjunto destes números. Os números 1 e 2^{-52} têm as representações normalizadas (só diferem nos expoentes)

$$+ (1.00\cdots 00)_2 \times 2^0$$

e

$$+ (1.00\cdots 00)_2 \times 2^{-52}$$

Para efeitos da adição, o número de menor expoente, isto é, o número 2^{-52}, terá de ser desnormalizado por forma a ficar com o mesmo expoente, neste caso 0. A representação obtida neste processo é então

$$+ (0.00\cdots 01)_2 \times 2^0$$

resultando que a soma $1 + 2^{-52}$ pertence a $\mathcal{F}$ uma vez que tem a representação

$$+ (1.00\cdots 01)_2 \times 2^0$$

Portanto, no Matlab, a execução de

```
>> 1+2^-52
```

produz o valor lógico 1. Já o mesmo não acontece com k=53, isto é,

```
>> 1+2^-53
```

produz o valor lógico 0. Porquê? A representação normalizada de 2^-53 é

$$+(1.00\cdots 00)_2 \times 2^{-53}$$

que, para efeitos da soma com 1, terá de ser desnormalizada para

$$+(0.00\cdots 00|1)_2 \times 2^0$$

O bit 1 está agora na posição 53 à direita do ponto, isto é, não "cabe na caixa"dos 52 bits reservados para a mantissa no formato duplo da norma IEEE 754. Por outras palavras, o número $1+2^{-53}$ não pertence a $\mathcal{F}$ e terá de ser arredondado. Do que se disse até agora, deverá estar claro que $1+2^{-52}$ é o sucessor de 1 em $\mathcal{F}$, portanto $1+2^{-53}$ será arredondado para 1 ou para $1+2^{-52}$. O arredondamento usual no Matlab (isto é, aquele que é implementado pelo sistema se o utilizador não o alterar) é o arredondamento "para o mais próximo". Mas $1+2^{-53}$ está à mesma distância, igual a 2^{-53}, de 1 e de $1+2^{-52}$ e por esta razão terá de ser usada a "regra de desempate"implementada na norma IEEE. Esta regra determina que o arredondamento é feito para o número que tem o bit na última posição igual a 0, que neste caso é o número 1. Confirmando no Matlab

```
>> 1+2^-53==1

ans =

     1
```

exercício 6.b)

```
>> 2^-52==eps

ans =

     1
```

Explicação: no Matlab, **eps** (abreviatura de epsilon) é a constante 2^-52 que é valor de um bit igual a 1 na última posição da mantissa, no formato duplo da norma IEEE 754. É também a distância entre os números de $\mathcal{F}$ que têm expoente zero e os respetivos sucessores. Mas a importância desta constante resulta do facto de se ter, qualquer que seja x não inferior a 2^{-1022},

$$\left|\frac{x - fl(x)}{x}\right| < eps$$

isto é, o erro relativo devido ao arredondamento é inferior a eps. No caso do arredondamento para o mais próximo, podemos melhorar o majorante deste erro e escrever

$$\left|\frac{x - fl(x)}{x}\right| \leq \frac{eps}{2}.$$

exercício 7) Para x entre 15 e o respetivo sucessor tem-se

$$|x - fl(x)| \leq 2^{-50}$$

Explicação: uma vez que

$$15 = 2^3 + 2^2 + 2^1 + 2^0$$

tem a representação

$$+\,(1.1110 \cdots 00)_2 \times 2^3$$

e o seu sucessor tem a representação (adicione-se uma unidade no último bit da mantissa)

$$+\,(1.1110 \cdots 01)_2 \times 2^3$$

que é o número $15 + 2^{-52} * 2^3$ ou seja, $15 + 2^{-49}$. No caso do arredondamento para o mais próximo, o erro absoluto $|x - fl(x)|$ não é superior a metade da amplitude 2^{-49} do intervalo $[15, 15 + 2^{-49}]$ e será igual a 2^{-50} se x for o ponto médio daquele intervalo.