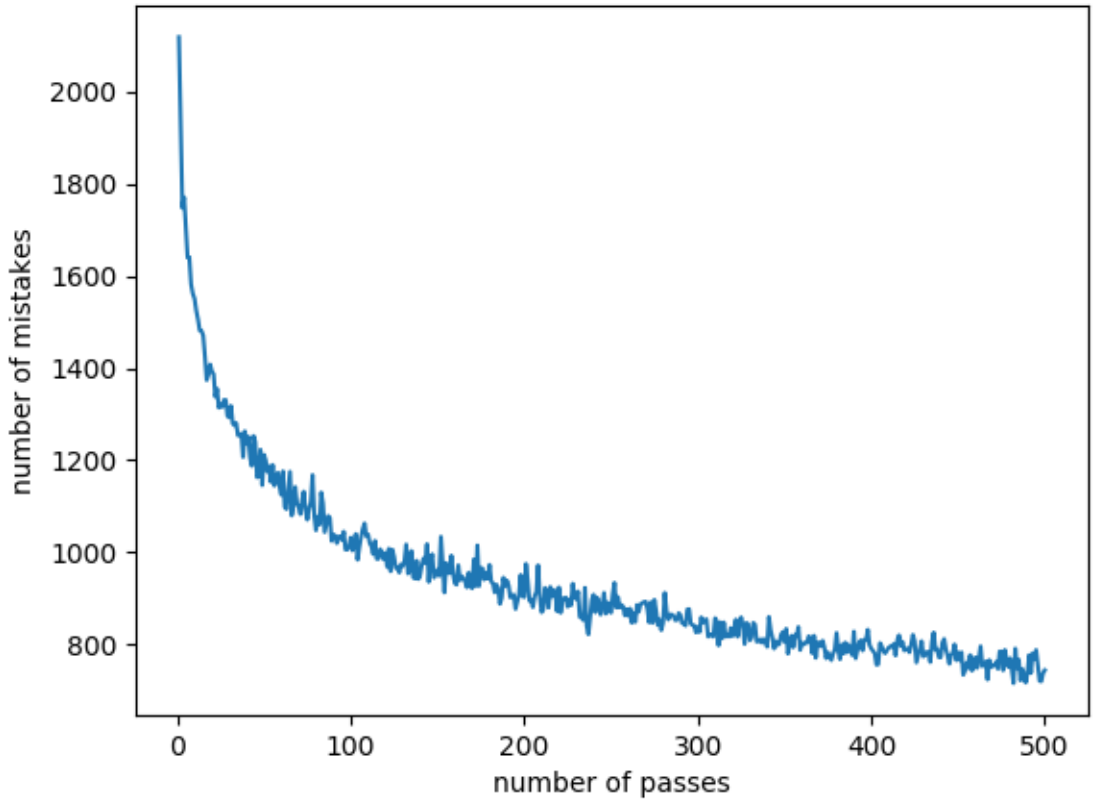
Exercise1: Perceptron Implementation

Ans: The implementation is shown in **"code_part.ipynb"**. The plot is shown as the following:



Exercise2: Linear Regression Recall that ridge regression refers to the following ℓ_2 norm regularized linear regression problem:

$$\min_{\mathbf{w}} \ \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2, \tag{1}$$

where $X \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^d$. Obviously, setting $\lambda = 0$ we recover ordinary linear regression.

1. (3 pts) Prove that ridge regression with any $\lambda > 0$ is equivalent to ordinary linear regression after performing the following data augmentation:

$$X \leftarrow \begin{bmatrix} X \\ \sqrt{2\lambda} I_{d\times d} \end{bmatrix} \tag{2}$$

$$\mathbf{y} \leftarrow \begin{bmatrix} \mathbf{y} \\ 0_d \end{bmatrix}, \tag{3}$$

where $I_{d\times d}$ is the $d \times d$ identity matrix and 0_d is the d-dimensional zero vector.

Ans: To perform the data augmentation on the ordinary linear regression, we need to substitute equation (2) & (3) into ordinary linear regression ($\min_{\mathbf{w}} \ \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2$):

$$\begin{aligned}
\min_{\mathbf{w}} \ \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2 &= \min_{\mathbf{w}} \ \frac{1}{2}\left\|\begin{bmatrix} X \\ \sqrt{2\lambda} I_{d\times d} \end{bmatrix}\mathbf{w} - \begin{bmatrix} \mathbf{y} \\ 0_d \end{bmatrix}\right\|_2^2 \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\left\|\begin{bmatrix} X\mathbf{w} \\ \sqrt{2\lambda} I_{d\times d}\mathbf{w} \end{bmatrix}_{(n+d)\times 1} - \begin{bmatrix} \mathbf{y} \\ 0_d \end{bmatrix}_{(n+d)\times 1}\right\|_2^2 \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\left\|\begin{bmatrix} X\mathbf{w}-\mathbf{y} \\ \sqrt{2\lambda} I_{d\times d}\mathbf{w} - 0_d \end{bmatrix}_{(n+d)\times 1}\right\|_2^2 \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\left\|\begin{bmatrix} X\mathbf{w}-\mathbf{y} \\ \sqrt{2\lambda} I_{d\times d}\mathbf{w} \end{bmatrix}_{(n+d)\times 1}\right\|_2^2 \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\left\|\begin{bmatrix} X\mathbf{w}-\mathbf{y} \\ \sqrt{2\lambda} I_{d\times d}\mathbf{w} \end{bmatrix}_{(n+d)\times 1}\right\|_2^2 \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\left(\|X\mathbf{w}-\mathbf{y}\|_2^2 + \left\|\sqrt{2\lambda} I_{d\times d}\mathbf{w}\right\|_2^2\right) \text{ by definition of norm} \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2 + \frac{1}{2}\left\|\sqrt{2\lambda} I_{d\times d}\mathbf{w}\right\|_2^2 \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2 + \frac{1}{2}\times 2\lambda\|I_{d\times d}\mathbf{w}\|_2^2 \\
&= \min_{\mathbf{w}} \ \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2
\end{aligned}$$

with $\lambda > 0$, which is ridge regression.
Therefore, ridge regression with any $\lambda > 0$ is equivalent to ordinary linear regression after performing the data augmentation.

2. (2 pts) Explain the data augmentation step (2)-(3). [E.g., what kind of data are we adding to the training set? how many of them? their dimension? what effect are they bringing to the weight vector $\mathbf{w}$?]

   Ans: Since $X \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^n$,

$$X \leftarrow \begin{bmatrix} X \\ \sqrt{2\lambda} I_{d\times d} \end{bmatrix}_{(n+d)\times d} \tag{4}$$

$$\mathbf{y} \leftarrow \begin{bmatrix} \mathbf{y} \\ 0_d \end{bmatrix}_{(n+d)\times 1} \tag{5}$$

   •The data we are adding to the dataset is like:
   $X_{n+1} = [\sqrt{2\lambda}\ 0\ ...\ 0]_{1\times d}$ with $y_{n+1} = 0$
   $X_{n+2} = [0\ \sqrt{2\lambda}\ ...\ 0]_{1\times d}$ with $y_{n+2} = 0$
   $X_{n+3} = [0\ 0\ \sqrt{2\lambda}\ ...\ 0]_{1\times d}$ with $y_{n+3} = 0$
   .....
   $X_{n+d} = [0\ 0\ 0...\ \sqrt{2\lambda}]_{1\times d}$ with $y_{n+d} = 0$
   •Number of data added = d
   •Dimension of data added = d
   •After adding these data, the norm of weight vector w will be reduced. In other words, the effect after adding these data is that the norm of weight vector w will not be too large since loss is added, and w need to be minimized.