Supplementary materials

Section 1 Hessian matrix form

The hessian matrix used has the following form:

$$H_{jk} = \nabla^2_{y^{(j)}y^{(k)}} f(\boldsymbol{y}) = \begin{cases} 2\left[\left(\sum_{i=1}^{l} w_{ij} + \sum_{i=1}^{u} \bar{w}_{ij}\right) - \bar{w}_{ij}\right] \text{ if } h = k \\ -2\bar{w}_{kj} \text{ if } j \neq k \end{cases} \quad \text{(I)}$$

Proof

the proof is split in two parts: a first part showing how to get the first term of the system (i.e., the diagonal elements, for which $j = k$) (1) and a second part in which showing how to get the second term of the system (i.e., the off-diagonal elements, for which $j \neq k$) (2).

(1) It is possible to get the right form for the diagonal elements by considering the derivative of the gradient of $f(\boldsymbol{y})$ (here shown for a single component of $\boldsymbol{y}$):

$$\frac{\partial}{\partial y^{(j)} \partial y^{(j)}} f(\boldsymbol{y}) = \frac{\partial}{\partial y^{(j)}} \nabla_{y^{(j)}} f(\boldsymbol{y}) = \frac{\partial}{\partial y^{(j)}} \left[2 \sum_{i=1}^{l} w_{ij} \left(y^{(j)} - \bar{y}^{(i)}\right) + 2 \sum_{i=1}^{u} \bar{w}_{ij} \left(y^{(j)} - y^{(i)}\right) \right] =$$

$$= 2 \frac{\partial}{\partial y^{(j)}} \sum_{i=1}^{l} w_{ij} \left(y^{(j)} - \bar{y}^{(i)}\right) + 2 \frac{\partial}{\partial y^{(j)}} \sum_{i=1}^{u} \bar{w}_{ij} \left(y^{(j)} - y^{(i)}\right)$$

From now on, it is possible to calculate the first term highlighted in light red and second term highlighted in light blue separately:

First term

$$2 \frac{\partial}{\partial y^{(j)}} \sum_{i=1}^{l} w_{ij} \left(y^{(j)} - \bar{y}^{(i)}\right) = 2 \sum_{i=1}^{l} w_{ij} \frac{\partial}{\partial y^{(j)}} \left(y^{(j)} - \bar{y}^{(i)}\right) = 2 \sum_{i=1}^{l} w_{ij} \frac{\partial}{\partial y^{(j)}} y^{(j)} = 2 \sum_{i=1}^{l} w_{ij}$$

Since $\frac{\partial}{\partial y^{(j)}} - \bar{y}^{(i)} = 0 \forall i \in \{1, \dots, l\}$ and $\frac{\partial}{\partial y^{(j)}} y^{(j)} = 1$.

Second term

$$2 \frac{\partial}{\partial y^{(j)}} \sum_{i=1}^{u} \bar{w}_{ij} \left(y^{(j)} - y^{(i)}\right) = 2 \sum_{i=1}^{u} \bar{w}_{ij} \frac{\partial}{\partial y^{(j)}} \left(y^{(j)} - y^{(i)}\right) = 2 \sum_{i=1}^{u} \bar{w}_{ij} \frac{\partial}{\partial y^{(j)}} y^{(j)}$$

Same as before, since $\frac{\partial}{\partial y^{(j)}} - y^{(i)} = 0 \forall i \in \{1, \dots, u\}$. However, in this case we get:

$$2 \sum_{i=1}^{u} \bar{w}_{ij} \frac{\partial}{\partial y^{(j)}} y^{(j)} = \begin{cases} 2 \sum_{i=1}^{u} \bar{w}_{ij} \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases} = 2 \left(\sum_{i=1}^{u} \bar{w}_{ij} - w_{jj} \right)$$

Since we have:

$$\sum_{i=1}^{u} \bar{w}_{ij} = \bar{w}_{1j} + \cdots + \bar{w}_{j-1j} + 0 + \bar{w}_{j+1j} + \cdots + \bar{w}_{uj} = 2\sum_{i=1}^{u} \bar{w}_{ij} - 2w_{jj}$$

Finally, if $j = k$, the two terms summed back again give:

$$\frac{\partial}{\partial y^{(j)}} \nabla_{y^{(j)}} f(\boldsymbol{y}) = 2\sum_{i=1}^{l} w_{ij} + 2\sum_{i=1}^{u} \bar{w}_{ij} - 2w_{jj} = 2\left[\left(\sum_{i=1}^{l} w_{ij} + \sum_{i=1}^{u} \bar{w}_{ij}\right) - \bar{w}_{ij}\right]$$

(2) For the second term of the system (I) it is possible to do the same:

$$\frac{\partial}{\partial y^{(j)} \partial y^{(k)}} f(\boldsymbol{y}) = \frac{\partial}{\partial y^{(k)}} \nabla_{y^{(j)}} f(\boldsymbol{y}) = \frac{\partial}{\partial y^{(k)}} \left[2\sum_{i=1}^{l} w_{ij}\left(y^{(j)} - \bar{y}^{(i)}\right) + 2\sum_{i=1}^{u} \bar{w}_{ij}\left(y^{(j)} - y^{(i)}\right)\right] =$$

$$= 2\frac{\partial}{\partial y^{(k)}} \sum_{i=1}^{l} w_{ij}\left(y^{(j)} - \bar{y}^{(i)}\right) + 2\frac{\partial}{\partial y^{(k)}} \sum_{i=1}^{u} \bar{w}_{ij}\left(y^{(j)} - y^{(i)}\right)$$

First term

$$2\frac{\partial}{\partial y^{(k)}} \sum_{i=1}^{l} w_{ij}\left(y^{(j)} - \bar{y}^{(i)}\right) = 2 \cdot 0$$

Since there all the sum is considered as a constant in this case.

Second term

$$2\frac{\partial}{\partial y^{(k)}} \sum_{i=1}^{u} \bar{w}_{ij}\left(y^{(j)} - \bar{y}^{(i)}\right) = 2\sum_{i=1}^{u} \bar{w}_{ij} \frac{\partial}{\partial y^{(k)}}\left(y^{(j)} - \bar{y}^{(i)}\right) = -2\sum_{i=1}^{u} \bar{w}_{ij} \frac{\partial}{\partial y^{(k)}} y^{(i)} = \begin{cases} 0 \text{ if } i \neq k \\ -2\bar{w}_{ij} \text{ if } i = k \end{cases}$$

Considering that $\frac{\partial}{\partial y^{(k)}} y^{(j)} = 0$. Also, $i \in \{1, \dots, u\}$ and $k \in \{1, \dots, u\}$, at a certain point it is possible to meet $i = k$. So:

$$-2\sum_{i=1}^{u} \bar{w}_{ij} \frac{\partial}{\partial y^{(k)}} y^{(j)} = 2\left(\bar{w}_{1j} \frac{\partial}{\partial y^{(k)}} y^{(1)} + \cdots + \bar{w}_{kj} \frac{\partial}{\partial y^{(k)}} y^{(k)} + \cdots + \bar{w}_{1j} \frac{\partial}{\partial y^{(k)}} y^{(k)}\right) =$$

$$= -2\left(0 + \cdots + \bar{w}_{kj} + \cdots + 0\right) = -2\bar{w}_{kj}$$

Finally, if $j \neq k$, the two terms summed back again give:

$$\frac{\partial}{\partial y^{(j)} \partial y^{(k)}} f(\boldsymbol{y}) = 0 - 2\bar{w}_{kj} = -2\bar{w}_{kj}$$

At last, by merging the results got for (1) and (2) we get:

$$H_{jk} = \nabla^2_{y^{(j)} y^{(k)}} f(\boldsymbol{y}) = \begin{cases} 2\left[\left(\sum_{i=1}^{l} w_{ij} + \sum_{i=1}^{u} \bar{w}_{ij}\right) - \bar{w}_{ij}\right] \text{ if } h = k \\ -2\bar{w}_{kj} \text{ if } j \neq k \end{cases}$$

That is the system (I). ■

Section 2 Mutual information calculation

Mutual information is a measure for determining the mutual dependence of two given variables, performed by quantifying amount of information gained about the first variable through observing the second variable.

Given two random variables X and Y, the mutual information I is calculated using the following formula:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} \mathbb{P}_{X,Y}[x,y] \log\left(\frac{\mathbb{P}_{X,Y}[x,y]}{\mathbb{P}_X[x]\mathbb{P}_Y[y]}\right)$$

Where $\mathbb{P}_X[x]$ is the probability of X, $\mathbb{P}_Y[y]$ is the probability of Y and $\mathbb{P}_{X,Y}[x,y]$ is the probability for X and Y to occur together (joint probability).

In this case we have:

$$X \perp\!\!\!\perp Y \Rightarrow I(X;Y) = 0$$

The mathematics around this concept will not be further explored in this paper, since it is beyond the purpose of this work.

By using this technique, it was possible to determine a ranking plot (Figure 1) for all the features of the Diabetes dataset. This allows to have a good idea of the amount of information for each variable.

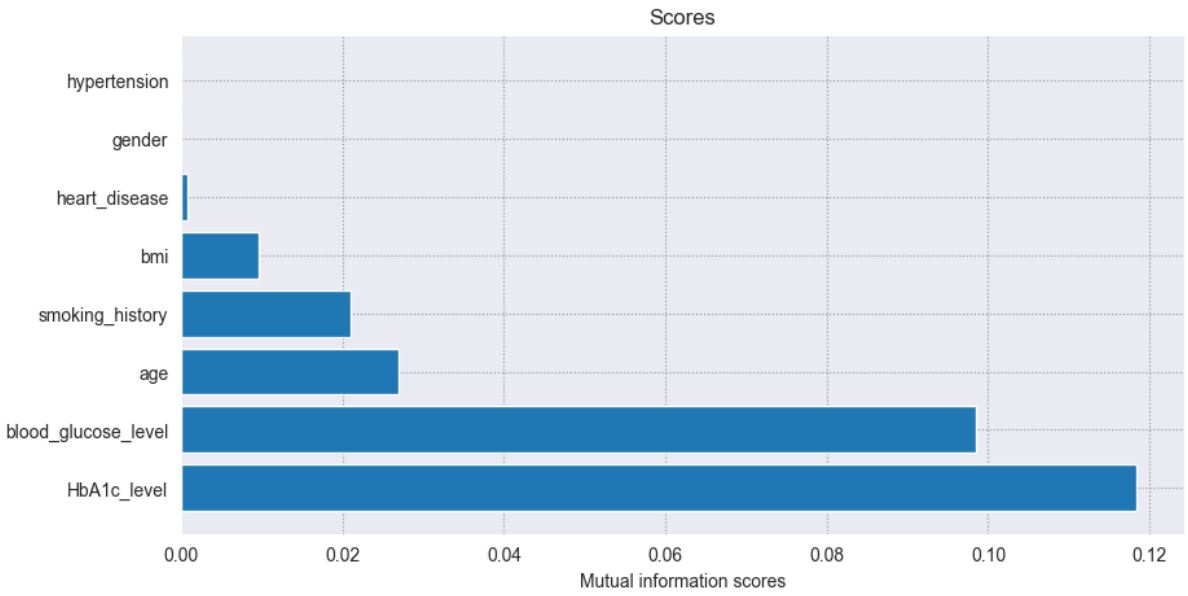


Figure 1 – By looking at the plot rank, the decision was to keep only the top 4 features, basing on the amount of information provided.

Section 3 Lipschitz constant calculation

For getting the Lipschitz constant, it would be necessary to solve the following characteristic equation for the regularized Hessian matrix:

$$\det\left(\boldsymbol{H_{jk,reg}} - \lambda \boldsymbol{I}_{\mathbb{R}^{u \times u}}\right) = 0$$

The solved equation provides the set of the eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ for the Hessian matrix and the Lipschitz constant can be found with the following:

$$L = \max\{|\lambda_1|, \dots, |\lambda_u|\}$$

However, the characteristic equation has not been solved manually and, using the NumPy Python library, as mentioned just before, it was possible to access to specific numeric efficient and sophisticated methods for an efficient computation of the eigenvalues, using the same underlying principle shown above.

Concerning the condition number, it is possible to see that, in practice, it is calculated by using the following ratio:

$$K(\boldsymbol{H_{jk,reg}}) = \frac{\sigma_{max}}{\sigma_{min}}$$

Where σ_{max} and σ_{min} are, respectively, the maximum and the minimum singular values, found by calculating the singular value decomposition (SVD) of $\boldsymbol{H_{jk,reg}}$. A large condition number means in this case that the matrix is ill-conditioned and, therefore, even a small error in the data can cause a large error in the solution when the matrix is used to solve a system of linear equations. Even in this case, the mathematics behind this specific concept will not be delved further since it is beyond the purpose of this discussion.

Section 4 Hyperparameters

The following table show the hyperparameters chosen for each dataset:

Dataset	Model	$\boldsymbol{\gamma}^*$	$\boldsymbol{\alpha}^{**}$	Iterations	Learning rates
Synthetic	GD	3	0.01	50	[0, 3/L]
	RP-BCGD	3	0.01	50	[0, 15/L]
	GS-BCGD	3	0.01	500	[0, 15/L]
Penguins	GD	5	2.5	500	[0, 3/L]
	RP-BCGD	5	2.5	250	[0, 30/L]
	GS-BCGD	5	2.5	2000	[0, 25/L]
Diabetes	GD	3	5	75	1/L
	RP-BCGD	3	5	250	10/L
	GS-BCGD	3	5	5000	15/L

* γ is the hyperparameter used in the RBF kernel function.

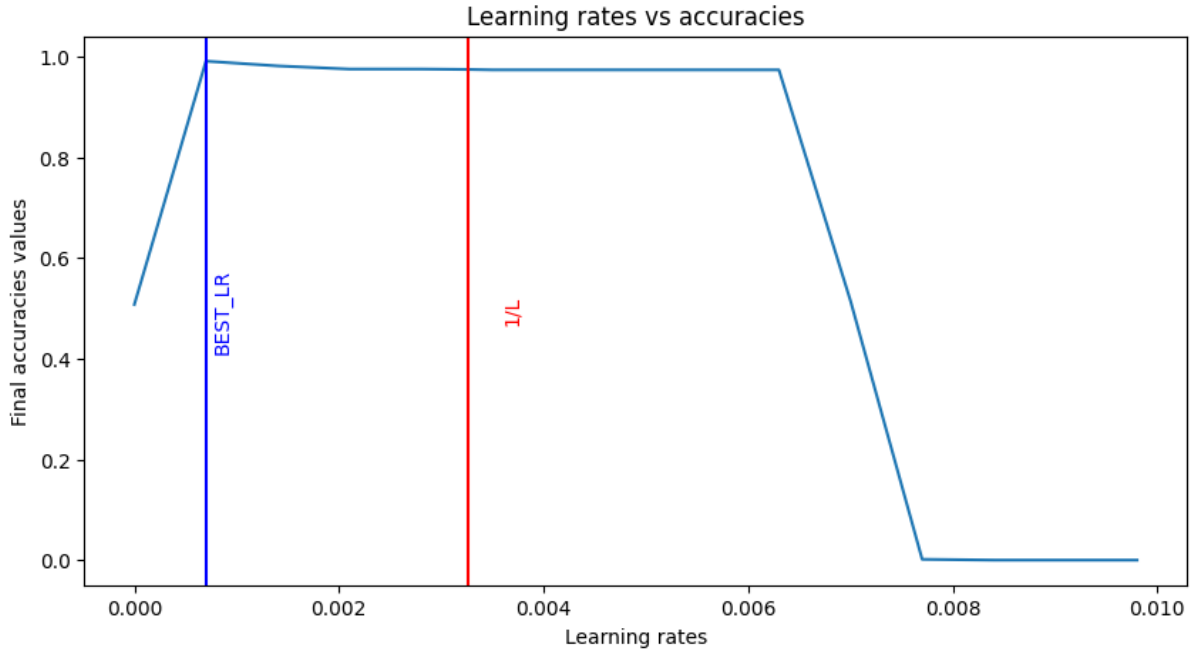
** α is the hyperparameter used for regularizing the Hessian matrix.

Section 5 First part plots

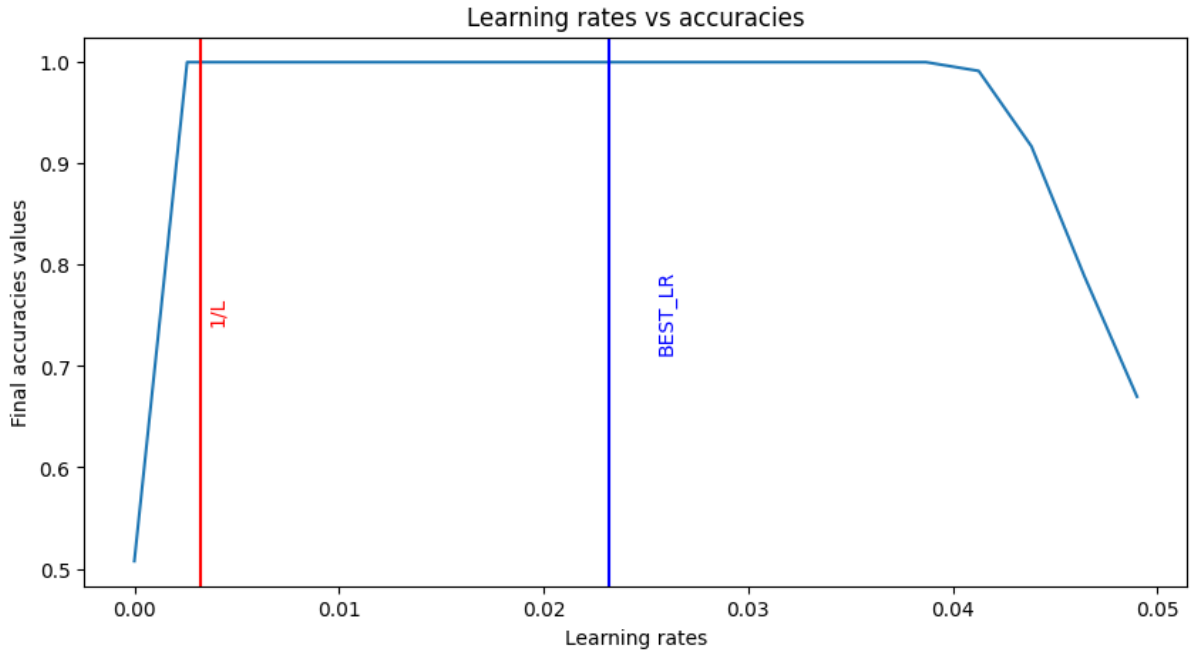
This section shows the final accuracies values variation in relation to learning rate increase for the synthetic dataset and for the Penguins Seaborn toy dataset. The blue vertical lines show the learning rate corresponding to the best accuracy found, while the red vertical lines show the learning rate found for the value of 1/L. The value of the accuracy found using the inverse of the Lipschitz constant seems nice for performing simple GD, while it slightly underestimate the best learning rate in the case of BCGD (mostly for the GS method), how can be easily seen from this plots.

Synthetic dataset

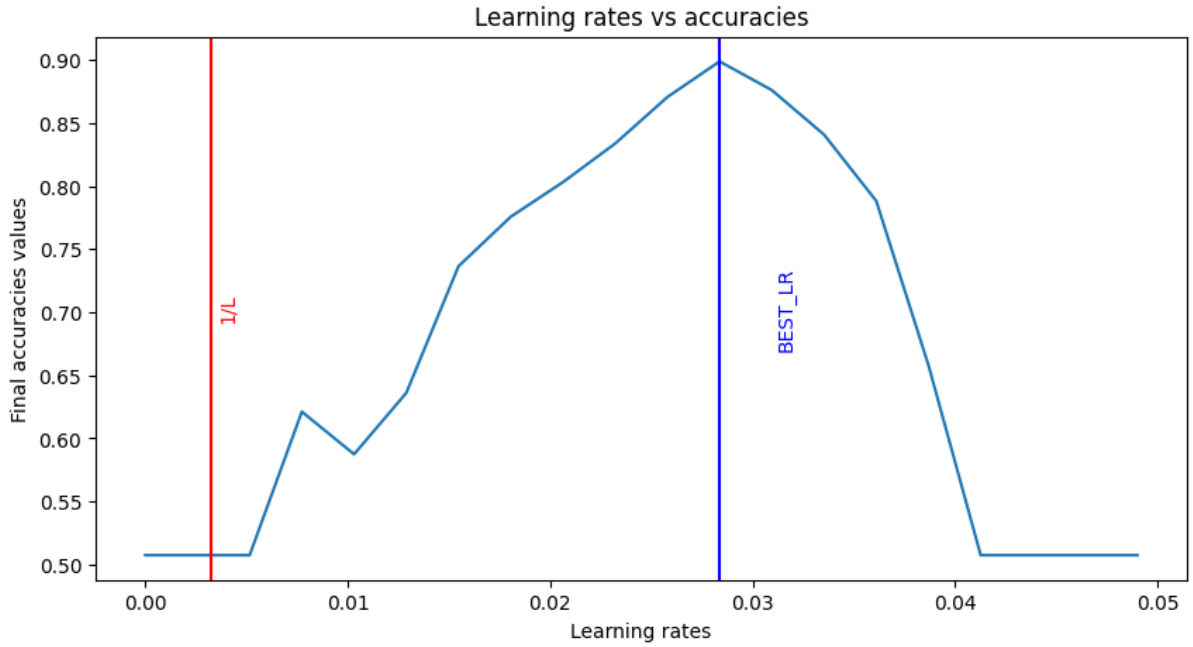
Simple GD



RP-BCGD

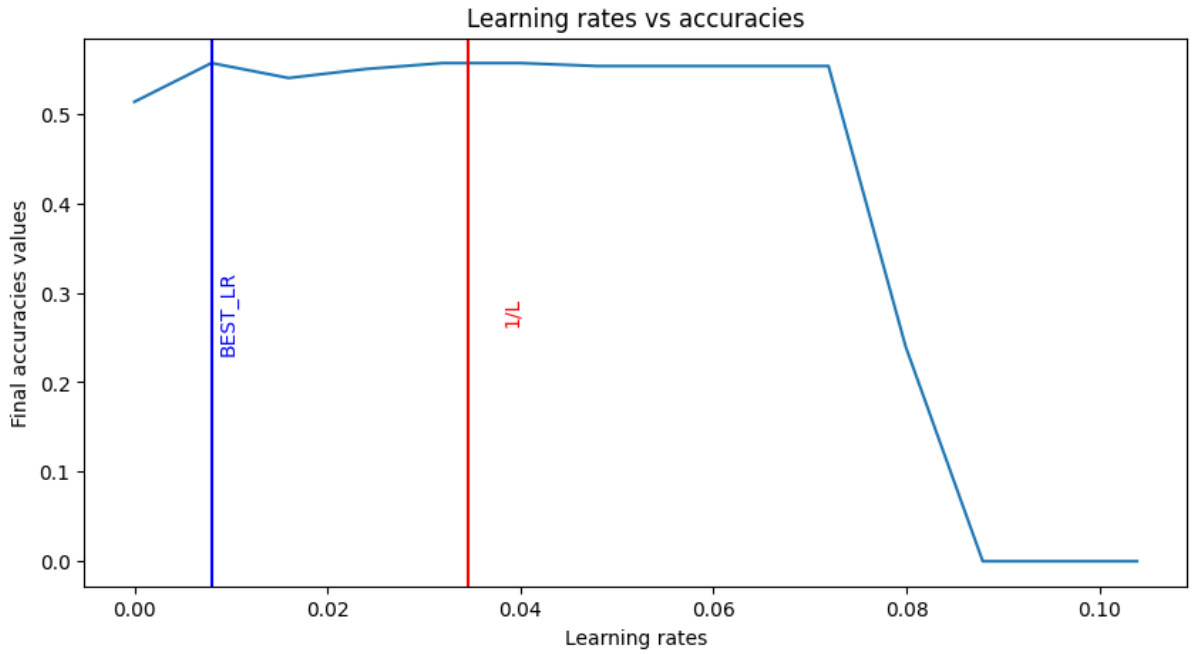


GS-BCGD

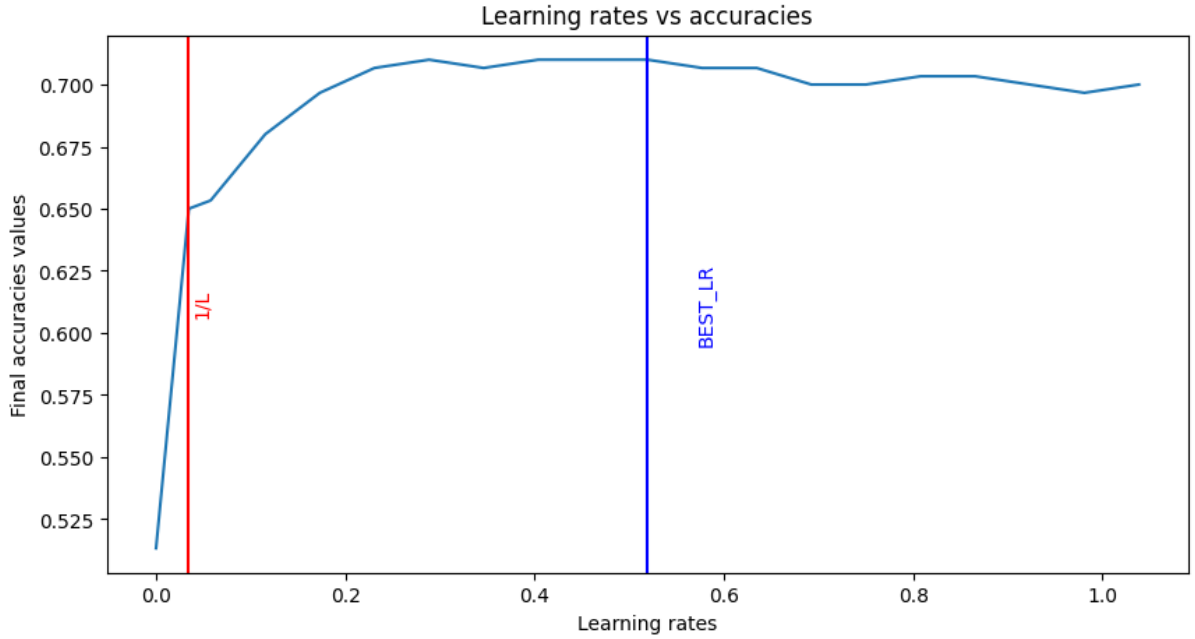


Penguins toy dataset pipeline

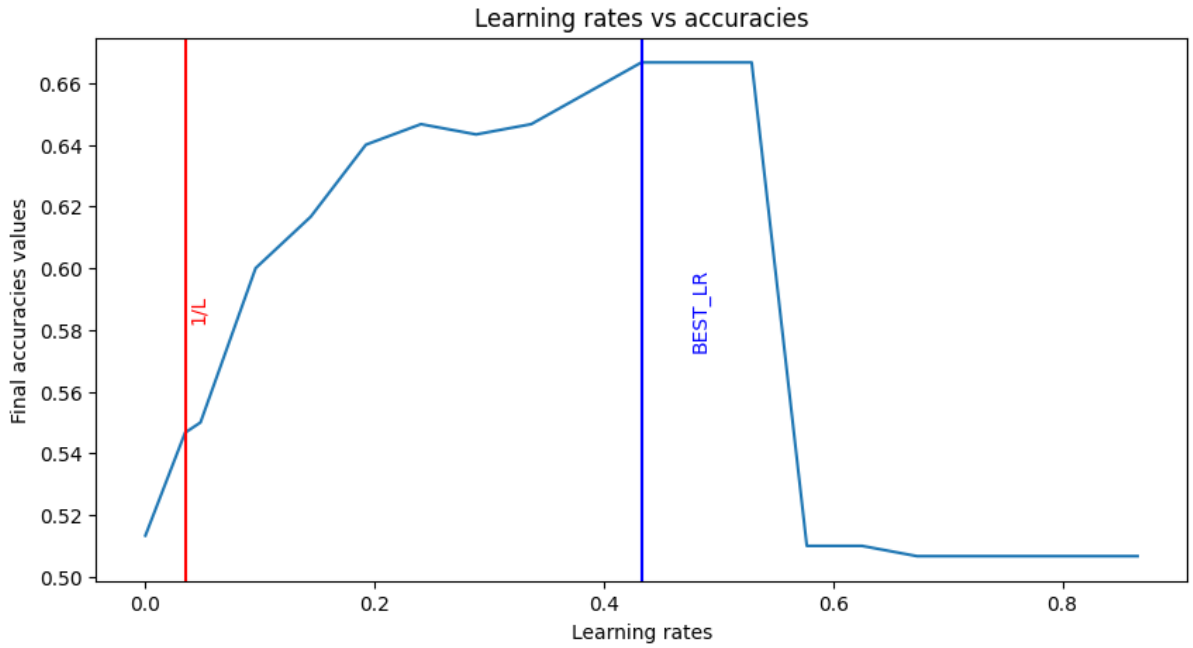
Simple GD



RP-BCGD



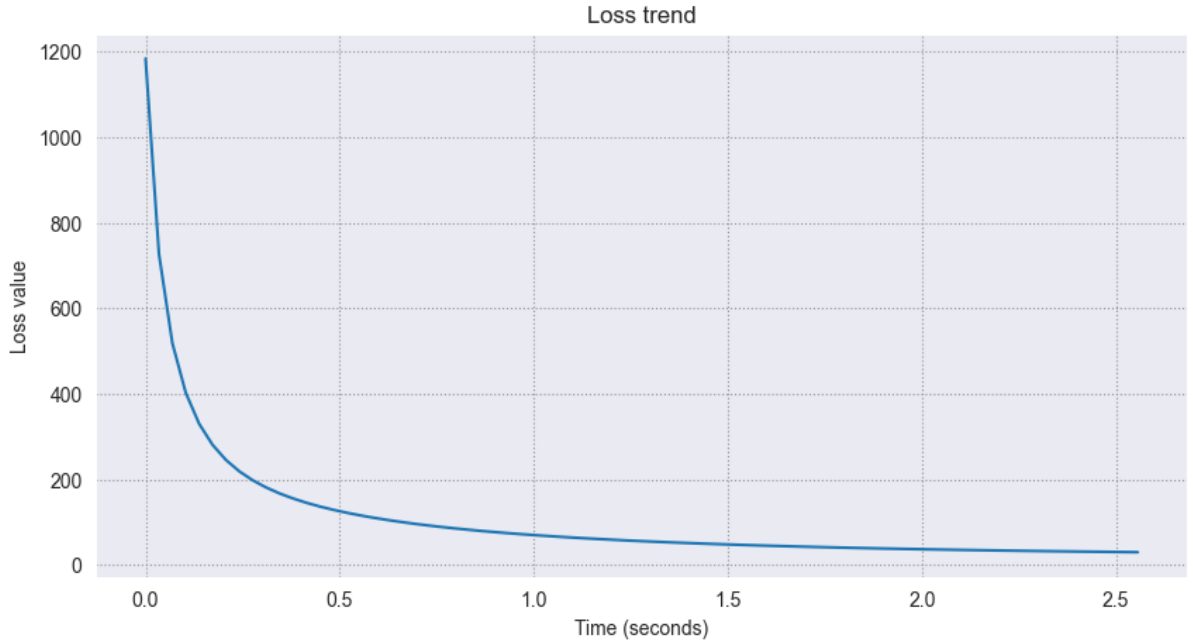
GS-BCGD

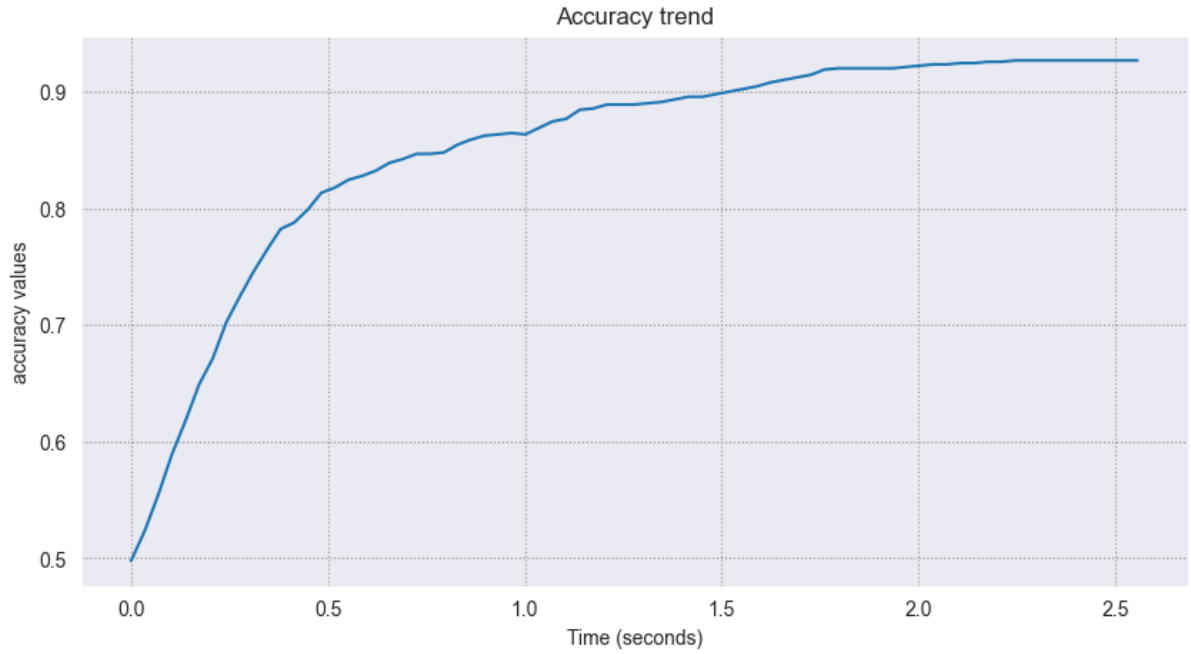


Section 6 Second part plots

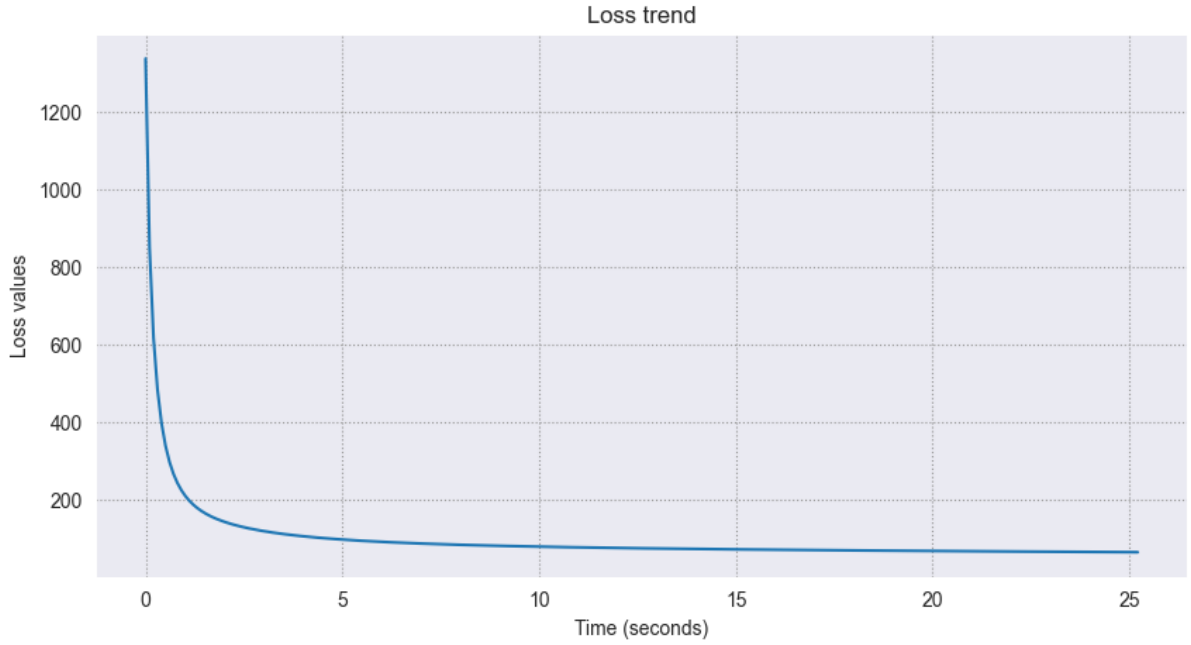
In this section are shown all the loss trend and accuracy trend plots returned from the application of the model to the Diabetes real dataset problem. All the plots highlight a nice trend, without specific problems and with a typical loss descent curve.

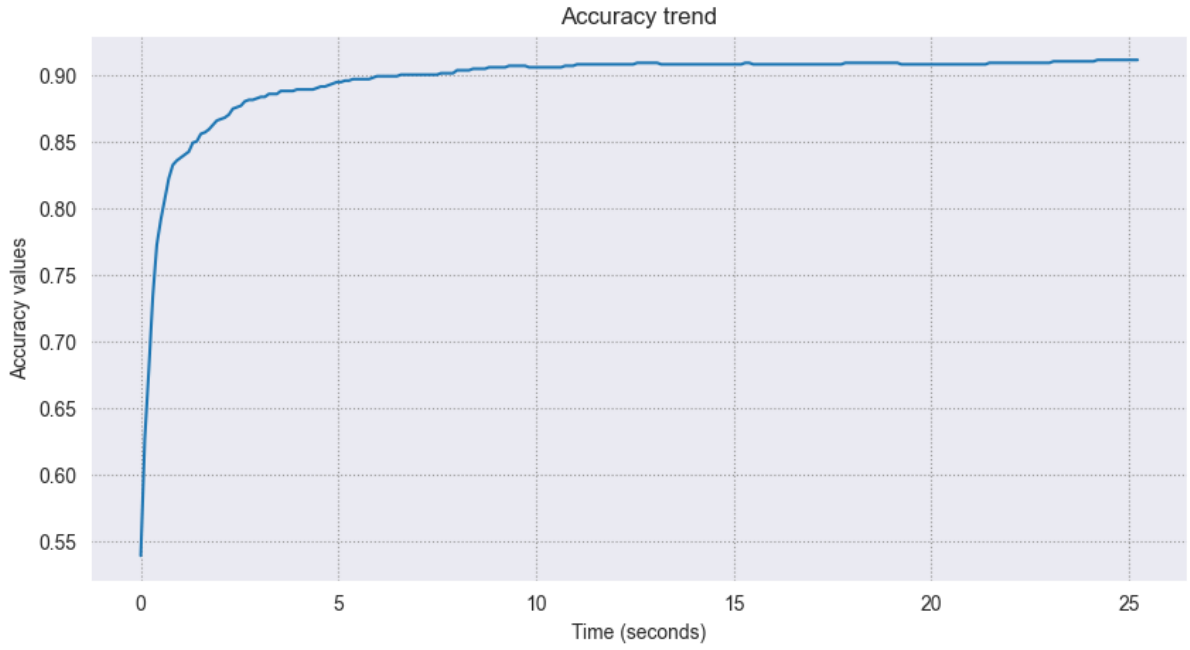
Simple GD



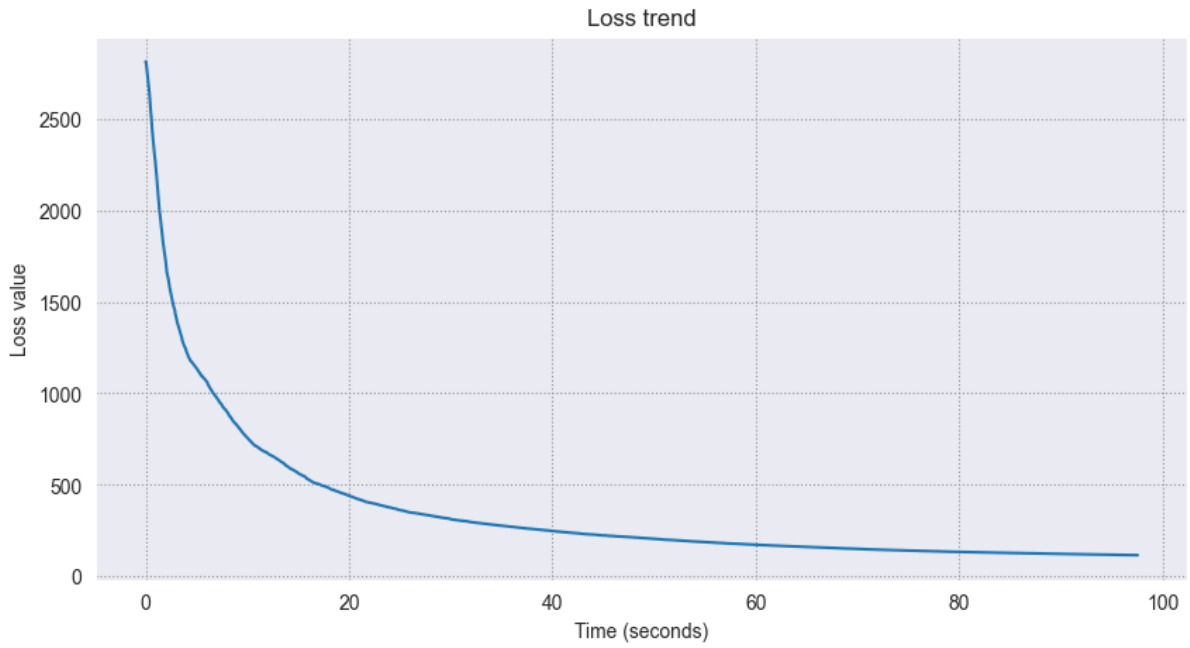

Accuracy trend
accuracy values
0.9
0.8
0.7
0.6
0.5
0.0
0.5
1.0
1.5
2.0
2.5
Time (seconds)

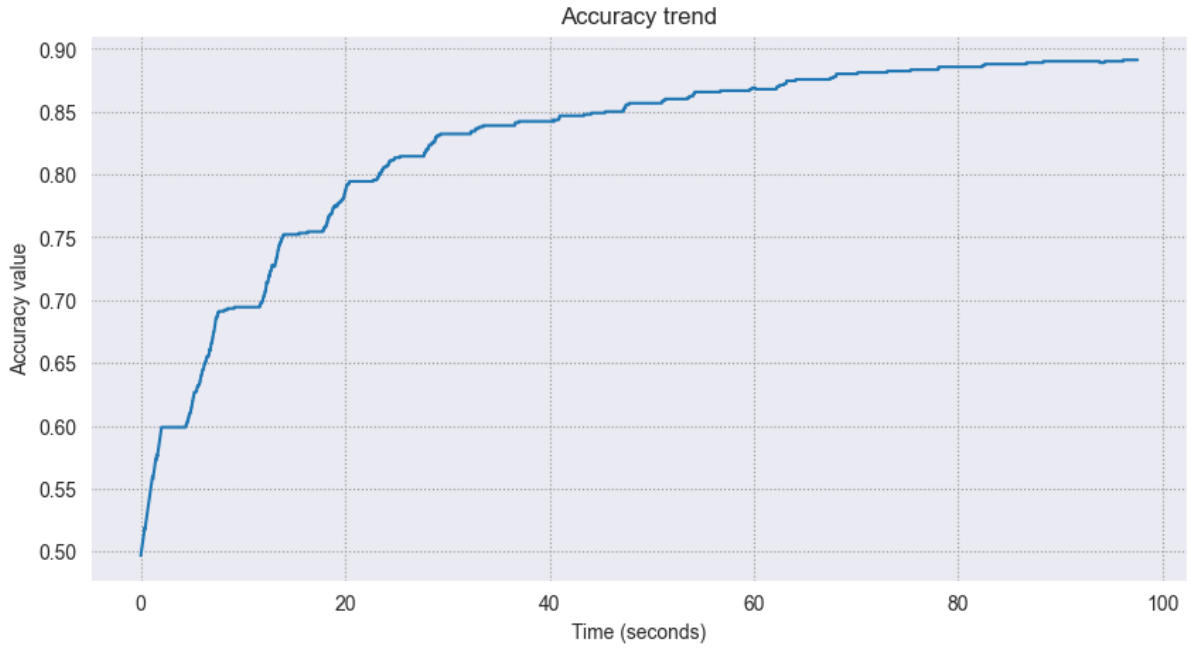
RP-BCGD


Loss trend
Loss values
1200
1000
800
600
400
200
0
5
10
15
20
25
Time (seconds)


Accuracy trend
Accuracy values
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
25
Time (seconds)

GS-BCGD


Loss trend
Loss value
2500
2000
1500
1000
500
0
0
20
40
60
80
100
Time (seconds)


Accuracy trend
Accuracy value
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
20
40
60
80
100
Time (seconds)