Homework 1

Firstname Lastname

April 17, 2020

This is a small template for homework.

Problem 1

> MNIST dataset is used for a classification tasks. Denote the dataset as $D = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{N}$, where $x^{(i)}$ is the image and $y^{(i)} \in \{0, 1, \ldots, 9\}$ is the groudtruth label. Lets stretch the image into a one-dimensional vector (from 28×28 to 784×1) and transform the labels into one-hot vectors, e.g. $y^{(i)} = 2 \rightarrow \begin{bmatrix}0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T$. Your task is to derive the gradient, and implement the gradient descent (GD) algorithm to optimize the following linear regression problem:
>
> $$L\left(\mathbf{W}\right) = \frac{1}{N}\sum_{i=1}^{N}\left|\left|\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right|\right|_2^2$$
>
> Here, $\mathbf{W}$ and $\mathbf{w_0}$ are the parameters of your linear model. Test different step sizes τ and discuss how it influences the result. Also, draw the curve of the training error and test error during optimization, and discuss how they evolve during optimization. You should program the algorithms by yourself. You are not allowed to use the predefined modules in PyTorch (e.g., `nn.Linear`, `optim.SGD`).

First, we must derive an expression of the gradient. We note that the gradient is linear so:

$$\nabla L(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^{N}\nabla\left(\left|\left|\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right|\right|_2^2\right)$$

The simplest way to evaluate this is to write the summand as the inner-product of a vector with itself:

$$\left|\left|\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right|\right|_2^2 = \left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right)^T\left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right)$$

Distributing this multiplication and, once again, utilizing the fact that the gradient is a linear operator, we may quickly compute the gradient with respect to the parameters $\mathbf{W}$ and $\mathbf{w}_0$:

$$\begin{aligned}
\nabla_{\mathbf{W}}\left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right)^T\left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right) &= 2\left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right)\left(x^{(i)}\right)^T \\
\nabla_{\mathbf{w}_0}\left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right)^T\left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right) &= 2\left(\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right)
\end{aligned}$$

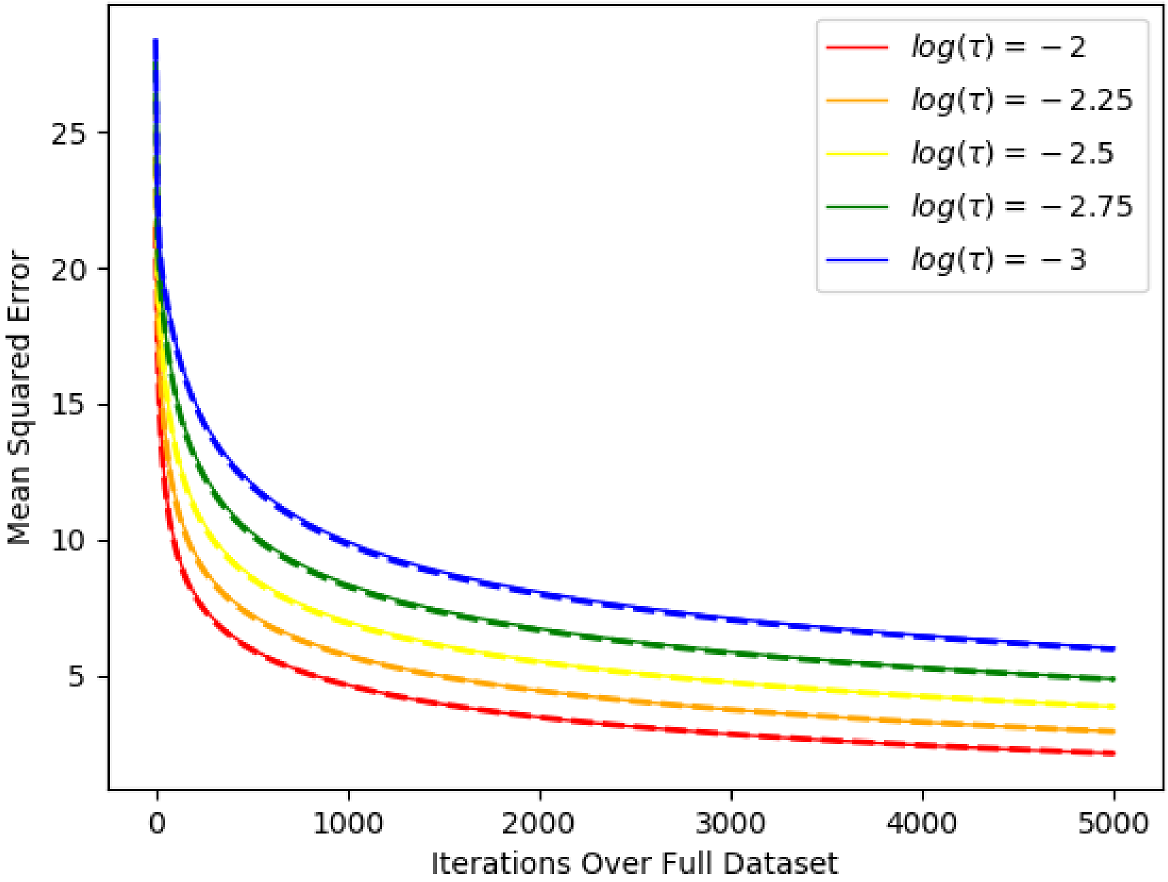

Figure 1: Comparison of gradient descent method for various learning rates. Solid lines indicate mean squared error on the training data, and dashed lines indicate MSE on the test data.

Problem 2

> Optimize the linear regression problem with Stochastic Gradient Descent (SGD).
>
> (a) Fixing the batch size m, test different step sizes τ, draw the curves of training error and testing error, and discuss the influence of step size.
>
> (b) Fixing the step size, test different batch sizes, draw the curves of training and testing error, discuss the influence of batch size
>
> (c) Use the same step size for GD and SGD. Draw the curves of training and testing error for both algorithms, where the x-axis is the number of training images used. Discuss the difference between both algorithms.

scooby

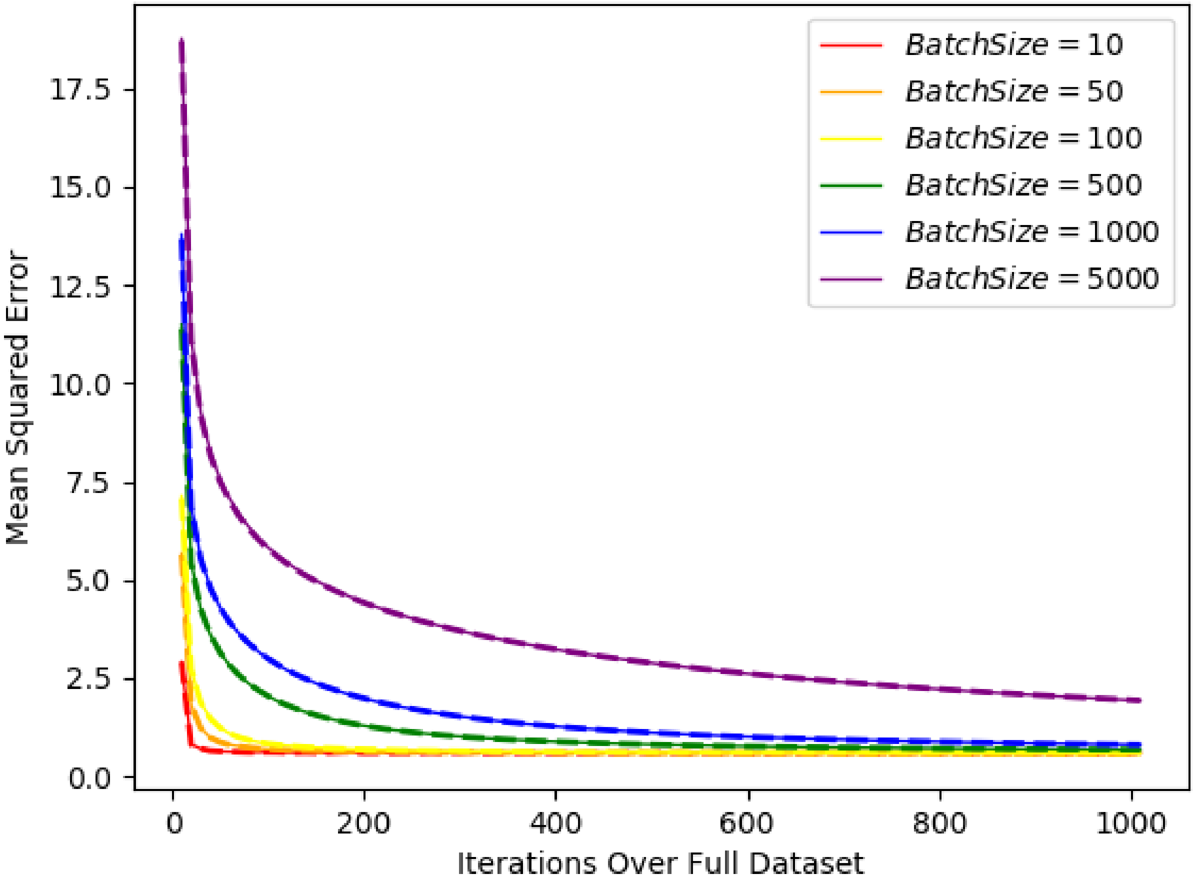

Figure 2: Comparison of stochastic gradient descent method for various batch sizes with a fixed learning rate of $\tau = 5e - 3$. Solid lines indicate mean squared error on the training data, and dashed lines indicate MSE on the test data.

Problem 3

Let's add $L2$ regularization to the optimization problem:

$$L(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^{N} \left|\left|\mathbf{W}x^{(i)} + \mathbf{w_0} - y^{(i)}\right|\right|_2^2 + \lambda\left(a\,||\mathbf{W}||_F^2 + ||\mathbf{w}_0||_2^2\right)$$

This problem is also called *ridge regression*.

(a) Optimize the problem with SGD when $\lambda = 1$. Draw the curve of training and testing error

(b) With $\lambda \in \{0.01, 0.1, 1, 10\}$, test how different choices of λ influence the final test error.

(c) Choose several hyperparameters with numbers you like, and use cross-validation to choose the best combination of λ, τ, and m.

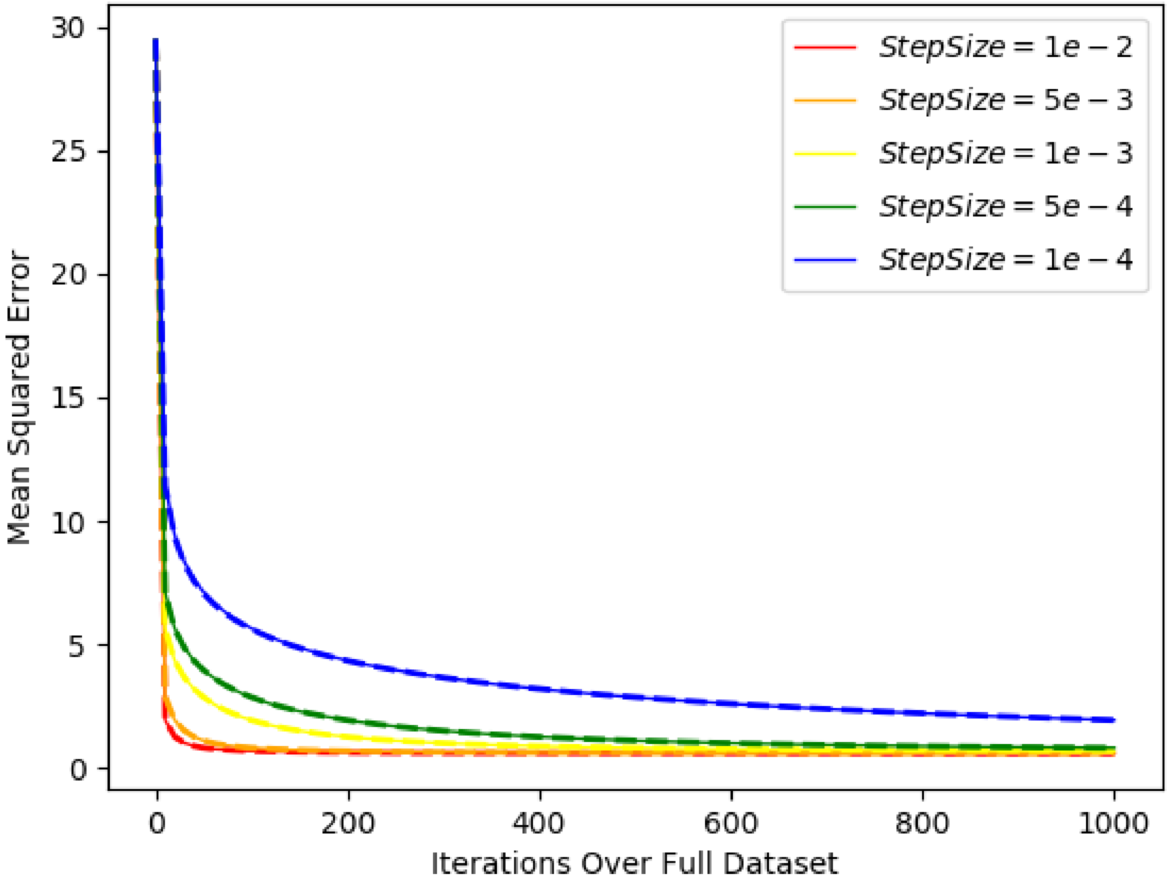

Figure 3: Comparison of stochastic gradient descent method for various learning rates with a fixed batch size of 100. Solid lines indicate mean squared error on the training data, and dashed lines indicate MSE on the test data.

asld;kfja;lskdfj

Problem 4

> Let's move from fixed step size to changing step size. Use the best combination you obtained from the previous task. For the optimization problem in Problem 3, implement and report hte performance of
>
> (a) decreasing step size
>
> (b) cosine step size scheduler
>
> with SGD. The cosine step size scheduler decays the step size with a cosine annealing for each iteration. Its expression is as follows:
>
> $$\tau_i = \gamma^i \left(\tau_{min} + \frac{1}{2} \left(\tau_{max} - \tau_{min} \right) \left(1 + \cos \left(\pi i / T \right) \right) \right)$$

where τ_{min} and τ_{max} are ranges for the learning rate, and T controls the period of the step size cyle. An annealing factor γ is adopted to ensure the step size approaches zero. Compare and discuss the performance between (a), (b), and constant step size.

asdfasdf

Batch Size	λ	τ	**MSE**
30	0.010	0.010	0.593
100	0.010	0.010	0.585
30	0.010	0.001	0.583
100	**0.010**	**0.001**	**0.5821**
30	0.100	0.010	0.622
100	0.100	0.010	0.616
30	0.100	0.001	0.613
100	0.100	0.001	0.613

Table 1: Cross-validation Results

and then the story continues.

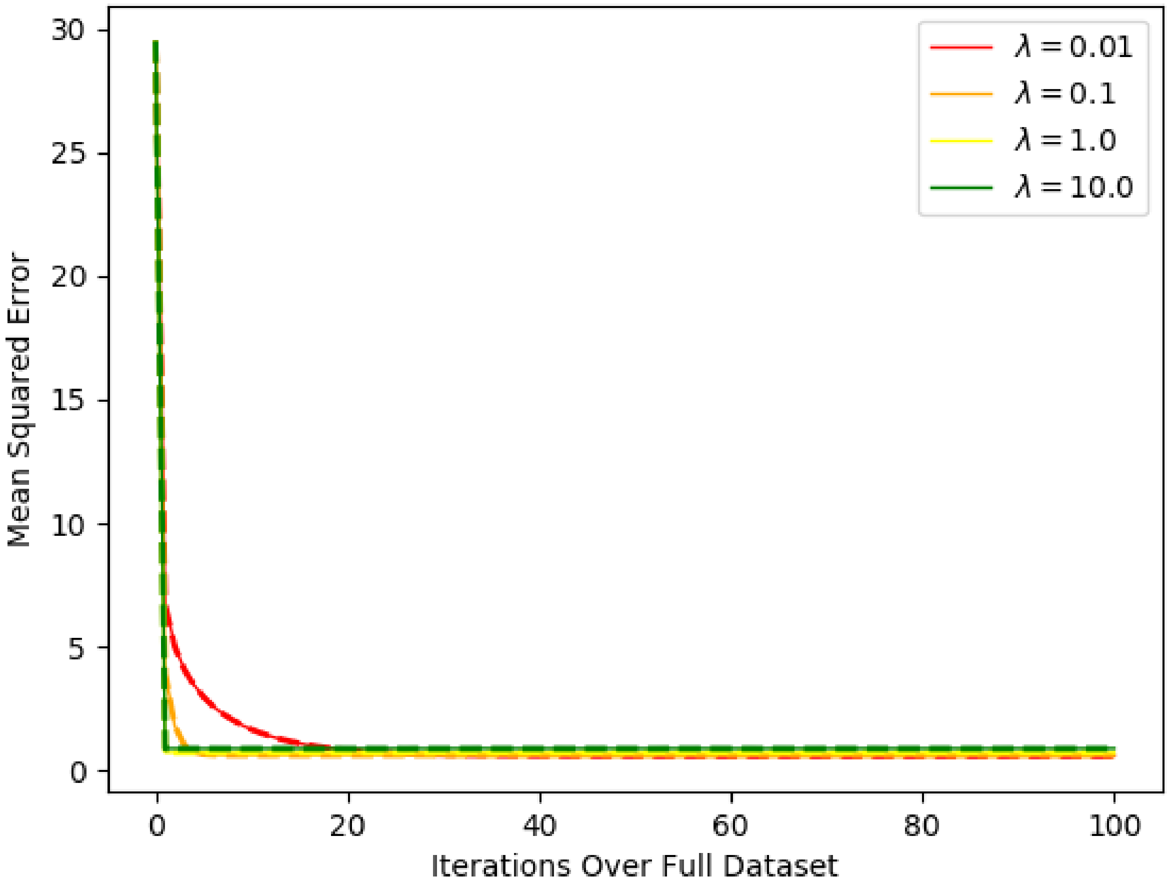

Figure 4: Comparison of stochastic gradient descent method applied to the ridge regression problem with various ridge coefficients. The batch size was fixed at 100, and the learning rate was fixed at $\tau = 5e - 3$. Solid lines indicate mean squared error on the training data, and dashed lines indicate MSE on the test data.

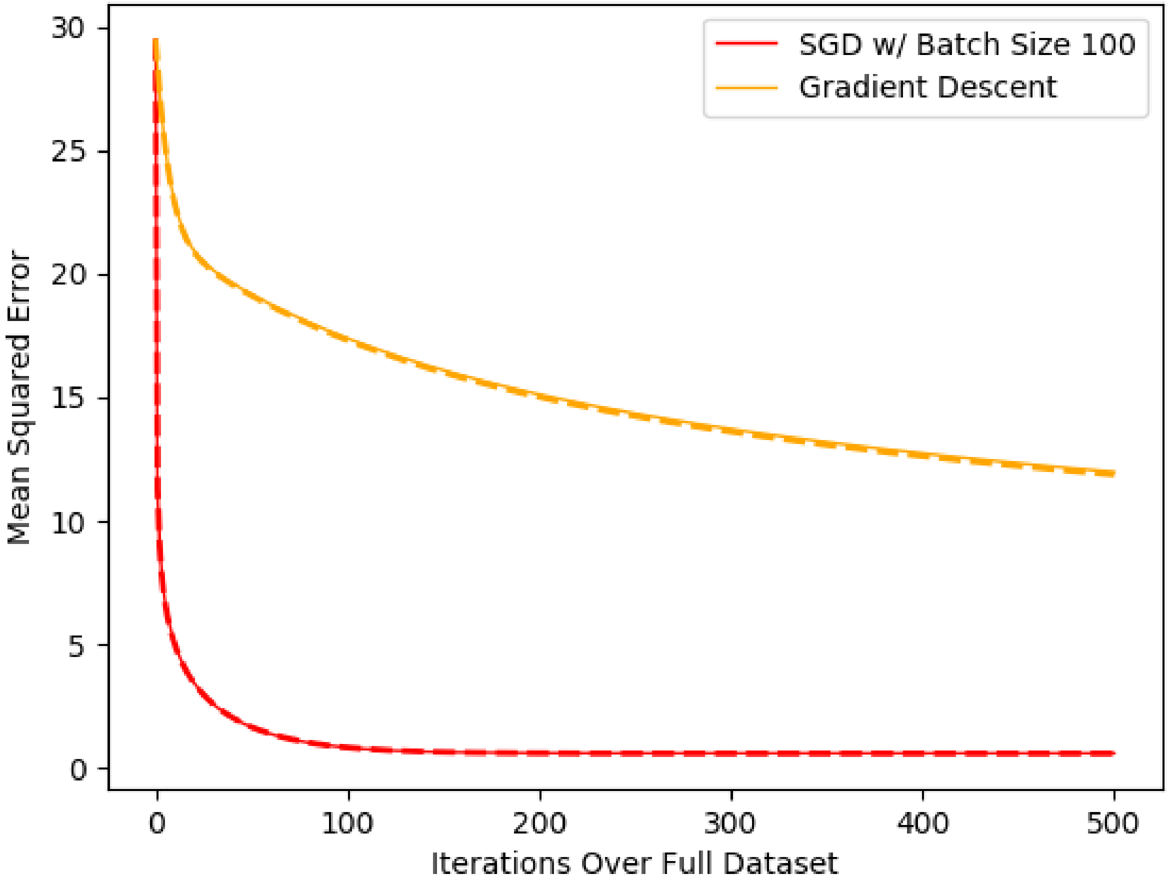

Figure 5: Comparison of gradient descent convergence with that of stochastic gradient descent with batch size of 100. The optimization was performed with $\tau = 0.001$ and $\lambda = 0.01$. Solid lines indicate mean squared error on the training data, and dashed lines indicate MSE on the test data.

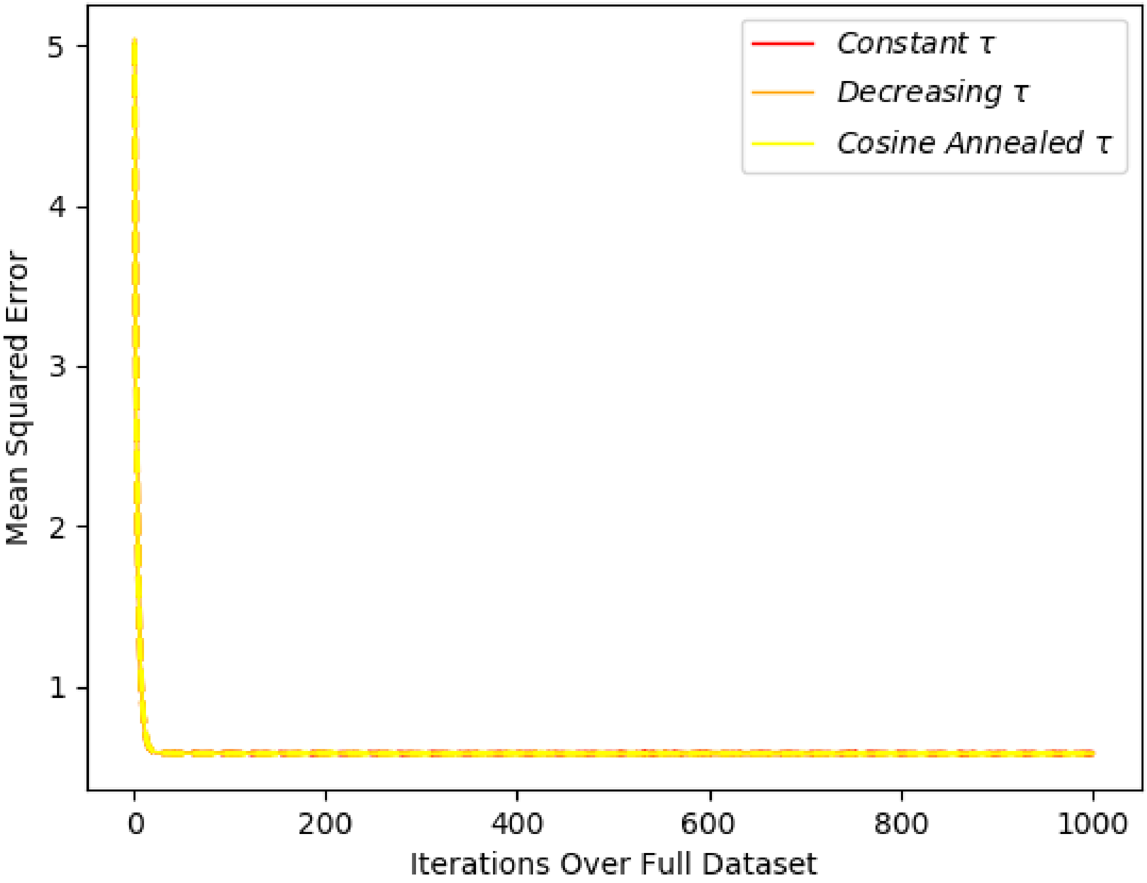

Figure 6: Comparison of stochastic gradient descent method applied to the ridge regression problem with various ridge coefficients. The batch size was fixed at 100, and the learning rate was fixed at $\tau = 1e - 2$. Solid lines indicate mean squared error on the training data, and dashed lines indicate MSE on the test data.