Michael Potter
UID:
mlpotter124@g.uclda.edu

Homework Assignment 1
CS260: Machine Learning Algorithms

2021-01-22

# 1 Model Training

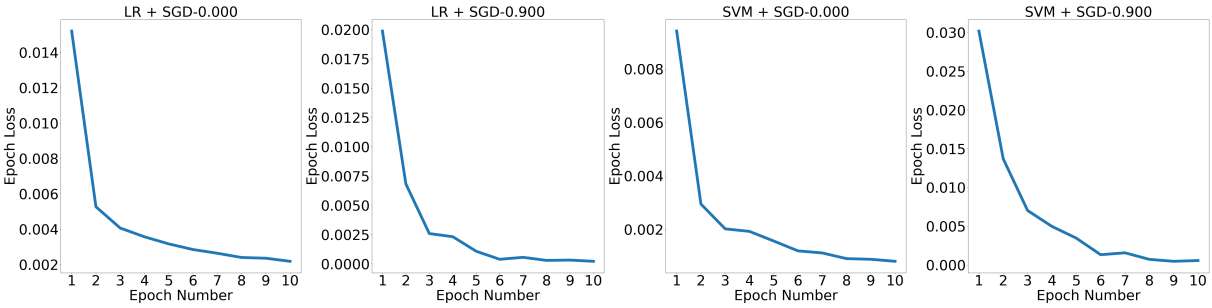
A Linear Logistic Regression (LR) model and Linear Support Vector Machine (SVM) model were trained on the MNIST dataset (subsetted on handrawn zeros and ones). The loss functions for Linear LR and Linear SVM are binary cross entropy (eqn. 1) and hinge loss with a margin of 1 (eqn. 2), respectively. The models are optimized with stochastic gradient descent (SGD) or stochastic gradient descent with momentum (SGD-Momentum).

$$BCE(\hat{y}_i, y_i) = y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(\hat{y}_i) \tag{1}$$

$$Hinge(\hat{y}_i, y_i) = max(0, 1 - y_i \cdot \hat{y}_i) \tag{2}$$

## 1.1 Training Curves

All permutations of model methods and gradient descent methods were optimized for a total of four experiments (Figure. 1). For fair comparison and reproducible results between different permutations of model methods and gradient descent methods, the random seed was set to '123' in PyTorch and Numpy. In all experiments the learning rate was set to 0.05.




(a) LR + SGD (b) LR + SGD-0.9 (c) SVM + SGD (d) SVM + SGD-0.9

Figure 1: Training Curves for 10 Epochs (SGD-x is SGD with momentum=x)

## 1.2 Final Test Accuracy

The final test accuracy of the trained models from all permutations of model methods and gradient descent methods did not vary, which suggests that the data was highly linearly separable in the vector space (Table. 2).

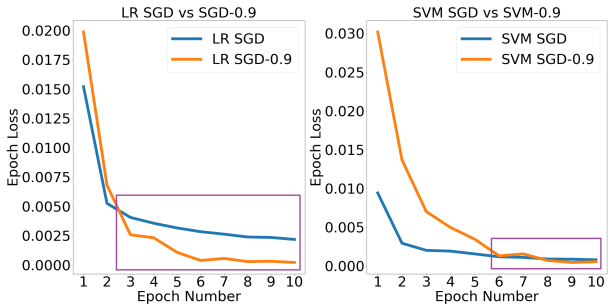
| Model Accuracy | | |
|---|---|---|
| Model | SGD | SGD-0.9 |
| LR | 99.905434% | 99.905434% |
| SVM | 99.952721% | 99.905434% |

Table 1: The Final Test Accuracy for Trained Model

## 1.3 Comparing SGD and SGD-Momentum

When comparing the results for the two optimizers (Figure. 2), SGD and SGD-Momentum, there is a transition point where the training loss becomes smaller with SGD-Momentum. The transition point is indicated in Figure. 2 by the purple box. As expected, as the training loss becomes smaller, the optimizer is approaching a flatter region in the loss function which creates smaller gradients with respect to the model parameters. Intuitively, SGD-momentum may be thought of as pushing a ball off a ledge of a bowl. As the ball moves down along the edge of the bowl, it picks up momentum and becomes faster and faster, and when reaching near the

bottom of the bowl has considerable speed. The bowl represents the convex loss function, the ball represents the current model parameters, and the speed represents the gradient of the loss function with respect to the model parameters.




(a) LR SGD vs SGD-0.9 (b) SVM SGD vs SGD-0.9

Figure 2: Comparison of SGD and SGD-Momentum

## 1.4 Comparing Step Sizes

As the step size decreases by orders of magnitude 10, it is shown that the model accuracy will decrease as well because when the step size is too small, the updates to the model parameter are insignificant and no learning occurs. As the step size increases by oders of magnitude 10, it is shown that the model accuracy does not decrease because of the loss function shape, which may be explained in Figure 3. As the step size increases, the curvature of g becomes smaller, but since the hinge loss and binary cross entropy loss do not curve upward when moving to the right in the x-axis, you decrease the loss (Figure 3).

| Model Accuracy | | |
|---|---|---|
| Step Size | LR | SVM |
| 5e8 | 99.007095% | 99.905434% |
| 5e4 | 99.007095% | 99.905434% |
| 5e3 | 99.007095% | 99.905434% |
| 5e1 | 99.716309% | 99.905434% |
| 5e0 | 99.905434% | 99.905434% |
| 5e-1 | 99.905434% | 99.905434% |
| 5e-2 | 99.905434% | 99.952721% |
| 5e-3 | 99.905434% | 99.952721% |
| 5e-4 | 99.716309% | 99.905434% |
| 5e-5 | 99.574471% | 99.858162% |
| 5e-6 | 90.070923% | 98.723404% |
| 5e-7 | 56.406616% | 87.470451% |
| 5e-8 | 52.576836% | 78.297874% |

Table 2: How Model Test Accuracy Changes w.r.t Step Size

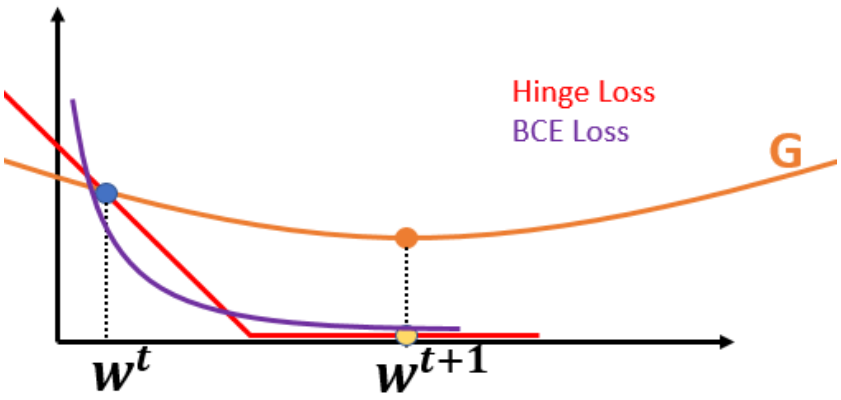



Figure 3: Illustration of Gradient Descent with Loss Functions