Gradient Methods for Muli-class Logistic Regression

Alessandro Pala - 2107800

Tanner Aaron Graves - 2073559

Alisa Snezskaia - 2107497

Anna Glado -

May 2024

1 Introduction

In this report we will analyze the relative performance of various gradient methods at optimizing a multi-class logistic regression model. Specifically, we are interested in comparing full gradient descent with block coordinate based methods (BCGD) which optimize one coordinate, or column in parameter matrix X, at a time. Multi-class logistic classification is an extension of traditional logistic regression. The task goes from predicting a Bernoulli random variable to predicting one label to classify each example from a set of k mutually exclusive labels. We encode the target $b_i \in \mathbb{R}^k$ where k is the number of output classes to be a one hot vector, corresponding to the correct classification of example $a_i \in \mathbb{R}^d$. It assigns these labels by learning the conditional probability distribution that the correct label is b_i given example a_i and learned parameter matrix $X \in \mathbb{R}^{d \times k}$ Where d is the number of features in each example. The output distribution is computed with the softmax function in this context:

$$p(b_i|a_i, X) = \frac{\exp(x_{b_i}^T a_i)}{\sum_{c=1}^{k} \exp(x_c^T a_i)}$$

Multi-class logistic regression is a simple yet effective method with ubiquitous use. Its training does, however, require learning the parameter matrix X, which can be done using various methods like maximum likelihood or gradient descent based methods. This corresponds to solving the following optimization problem:

$$\min_{x \in \mathbb{R}^{d \times k}} \sum_{i=1}^{m} [-x_{b_i}^T a_i + \log(\sum_{c=1}^{k} \exp(x_c^T a_i))]$$

which is the cross-entropy of the predicted class distribution and the target label b_i. Of importance to the application of gradient methods is this problems convexity. This is nice for many reasons, but most importantly we can guarantee that a local minima found by GD methods will be the global minimum.

2 Algorithms

Here we provide an overview to the methods studied. We implemented the various methods in python using either pyTorch or Numpy. The three algorithms studied are full gradient descent and two BCGD methods using both Gauss-Southwell and randomized rule.

2.1 Gradient Descent

Algorithm 1 Full Gradient Descent

```
Initialize $X_1 \in \mathbb{R}^{d\times k}$
**for** $k = 1, ...$ **do**
    If $X \in X^*$ **STOP** (Optimality conditions)
    Set $X_{k+1} = x_k - \alpha_k \nabla$
**end for**
```

2.2 Block coordinate Gradient Descent

BCGD methods are specialized in optimizing high dimensional problems where where calculation of the gradient can be split up in to blocks which are easier to compute. These methods exploit the convex nature of the problem to optimize one coordinate at a time and approach the global optimum. Above $\alpha_k \geq 0$ is

Algorithm 2 BCGD

```
Initialize $X_1 \in \mathbb{R}^{d\times k}$
**for** $k = 1, ...$ **do**
    If $X \in X^*$ **STOP** (Optimality conditions)
    Select $i_k \in$ block according to some **RULE**
    Set $x_{k+1} = x_k - \alpha_k [\nabla_x f(x_k)]_{i_k}$
**end for**
```

a stepsize for the kth iteration. We later analize the effect of different methods for calculating this. The choice of **RULE** will distinguish the Gauss-Southwell and Randomized variants of BCGD. For the GS rule we compute

$$i_k = \arg\max_j ||\nabla_j f(x_k)||$$

This has the notable disadvantage of requiring computation of the full gradient. Taking i_k uniform random samples to be the block index give the randomized rule.

3 Synthetic Dataset

To initially very the convergence of our algorithms we create a high-dimensional dataset for the multi-class classification problem. This consists of our data matrix $A \in \mathbb{R}^{1000\times 1000}$ and target classes b. The rows of A correspond individual examples in the dataset and the columns their numeric features. Entries for A were drawn i.i.d. from a $N(0,1)$ distribution. Here $b \in \mathbb{Z}_5 0$ represents the vector of target classes corresponding to their respective rows in A. We define our starting parameter matrix $X \in \mathbb{R}^{1000\times 50}$. We can then define a error matrix with standard normal entries to inject noise into our calculation of the target classes

$$b = AX + E$$

and further process b to be the argmax of each row. This dataset has the feature of having an expensive gradient calculation due to the high number of features and output classes. This creates the potential for BCGD methods to be advantaged as ones like randomized rule may compute a smaller block of the gradient.

3.1 Performance

3.2 Stepsize Analysis

4 Real-world Dataset