Convex Optimization

Nima

Assignment 4: Group-Regularized Regression for Parkinson's Disease

In this assignment, we aim to solve a regularized regression problem that promotes sparsity—not at the level of individual predictors, but at the level of predefined groups of predictors. This is particularly useful when predictors are naturally grouped and we want to select or discard entire groups based on their relevance to the prediction task.

Objective

We are given a dataset related to Parkinson's disease and asked to predict the total symptom score (UPDRS) using voice measurements and demographic features. The goal is to:

- Implement ridge regression using stochastic gradient descent.
- Implement group LASSO using proximal gradient descent.
- Compare the performance and sparsity patterns of these methods.
- Explore acceleration techniques for faster convergence.

Dataset Description

The dataset contains:

- $N = 5785$ observations.
- $p = 18$ predictors (features), stored in `X_train.csv`.
- The target variable is the total UPDRS score, stored in `y_train.csv`.

The predictors are grouped as follows:

1. **Demographics**: `age`, `sex`
2. **Jitter Features**: `Jitter(%)`, `Jitter(Abs)`, `Jitter:RAP`, `Jitter:PPQ5`, `Jitter:DDP`
3. **Shimmer Features**: `Shimmer`, `Shimmer(dB)`, `Shimmer:APQ3`, `Shimmer:APQ5`, `Shimmer:APQ11`, `Shimmer:DDA`

4. **Noise-to-Harmonics Ratio**: `NHR`, `HNR`

5. **Nonlinear and Dynamical Features**: `RPDE`, `DFA`, `PPE`

(a) Ridge Regression

We solve the ridge regression problem:

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{2N} \|X\beta + b - y\|^2 + \lambda \|\beta\|^2$$

where b is the bias term, treated separately from the regularized parameters β.

i. Gradient Updates

The gradient of the objective function with respect to β and b are computed separately:

$$\nabla_\beta f(\beta, b) = X^\top (X\beta + b - y) + 2\lambda\beta$$

$$\nabla_b f(\beta, b) = \mathbf{1}^\top (X\beta + b - y)$$

The update rules are:

$$\beta^{(k+1)} = \beta^{(k)} - t \cdot \nabla_\beta f(\beta^{(k)}, b^{(k)})$$

$$b^{(k+1)} = b^{(k)} - t \cdot \nabla_b f(\beta^{(k)}, b^{(k)})$$

ii–iii. Implementation and Observations

The implementation was completed in the submitted code files. Based on the convergence plots for 16 different configurations, we observed:

- The step size should be moderate: too small slows convergence, too large causes oscillation. - Smaller batch sizes generally improve convergence speed. - Larger batch sizes can help stabilize training when step size is large, but at the cost of slower convergence.

(b) Group LASSO

We solve the group LASSO problem:

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{2N} \|X\beta + b - y\|^2 + \lambda \sum_{j=1}^{J} w_j \|\beta^{(j)}\|_2$$

where w_j is the weight for group j, typically set to the number of features in that group.

i. Proximal Operator

Using the group-wise proximal operator:

$$\text{prox}_{\lambda w_j \|\cdot\|_2}(v^{(j)}) = \left(1 - \frac{\lambda w_j}{\|v^{(j)}\|_2}\right)_+ v^{(j)}$$

The full proximal update is applied group-wise:

$$\beta^{(k+1)} = \text{prox}_{th}\left(\beta^{(k)} - t \cdot \nabla g(\beta^{(k)})\right)$$

ii–iii. Implementation and Group Selection

The implementation was completed in the submitted code files. Based on the output, the selected groups were:

- RPDE - Age - PPE

iv. Comparison with LASSO

Group LASSO showed better sparsity and faster convergence compared to standard LASSO. The selected groups were more interpretable and consistent with domain knowledge.

v. Accelerated Proximal Gradient

We implemented Nesterov's acceleration. The update rule for β becomes:

$$\beta^{(k+1)} = \text{prox}_{th}\left(z^{(k)} - t \cdot \nabla g(z^{(k)})\right)$$

where $z^{(k)}$ is a momentum term. The bias term b is updated separately.

vi. Observations

The accelerated algorithm converged significantly faster—approximately $10\times$ closer to the optimal value compared to the unaccelerated method. It also preserved the sparsity structure effectively.