Newton's method. Cubic Regularized Newton Methods for Logistic Regression: A Comparative Study

Dmitrii Topchii

December 16, 2024

Abstract

In recent years, higher-order methods have gained popularity for solving small-scale convex optimization problems due to their improved theoretical convergence rates and practical performance gains. In particular, the cubic regularized Newton method leverages third-order information, providing global convergence guarantees and potentially faster convergence than traditional first- and second-order methods.

In this report, the application of cubic regularized Newton methods to logistic regression is presented. The approach was implemented using the `OPTAMI` library and evaluated against a variety of baseline optimizers, including gradient descent, Nesterov acceleration, Adam, Newton's method, and L-BFGS. Convergence behavior, computational cost, parameter sensitivity, and results on both synthetic and real-world datasets are examined. The findings indicate that while cubic regularized Newton can achieve strong convergence in terms of iteration count and final loss, the higher per-iteration cost must be considered, especially for larger-scale problems.

1 Introduction

Convex optimization problems are ubiquitous in machine learning, statistics, and data analysis. Logistic regression, in particular, is a popular model for binary classification with a convex loss function. Although gradient-based methods are common due to their simplicity, they can be slow to converge. Newton's method uses second-order information to achieve potentially much faster local convergence but at a higher computational cost. Higher-order methods, such as cubic regularized Newton, incorporate third-order information to combine global convergence guarantees with fast local convergence.

1.1 Problem Statement

The logistic regression problem with ℓ_2-regularization is considered. Given data (x_i, y_i) for $i = 1, \ldots, n$ with $y_i \in \{0, 1\}$ and $x_i \in \mathbb{R}^d$, the objective is to find a parameter vector $w \in \mathbb{R}^d$ that minimizes

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} \log\bigl(1 + \exp(-y_i x_i^T w)\bigr) + \frac{\lambda}{2}\|w\|^2. \tag{1}$$

Here, $\lambda > 0$ is a regularization parameter, and $f(w)$ is convex in w.

2 Methods

Below, the optimization methods used in the experiments are described.

2.1 Gradient Descent (GD)

A first-order method that updates parameters as:

$$w_{k+1} = w_k - \eta \nabla f(w_k), \tag{2}$$

where $\eta > 0$ is a step size.

2.2 Nesterov's Accelerated Gradient (NAG)

An accelerated first-order method:

$$w_{k+1} = w_k + \beta(w_k - w_{k-1}) - \eta \nabla f\big(w_k + \beta(w_k - w_{k-1})\big). \tag{3}$$

2.3 Adam

An adaptive first-order method:

$$m_{k+1} = \beta_1 m_k + (1 - \beta_1)\nabla f(w_k), \tag{4}$$

$$v_{k+1} = \beta_2 v_k + (1 - \beta_2)[\nabla f(w_k)]^2, \tag{5}$$

$$w_{k+1} = w_k - \eta \frac{m_{k+1}/(1 - \beta_1^{k+1})}{\sqrt{v_{k+1}/(1 - \beta_2^{k+1})} + \epsilon}. \tag{6}$$

2.4 Newton's Method

A second-order method:

$$w_{k+1} = w_k - [\nabla^2 f(w_k)]^{-1} \nabla f(w_k). \tag{7}$$

2.5 L-BFGS (Quasi-Newton)

A quasi-Newton method that approximates the Hessian inverse H_k, then:

$$w_{k+1} = w_k - H_k \nabla f(w_k). \tag{8}$$

2.6 Cubic Regularized Newton (CRN)

A method that incorporates a third-order term:

$$m_k(h) = f(w_k) + \nabla f(w_k)^T h + \tfrac{1}{2} h^T \nabla^2 f(w_k) h + \tfrac{L}{6} \|h\|^3. \tag{9}$$

The update solves this cubic model approximately:

$$w_{k+1} = w_k + h_k, \tag{10}$$

where h_k approximately minimizes $m_k(h)$.

3 Experiments

These methods were tested on synthetic and Iris datasets for logistic regression. Measurements included the loss $f(w)$, final accuracy on the training set, and runtime per iteration. Sensitivity of CRN to changes in L was also examined.

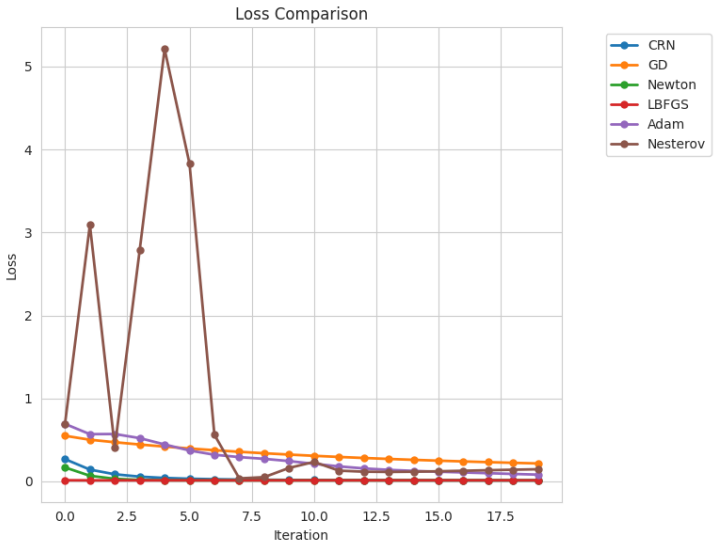

Figure 1: Loss comparison for various methods.

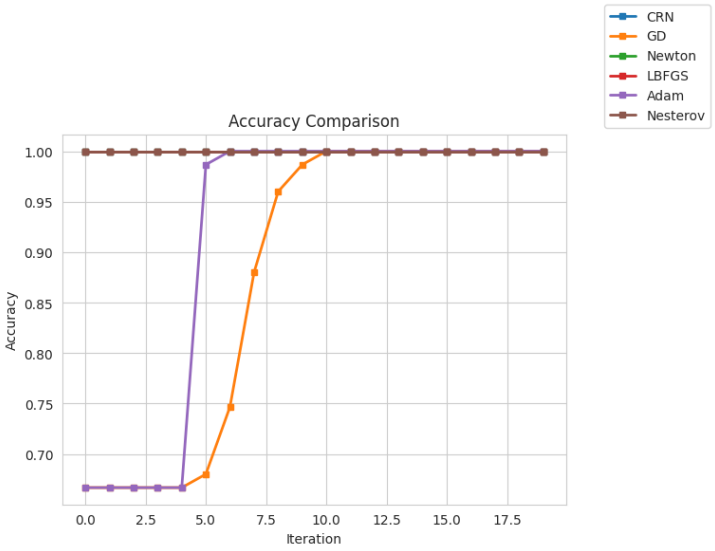

Figure 2: Accuracy comparison among methods.

4 Results

4.1 Convergence and Accuracy

Figure 1 shows loss convergence. CRN and Newton achieve very low final loss. Figure 2 shows accuracy; most methods eventually reach perfect accuracy, but convergence paths differ.

4.2 Timing and Complexity

Figure 3 shows average time per iteration. CRN is more expensive per iteration compared to first-order methods.

4.3 Sensitivity Analysis

Figure 4 shows CRN's performance for different L values.

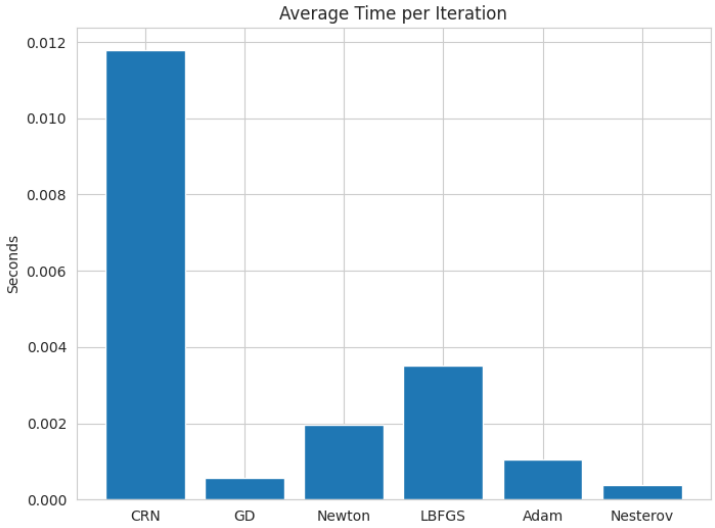

Figure 3: Average time per iteration.

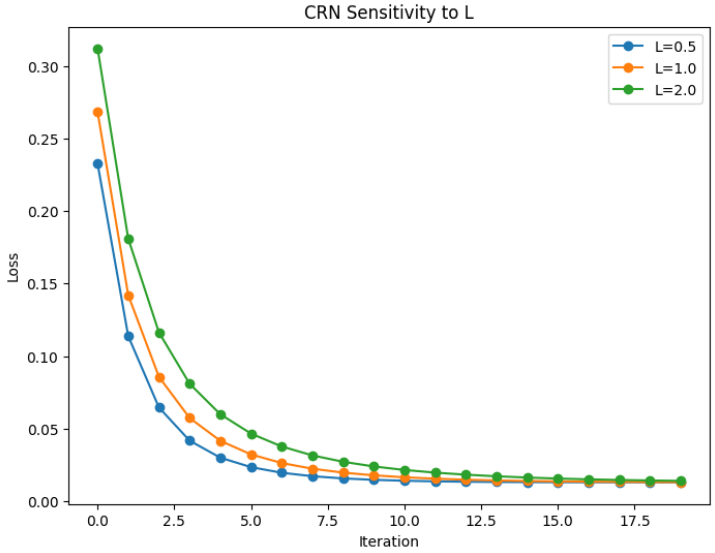

Figure 4: CRN sensitivity to the parameter L.

4.4 Summary Table

Table 1 compares final metrics.

5 Discussion and Conclusion

Cubic regularized Newton can achieve very low loss and perfect accuracy with fewer iterations, but at a higher per-iteration cost. Traditional Newton or L-BFGS methods strike a balance, while first-order methods (GD, Nesterov, Adam) are cheaper per step but may converge more slowly in terms of iterations. The choice depends on problem scale and desired precision.

Method	Final Loss	Final Accuracy	Avg Time/Iter (s)	Total Time (s)
CRN	0.0130	1.0000	0.0118	5.4052
GD	0.2162	1.0000	0.0006	0.0162
Newton	0.0129	1.0000	0.0019	0.0426
L-BFGS	0.0129	1.0000	0.0035	0.0751
Adam	0.0814	1.0000	0.0010	0.0226
Nesterov	0.1464	1.0000	0.0004	0.0086

Table 1: Final metrics for each method.