A Third-Order Majorization Algorithm for Logistic Regression With Convergence Rate Guarantees

Licheng Zhao, Wenqiang Pu, *Member, IEEE*, Rui Zhou, *Member, IEEE*, and Qingjiang Shi

***Abstract*—In this paper, we study the classical Logistic Regression (LR) problem in machine learning. Traditionally, the solving algorithms are based on either the first- or second-order approximation of the objective. For instance, the Fixed-Hessian Newton (FHN) method approximates the true Hessian with a constant estimate. In contrast, our design additionally exploits the third-order information. Applying the majorization–minimization (MM) framework, we construct a novel majorizing function based on the third-order Taylor expansion and the minimization solution is in closed-form with perseverance of the true gradient and Hessian structures. In analysis, we prove the convergence rate of the proposed algorithm. The enhanced numerical performance can be verified through simulation results.**

***Index Terms*—Convergence rate, logistic regression, majorization–minimization, third-order.**

I. Introduction

LOGISTIC Regression (LR) is a widely-used linear model for binary classification in the field of machine learning [1], [2], [3]. Given a feature vector $\mathbf{x} \in \mathbb{R}^D$ and a binary label $y \in \{0, 1\}$,[1] we linearly combine elements of $\mathbf{x}$ with a weight vector $\mathbf{w} \in \mathbb{R}^D$ and use the quantity $\mathbf{w}^T\mathbf{x}$ to fit a class label y. The only parameter in LR is a D-dimensional vector, so LR is a lightweight model and easy to deploy online [4]. Another advantage in LR is high interpretability. Once the weight vector $\mathbf{w}$ is obtained, for any feature index d, we can evaluate the ratio $|w_d x_d| / \sum_{d=1}^{D} |w_d x_d|$ for feature importance. Due to great convenience in cause tracing, LR plays a crucial role in impact factor analysis in many fields, like financial risk management [5], online transaction supervision [6], and medical disease diagnosis [7].

To illuminate the LR model, we denote the dataset as $\{\mathbf{x}_n, y_n\}_{n=1}^N$ $(N > D)$ and formulate the optimization problem as

$$\underset{\mathbf{w}}{\text{minimize}} \quad f(\mathbf{w}) = \sum_{n=1}^{N} f_n(\mathbf{w}) = \sum_{n=1}^{N} g_n\left(\mathbf{x}_n^T\mathbf{w}\right) = \sum_{n=1}^{N} \Big[-y_n \times \log\left(\sigma\left(\mathbf{x}_n^T\mathbf{w}\right)\right) - (1-y_n)\log\left(1-\sigma\left(\mathbf{x}_n^T\mathbf{w}\right)\right)\Big] \tag{1}$$

where $\sigma(x) = 1/(1+\exp(-x)) \in (0,1)$ is the Sigmoid function. This formulation is driven by the maximum likelihood estimation and the objective is known as the cross-entropy loss. Note that problem (1) is convex and thus its global optimum can be attained by polynomial-time algorithms. Existing LR solution methods are twofold: Hessian-based and non-Hessian-based. Non-Hessian-based methods only exploit the first-order information, like gradient descent, conjugate gradient [8], iterative scaling [9], [10], and dual optimization [11]. Hessian-based methods further utilize the second-order information, including Newton' s method, Quasi-Newton [12], Iteratively Re-weighted Least Squares (IRLS) [13], and Fixed-Hessian Newton (FHN) method [14]. As has been revealed in many technical reports, the introduction of Hessian brings about beneficial outcomes in numerical performance and the variants of Newton' s method beat the vanilla version empirically.

We have also studied a few closely-related algorithmic frameworks for LR solution. Subspace optimization on generalized linear models [15] works well with a superlinear convergence rate. With regard to third-order methods, Nesterov and Polyak [16] proposed the Cubic-Regularized Newton' s method (CRN) which derives a global upper bound via the Hessian Lipschitz constant. The Adaptive Regularization algorithm using Cubics (ARC) framework [17], [18] iteratively constructs a local cubic overestimator. However, the Hessian Lipschitz constant in CRN is difficult to calculate on a specific objective and the overestimator construction rule of ARC is very intricate and involves parameter tuning issues. In this paper, we focus on improving Hessian-based methods with higher-order information. We point out the weaknesses of FHN from a majorization minimization (MM) [19] perspective and enhance the performance of LR solution herein.

The Hessian-based method FHN approximates the iteration-varying Hessian $\mathbf{H}$ with a constant matrix $\hat{\mathbf{H}}$:

$$\mathbf{H} = \mathbf{X}\mathbf{A}\mathbf{X}^T \tag{2}$$

$$\text{and} \quad \hat{\mathbf{H}} = \frac{1}{4}\mathbf{X}\mathbf{X}^T, \tag{3}$$

where all $\mathbf{x}_n$'s are horizontally stacked as $\mathbf{X} \in \mathbb{R}^{D\times N}$ and $\mathbf{A} \in \mathbb{R}^{N\times N}$ is diagonal with $A_{nn} = \sigma(\mathbf{x}_n^T\mathbf{w})(1-\sigma(\mathbf{x}_n^T\mathbf{w}))$. Since $A_{nn} \le 1/4$, we naturally have $\mathbf{H} \preceq \hat{\mathbf{H}}$ and convergence of FHN is thus guaranteed. Admittedly, $\hat{\mathbf{H}}$ is a good estimate of $\mathbf{H}$ in the early stage of iteration if $\mathbf{w}$ is elementwisely initialized around zero. However, when $\mathbf{w}$ is close to the optimum, A_{nn} should approach either 0 or 1 and $\hat{\mathbf{H}}$ deviates from the true Hessian to a great extent, leading to slow convergence.

Manuscript received 23 March 2024; revised 31 May 2024; accepted 8 June 2024. Date of publication 12 June 2024; date of current version 3 July 2024. This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 62206182, Grant 62101350, and Grant 62201362, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010154, and in part by the Shenzhen Science and Technology Program under Grant RCBS20221008093126071. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christos Thrampoulidis. *(Corresponding author: Rui Zhou.)*

Licheng Zhao, Wenqiang Pu, and Rui Zhou are with the Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: zhaolicheng@sribd.cn; wpurui.zhou@sribd.cn; rui.zhou@sribd.cn).

Qingjiang Shi is with the School of Software Engineering, Tongji University, Shanghai 201804, China, and also with the Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: shiqj@tongji.edu.cn).

Digital Object Identifier 10.1109/LSP.2024.3413306

[1]The intercept term is already absorbed in $\mathbf{x}$.

The main contribution of this paper resides in a more efficient LR solution. On the algorithmic level, we construct a third-order global majorizing function and find a closed-form update preserving the true gradient and Hessian. Compared with the existing third-order methods, we develop an innovative implementation via MM and a parameter-free acceleration strategy. In analysis, we derive an $\mathcal{O}(1/k^2)$ convergence rate for the proposed algorithm where k is the iteration count. The superior numerical performance is illustrated in simulations.

II. Third-Order Majorization

In this section, we develop an iterative algorithm for the LR model. We adopt the MM framework as in [19], [20] and obtain a closed-form update scheme in every iteration.

A. Majorizing Function Construction

The majorization of $f(\mathbf{w})$ is carried out individually on each $f_n(\mathbf{w})$. The current value of $\mathbf{w}$ in iteration k is written as $\mathbf{w}^{(k)}$. Note that $f_n(\mathbf{w}) = g_n(\mathbf{x}_n^T\mathbf{w})$ by definition. We represent $\mathbf{x}_n^T\mathbf{w}$ as z_n and $\mathbf{x}_n^T\mathbf{w}^{(k)}$ as $z_n^{(k)}$, and then majorize $g_n(z_n)$ at $z_n^{(k)}$. The second-order Taylor expansion of $g_n(z_n)$ around $z_n^{(k)}$ with a Lagrange remainder is given as

$$g_n(z_n) = g_n(z_n^{(k)}) + g_n'(z_n^{(k)})(z_n - z_n^{(k)}) + \frac{1}{2}g_n''(\xi_n^{(k)})(z_n - z_n^{(k)})^2 \tag{4}$$

with $\xi_n^{(k)}$ sitting between $z_n^{(k)}$ and z_n. Since $g_n''(\xi_n^{(k)}) = \sigma(\xi_n^{(k)})(1-\sigma(\xi_n^{(k)})) \le 1/4$, one reasonable choice of majorizing function $\bar{g}_n(z_n, z_n^{(k)})$ is

$$\bar{g}_n(z_n, z_n^{(k)}) = g_n(z_n^{(k)}) + g_n'(z_n^{(k)})(z_n - z_n^{(k)}) + \frac{1}{8}(z_n - z_n^{(k)})^2.$$

Then we undo z_n and $z_n^{(k)}$ to formally present the majorizing function $\bar{f}_n(\mathbf{w}, \mathbf{w}^{(k)})$:

$$\bar{f}_n(\mathbf{w}, \mathbf{w}^{(k)}) = f_n(\mathbf{w}^{(k)}) + \left(\sigma\left(\mathbf{x}_n^T\mathbf{w}^{(k)}\right) - y_n\right) \times \mathbf{x}_n^T(\mathbf{w} - \mathbf{w}^{(k)}) + \frac{1}{8}(\mathbf{w} - \mathbf{w}^{(k)})^T \cdot \left[\mathbf{x}_n\mathbf{x}_n^T\right] \cdot (\mathbf{w} - \mathbf{w}^{(k)}).$$

Eventually, the majorizing function of $f(\mathbf{w})$, denoted by $\bar{f}(\mathbf{w}, \mathbf{w}^{(k)})$, is merely a summation of $\bar{f}_n(\mathbf{w}, \mathbf{w}^{(k)})$ and expressed as

$$\bar{f}(\mathbf{w}, \mathbf{w}^{(k)}) = f(\mathbf{w}^{(k)}) + \left(\sum_{n=1}^{N}\left(\sigma\left(\mathbf{x}_n^T\mathbf{w}^{(k)}\right) - y_n\right)\cdot\mathbf{x}_n\right)^T \times (\mathbf{w} - \mathbf{w}^{(k)}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^{(k)})^T \cdot \left[\frac{1}{4}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T\right] \cdot (\mathbf{w} - \mathbf{w}^{(k)}). \tag{5}$$

We can easily find out that $\frac{1}{4}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T = \frac{1}{4}\mathbf{X}\mathbf{X}^T = \hat{\mathbf{H}}$, so the FHN method fits in the MM framework by means of second-order majorization. To further improve FHN, we need even higher orders of approximation for greater accuracy.

Hence, we extend the Taylor expansion (4) to the third order:

$$g_n(z_n) = g_n(z_n^{(k)}) + g_n'(z_n^{(k)})(z_n - z_n^{(k)}) + \frac{1}{2}g_n''(z_n^{(k)})(z_n - z_n^{(k)})^2 + \frac{1}{6}g_n'''(\xi_n^{(k)})(z_n - z_n^{(k)})^3.$$

The third-order derivative is also bounded: $g_n'''(\xi_n^{(k)}) = \sigma(\xi_n^{(k)})(1-\sigma(\xi_n^{(k)}))(1-2\sigma(\xi_n^{(k)})) \in [-\bar{L}, \bar{L}]$ and $\bar{L} = \sqrt{3}/18$. Thus, we majorize the third-order term:

$$\frac{1}{6}g_n'''(\xi_n^{(k)})(z_n - z_n^{(k)})^3 \le \frac{\bar{L}}{6}|z_n - z_n^{(k)}|^3. \tag{6}$$

We undo z_n and $z_n^{(k)}$, and separate $\mathbf{x}_n$ and $\mathbf{w} - \mathbf{w}^{(k)}$ with a further majorization using the Cauchy-Schwarz inequality:

$$\frac{\bar{L}}{6}\left|\mathbf{x}_n^T(\mathbf{w} - \mathbf{w}^{(k)})\right|^3 \le \frac{\bar{L}}{6}\|\mathbf{x}_n\|_2^3\left\|\mathbf{w} - \mathbf{w}^{(k)}\right\|_2^3.$$

In the end, we display an improved version of $\bar{f}(\mathbf{w}, \mathbf{w}^{(k)})$ as

$$\bar{f}(\mathbf{w}, \mathbf{w}^{(k)}) = f(\mathbf{w}^{(k)}) + [\mathbf{h}^{(k)}]^T(\mathbf{w} - \mathbf{w}^{(k)}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^{(k)})^T\mathbf{H}^{(k)}(\mathbf{w} - \mathbf{w}^{(k)}) + \frac{L}{6}\|\mathbf{w} - \mathbf{w}^{(k)}\|_2^3, \tag{7}$$

where $\mathbf{h}^{(k)} = \sum_{n=1}^{N}(\sigma(\mathbf{x}_n^T\mathbf{w}^{(k)}) - y_n)\cdot\mathbf{x}_n$, $\mathbf{H}^{(k)}$ follows (2), and $L = \bar{L}\cdot\sum_{n=1}^{N}\|\mathbf{x}_n\|_2^3$. Compared to (5), $\bar{f}(\mathbf{w}, \mathbf{w}^{(k)})$ in (7) is a closer approximation of $f(\mathbf{w})$ in the neighborhood centering $\mathbf{w}^{(k)}$.

B. Minimization Solution Pursuit

Upon obtaining the majorizing function, we conduct the minimization operation at each iteration. Because the LR model is an unconstrained optimization problem, the minimization solution $\mathbf{w}^{(k+1)}$ must satisfy the zero gradient condition:

$$\mathbf{h}^{(k)} + \mathbf{H}^{(k)}(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}) + \frac{L}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|_2(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}) = \mathbf{0}.$$

Thus,

$$\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)} = -\left(\mathbf{H}^{(k)} + \frac{L\rho^{(k)}}{2}\mathbf{I}\right)^{-1}\mathbf{h}^{(k)} \tag{8}$$

and

$$\rho^{(k)} = \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|_2.$$

Taking squared ℓ_2-norm on both sides of (8), we have an equation for $\rho^{(k)}$:

$$(\rho^{(k)})^2 = \left\|\left(\mathbf{H}^{(k)} + \frac{L\rho^{(k)}}{2}\mathbf{I}\right)^{-1}\mathbf{h}^{(k)}\right\|_2^2. \tag{9}$$

With $\rho^{(k)} > 0$ and $\mathbf{H}^{(k)} \succeq \mathbf{0}$, the function of the left-hand side minus right-hand side monotonically increases in $\rho^{(k)}$ on domain $(0, \|(\mathbf{H}^{(k)})^{-1}\mathbf{h}^{(k)}\|_2)$ and 0 falls within the function range. Therefore, equation (9) yields a unique solution and this solution can be found via bisection search. The solving procedure involves eigenvalue decomposition of $\mathbf{H}^{(k)}$: $\mathbf{H}^{(k)} = \mathbf{U}^{(k)}\mathbf{\Lambda}^{(k)}(\mathbf{U}^{(k)})^T$. $\mathbf{\Lambda}^{(k)}$ is diagonal with eigenvalues arranged in descending order and the right-hand side of (9) becomes

$$\left\|\left(\mathbf{\Lambda}^{(k)} + \frac{L\rho^{(k)}}{2}\mathbf{I}\right)^{-1}\tilde{\mathbf{h}}^{(k)}\right\|_2^2 = \sum_{d=1}^{D}\frac{\left(\tilde{h}_d^{(k)}\right)^2}{\left(\lambda_d^{(k)} + \frac{L\rho^{(k)}}{2}\right)^2}.$$

where $\tilde{\mathbf{h}}^{(k)} = (\mathbf{U}^{(k)})^T\mathbf{h}^{(k)}$ and $\Lambda_{dd}^{(k)} = \lambda_d^{(k)}$. After bisection, we substitute $\rho^{(k)}$ in (8) for $\mathbf{w}^{(k+1)}$ and this minimization solution yields a decrease in objective.

Upon implementation, we may want to improve the solving procedure and decrease the objective further. With a slight abuse of notation, we denote $\tilde{\mathbf{w}}^{(k+1)} = \mathbf{w}^{(k)} - (\mathbf{H}^{(k)} +$

Algorithm 1: Third-Order Majorization Algorithm with line Search (TOMAS).

Require: Input: $\mathbf{X} \in \mathbb{R}^{D\times N}$ and $\mathbf{y} \in \mathbb{R}^{N\times 1}$. Initialization: $\mathbf{w}^{(0)}$. Line search division parameter $J = 2^{16}$. $k = 0$.

repeat
  $\mathbf{h}^{(k)} = \sum_{n=1}^{N}(\sigma(\mathbf{x}_n^T\mathbf{w}^{(k)}) - y_n)\cdot\mathbf{x}_n$;
  $\mathbf{H}^{(k)} = \mathbf{X}\mathbf{A}^{(k)}\mathbf{X}^T$ with $\mathbf{A}^{(k)}$ being diagonal and $A_{nn}^{(k)} = \sigma(\mathbf{x}_n^T\mathbf{w}^{(k)})(1-\sigma(\mathbf{x}_n^T\mathbf{w}^{(k)}))$;
  Eigenvalue decomposition: $\mathbf{H}^{(k)} = \mathbf{U}^{(k)}\mathbf{\Lambda}^{(k)}(\mathbf{U}^{(k)})^T$ with $\mathbf{\Lambda}^{(k)} = \mathrm{Diag}(\boldsymbol{\lambda}^{(k)})$;
  $\tilde{\mathbf{h}}^{(k)} = (\mathbf{U}^{(k)})^T\mathbf{h}^{(k)}$;
  Solve for $\rho^{(k)} \in (0, \|(\mathbf{H}^{(k)})^{-1}\mathbf{h}^{(k)}\|_2)$ with bisection search: $(\rho^{(k)})^2 = \sum_{d=1}^{D}(\tilde{h}_d^{(k)})^2/(\lambda_d^{(k)} + \frac{L\rho^{(k)}}{2})^2$;
  $\tilde{\mathbf{w}}^{(k+1)} = \mathbf{w}^{(k)} - \mathbf{U}^{(k)}(\tilde{\mathbf{h}}^{(k)}/(\boldsymbol{\lambda}^{(k)} + \frac{L\rho^{(k)}}{2}\mathbf{1}))$ (elementwise division);
  $\xi = \|(\mathbf{H}^{(k)})^{-1}\mathbf{h}^{(k)}\|_2/J$; Flag $= 0$;
  $\tilde{\mathbf{w}}_\xi = \mathbf{w}^{(k)} - \mathbf{U}^{(k)}(\tilde{\mathbf{h}}^{(k)}/\boldsymbol{\lambda}^{(k)})$;
  while $\xi \le \|(\mathbf{H}^{(k)})^{-1}\mathbf{h}^{(k)}\|_2$ **do**
    if $f(\tilde{\mathbf{w}}_\xi) \le f(\tilde{\mathbf{w}}^{(k+1)})$ **then**
      Flag $= 1$; **Break**;
    end if
    $\tilde{\mathbf{w}}_\xi = \mathbf{w}^{(k)} - \mathbf{U}^{(k)}(\tilde{\mathbf{h}}^{(k)}/(\boldsymbol{\lambda}^{(k)} + \frac{L\xi}{2}\mathbf{1}))$; $\xi = 2\xi$;
  end while
  $\mathbf{w}^{(k+1)} = \begin{cases}\tilde{\mathbf{w}}^{(k+1)} & \text{Flag} = 0\\ \tilde{\mathbf{w}}_\xi & \text{Flag} = 1\end{cases}$;
  $k = k+1$;
until convergence

$\frac{L\rho^{(k)}}{2}\mathbf{I})^{-1}\mathbf{h}^{(k)}$. We consider an objective-oriented line search for a further decrement with respect to $f(\tilde{\mathbf{w}}^{(k+1)})$. The potential update is in the form of $\mathbf{w}^{(k)} - (\mathbf{H}^{(k)} + \frac{L\xi}{2}\mathbf{I})^{-1}\mathbf{h}^{(k)}$ and the line search is conducted on ξ within $(0, \|(\mathbf{H}^{(k)})^{-1}\mathbf{h}^{(k)}\|_2)$ in a geometric series. If a further decrement is observed, we set the potential update as $\mathbf{w}^{(k+1)}$; otherwise, set $\mathbf{w}^{(k+1)} = \tilde{\mathbf{w}}^{(k+1)}$. To this point, we summarize the proposed algorithm in Algorithm 1, named TOMAS.

III. Convergence Analysis

In this section, we provide the convergence analysis of the proposed algorithm TOMAS. Specifically, we study the convergence rate in terms of optimality gap. We assume that the LR problem is well defined and the sublevel set of $f(\mathbf{w})$ is bounded, i.e., ($\mathbf{w}^\star$ is the global optimal solution)

$$\|\mathbf{w} - \mathbf{w}^\star\|_2 \le S,\ \forall \mathbf{w} : f(\mathbf{w}) \le f\left(\mathbf{w}^{(0)}\right). \tag{10}$$

This assumption holds true for inseparable datasets as well as separable datasets with ℓ_2 regularization in $f(\mathbf{w})$. The convergence rate is summarized in the theorem below.

Theorem 1: Suppose the sublevel set of $f(\mathbf{w})$ is bounded like (10). The convergence rate of Algorithm 1 in terms of optimality gap is given by

$$f(\mathbf{w}^{(k)}) - f(\mathbf{w}^\star) \le \frac{\sqrt{3}S^3\cdot\sum_{n=1}^{N}\|\mathbf{x}_n\|_2^3}{2\left(k+3\sqrt{3}-1\right)^2} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{11}$$

Proof: To start with, we bound the residual of $f(\mathbf{w}) - \bar{f}(\mathbf{w}, \mathbf{w}^{(k)})$. Notice that the complete inequality relation of (6) is $-\frac{\bar{L}}{6}|z_n - z_n^{(k)}|^3 \le \frac{1}{6}f_n'''(\xi_n^{(k)})(z_n - z_n^{(k)})^3 \le \frac{\bar{L}}{6}|z_n - z_n^{(k)}|^3$, so it holds that

$$\left|f(\mathbf{w}) - \bar{f}(\mathbf{w}, \mathbf{w}^{(k)})\right| \le \frac{L}{3}\left\|\mathbf{w} - \mathbf{w}^{(k)}\right\|_2^3.$$

As a result,

$$f\left(\mathbf{w}^{(k+1)}\right) \le f\left(\tilde{\mathbf{w}}^{(k+1)}\right) = \min_{\mathbf{w}}\bar{f}(\mathbf{w}, \mathbf{w}^{(k)})$$
$$\le \min_{\mathbf{w}}\left[f(\mathbf{w}) + \frac{L}{3}\left\|\mathbf{w} - \mathbf{w}^{(k)}\right\|_2^3\right].$$

By the assumption of a bounded sublevel set, we further have

$$f\left(\mathbf{w}^{(1)}\right) \le f(\mathbf{w}^\star) + \frac{L}{3}S^3. \tag{12}$$

According to the MM framework, it is obvious that $f(\mathbf{w}^{(k+1)}) \le f(\mathbf{w}^{(k)}), \forall k$. In the following, we link $f(\mathbf{w}^{(k+1)})$ and $f(\mathbf{w}^{(k)})$ with an inequality chain. Let $\mathbf{w}_\beta^{(k)} = \beta\mathbf{w}^\star + (1-\beta)\mathbf{w}^{(k)}$ and $\beta \in [0,1]$. It can be established that

$$f\left(\mathbf{w}^{(k+1)}\right) \le f\left(\mathbf{w}_\beta^{(k)}\right) + \frac{L}{3}\left\|\mathbf{w}_\beta^{(k)} - \mathbf{w}^{(k)}\right\|_2^3$$
$$\overset{(a)}{\le} f(\mathbf{w}^{(k)}) - \beta\left(f(\mathbf{w}^{(k)}) - f(\mathbf{w}^\star)\right) + \frac{L}{3}\beta^3S^3.$$

where (a) is because $f(\mathbf{w}_\beta^{(k)}) - f(\mathbf{w}^{(k)}) \le \beta(f(\mathbf{w}^\star) - f(\mathbf{w}^{(k)}))$ due to convexity of $f(\mathbf{w})$, and $\|\mathbf{w}^\star - \mathbf{w}^{(k)}\|_2 \le S$. Minimizing the right-hand side with respect to $\beta \in [0,1]$, we obtain

$$f\left(\mathbf{w}^{(k+1)}\right) - f(\mathbf{w}^{(k)}) \le -\frac{2}{3}\frac{\left(f(\mathbf{w}^{(k)}) - f(\mathbf{w}^\star)\right)^{3/2}}{\sqrt{LS^3}} \tag{13}$$

with

$$0 \le \beta = \sqrt{\frac{f(\mathbf{w}^{(k)}) - f(\mathbf{w}^\star)}{LS^3}} \le \sqrt{\frac{f\left(\mathbf{w}^{(1)}\right) - f(\mathbf{w}^\star)}{LS^3}} \le 1.$$

Finally, we derive the convergence rate. Let $r_k = f(\mathbf{w}^{(k)}) - f(\mathbf{w}^\star)$ be the optimality gap, and thus

$$\frac{1}{\sqrt{r_{k+1}}} - \frac{1}{\sqrt{r_k}} = \frac{r_k - r_{k+1}}{\sqrt{r_k r_{k+1}}\left(\sqrt{r_k} + \sqrt{r_{k+1}}\right)}$$
$$\overset{(a)}{\ge} \frac{2}{3\sqrt{LS^3}}\frac{r_k^{3/2}}{\sqrt{r_k r_{k+1}}\left(\sqrt{r_k} + \sqrt{r_{k+1}}\right)} \ge \frac{1}{3\sqrt{LS^3}}$$

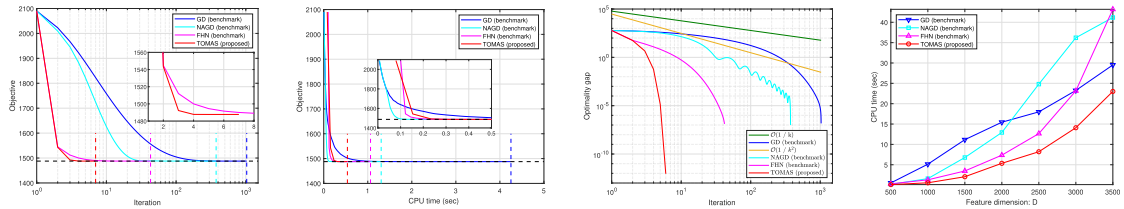
where (a) is because of (13): $r_k - r_{k+1} \ge \frac{2}{3}\frac{r_k^{3/2}}{\sqrt{LS^3}}$. Therefore,

$$\frac{1}{\sqrt{r_k}} \ge \frac{1}{\sqrt{r_1}} + \frac{k-1}{3\sqrt{LS^3}} \overset{(a)}{\ge} \frac{1}{\sqrt{LS^3}}\left(\sqrt{3} + \frac{k-1}{3}\right)$$
$$= \frac{1}{\sqrt{LS^3}}\cdot\frac{k+3\sqrt{3}-1}{3},\ \forall k,$$

where (a) is due to (12). Substitute $\sqrt{3}/18\cdot\sum_{n=1}^{N}\|\mathbf{x}_n\|_2^3$ for L, and we get (11). ■

IV. Numerical Simulations

In this section, we present numerical results on both synthetic and real-data experiments to illustrate the performance of Algorithm TOMAS. All simulations are performed on a PC with



(a) Objective versus iteration. (b) Objective versus CPU time. (c) Optimality gap versus iteration. (d) CPU time versus different choices of feature dimension D.

Fig. 1. Numerical performance on synthetic datasets.

TABLE I
NUMERICAL PERFORMANCE ON REAL DATASETS. THE EVALUATION METRICS ARE TRAIN/TEST ACCURACY AND AVERAGE CPU TIME.

Dataset	GD		NAGD		FHN		TOMAS	
	Train/Test accuracy	Time (s)	Train/Test accuracy	Time (s)	Train/Test accuracy	Time (s)	Train/Test accuracy	Time (s)
australian	87.08%/88.10%	0.6117	87.08%/88.10%	0.1485	87.08%/88.10%	0.0098	87.08%/88.10%	0.0061
breast-cancer	97.20%/97.27%	0.0534	97.20%/97.27%	0.0109	97.20%/97.27%	0.0152	97.20%/97.27%	0.0059
mushroom	99.95%/99.85%	0.6735	100%/99.85%	0.4211	100%/99.85%	1.7060	100%/99.85%	0.1495
phishing	94.27%/93.94%	7.0165	94.27%/93.94%	0.4391	94.27%/93.94%	0.7315	94.27%/93.94%	0.1850

a 2.90 GHz i7-10700 CPU and 16.0 GB RAM. The compared algorithms are listed as follows:

- ***GD*** *(benchmark):* Gradient descent. Complexity:[2] $\mathcal{O}(DN)$. Backtracking line search [21] is applied to ensure monotonicity.
- ***NAGD*** *(benchmark):* Nesterov Accelerated Gradient Descent [22], [23]. Complexity: $\mathcal{O}(DN)$. Line search is applied for a larger step size rather than monotonicity.
- ***FHN*** *(benchmark):* a second-order method. Complexity: $\mathcal{O}(DN + D^2)$. Line search is not applicable.
- ***TOMAS*** *(proposed):* a third-order algorithm with line search. Complexity: $\mathcal{O}(DN + D^2N + D^3)$.

A. Synthetic Experiments

Let $D = 1000$ and $N = 3000$. We generate one instance of $\mathbf{X}$ and $\mathbf{y}$: $\mathbf{X}$ is drawn columnwisely from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{y}$ elementwisely from a Bernoulli distribution $B(1, 0.5)$. We present the convergence property of the compared algorithms in Fig. 1. Fig. 1(a) shows the convergence curves of GD, NAGD, FHN, and TOMAS. The horizontal black dashed line indicates that all three algorithms converge to the same objective value. This phenomenon matches our expectation because the LR model produces a convex problem and all optimization algorithms should reach the same optimal value. With regard to computational efficiency, we can observe in Figs. 1(a) and 1(b) that the proposed algorithm TOMAS takes the fewest iterations and least CPU time upon convergence. In detail, TOMAS converges within 10 iterations and 0.6 seconds while FHN, NAGD, and GD take (40 iterations, 1 second), (400 iterations, 1.3 seconds), and (1000 iterations, 4.2 seconds) to complete the whole process. Note that the empirical convergence speed could be faster than the worst-case theoretical rate derived above, cf. Fig. 1(c).

To further compare computational efficiency, we run the algorithms in different scenarios with multiple realizations. We vary D in $\{500, 1000, 1500, \ldots, 3500\}$ and set $N = 3D$. The reported performance is averaged over 100 random instances of $\mathbf{X}$ and $\mathbf{y}$. Fig. 1(d) reflects the running time of the compared algorithms under different choices of feature dimension D. Obviously, TOMAS is the most efficient across the entire range of D. Quantitatively, TOMAS shortens the running time of GD, NAGD, and FHN by 60%, 55%, and 40% in average. So numerically, TOMAS is superior to the other studied benchmarks if implemented properly.

B. Real Data Experiments

In real data experiments, we choose four binary classification datasets named australian, breast-cancer, mushroom, and phishing.[3] The detailed information of the chosen datasets are provided in the table below ("#" means number). The comparison results are elaborated in Table I. In real-world datasets, the LR objective may not have a bounded sublevel set and the converged variables could be slightly different for miscellaneous algorithms. Whenever the Hessian matrix is close to singular, the Moore–Penrose inverse is alternatively applied. The proposed algorithm TOMAS achieves as good classification performance as the other compared benchmarks but is the most computationally efficient. In detail, TOMAS can accelerate the benchmarks by 30% at a conservative estimate.

Dataset	#features	#training samples	#test samples
australian	14	480	210
breast-cancer	10	500	183
mushroom	112	4062	4062
phishing	68	8000	3055

V. CONCLUSION

In this paper, we have investigated into the LR optimization problem. Unlike the traditional solving algorithms which only utilize the first- or second-order information of the objective, we have considered the third-order derivatives in extra. The proposed algorithm is named TOMAS and based on the MM algorithmic framework. We have derived a novel majorizing function using the third-order Taylor expansion. The minimization solution is in closed-form and includes the information of the true gradient and Hessian. Moreover, we have provided an analysis on the convergence rate of TOMAS, which is $\mathcal{O}(1/k^2)$. Numerical experiments have shown the superior performance of the proposed algorithm compared with the other studied benchmarks.

[2]Computation complexity is assessed on a per-iteration basis.

[3]http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

REFERENCES

[1] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4, Berlin, Germany: Springer, 2006.

[2] F. Abramovich and V. Grinshtein, "High-dimensional classification by sparse logistic regression," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3068–3079, May 2019.

[3] R. Wang, N. Xiu, and C. Zhang, "Greedy projected gradient-Newton method for sparse logistic regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 527–538, Feb. 2020.

[4] P. Harrington, *Machine Learning in Action*. USA: Manning Publications Co., 2012.

[5] P. Christoffersen, *Elements of Financial Risk Management*. Cambridge, MA, USA: Academic, 2011.

[6] S. Dhankhad, E. Mohammed, and B. Far, "Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study," in *Proc. IEEE Int. Conf. Inf. Reuse Integration*, 2018, pp. 122–125.

[7] R. Xiao, X. Cui, H. Qiao, X. Zheng, and Y. Zhang, "Early diagnosis model of Alzheimer's disease based on sparse logistic regression," *Multimedia Tools Appl.*, vol. 80, pp. 3969–3980, 2021.

[8] B. T. Polyak, "The conjugate gradient method in extremal problems," *USSR Comput. Math. Math. Phys.*, vol. 9, no. 4, pp. 94–112, 1969.

[9] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *Proc. Workshop Mach. Learn. Inf. Filtering*, vol. 1, no. 1, 1999, pp. 61–67.

[10] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, adaboost and Bregman distances," *Mach. Learn.*, vol. 48, pp. 253–285, 2002.

[11] T. S. Jaakkola and D. Haussler, "Probabilistic kernel regression models," in *Proc. 7th Int. Workshop Artif. Intell. Statist.*, 1999. [Online]. Available: https://proceedings.mlr.press/r2/jaakkola99a.html

[12] J. E. Dennis Jr and J. J. Moré, "Quasi-Newton methods, motivation and theory," *SIAM Rev.*, vol. 19, no. 1, pp. 46–89, 1977.

[13] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math. A J. Issued Courant Inst. Math. Sci.*, vol. 63, no. 1, pp. 1–38, 2010.

[14] D. Böhning, "The lower bound method in probit regression," *Comput. Statist. Data Anal.*, vol. 30, no. 1, pp. 13–17, 1999.

[15] G. Narkiss and M. Zibulevsky, *Sequential Subspace Optimization Method for Large-Scale Unconstrained Problems*, Technion - Israel Institute of Technology, Tech. Rep., 2005.

[16] Y. Nesterov and B. T. Polyak, "Cubic regularization of Newton method and its global performance," *Math. Program.*, vol. 108, no. 1, pp. 177–205, 2006.

[17] C. Cartis, N. I. Gould, and P. L. Toint, "Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results," *Math. Program.*, vol. 127, no. 2, pp. 245–295, 2011.

[18] C. Cartis, N. I. Gould, and P. L. Toint, "Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity," *Math. Program.*, vol. 130, no. 2, pp. 295–319, 2011.

[19] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[20] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[21] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[22] Y. E. Nesterov, "A method of solving a convex programming problem with convergence rate O $(1/k^2)$," *Doklady Akademii Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.

[23] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.