Convergence analysis of an adaptively regularized natural gradient method

Jiayuan Wu, Jiang Hu, Hongchao Zhang, and Zaiwen Wen

***Abstract*—In this paper, we study the convergence properties of the natural gradient methods. By reviewing the mathematical condition for the equivalence between the Fisher information matrix and the generalized Gauss-Newton matrix, as well as the comparisons on the computation and storage, we reveal the popularity of the natural gradient method. To ensure the global convergence, an adaptively regularized natural gradient method is proposed. By requiring sufficient probabilistic accurate estimations on both the function and the gradient evaluations, we establish the almost sure convergence. In the local convergence, we employ the local error bound condition and show the convergence rate can be quadratic by adding mild assumptions on the stochastic estimates of gradients and Fisher information matrices. Preliminary numerical experiments on the regularized logistic regression are performed to support our findings.**

***Index Terms*—Fisher information matrix, natural gradient method, adaptive regularization, local error bound, quadratic convergence rate**

I. Introduction

We consider the optimization problem

$$\min_{\theta \in \mathbb{R}^n} h(\theta) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i, \theta)), \tag{1}$$

where $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^m$ is a set of data points satisfying $(x_i, y_i) \sim Q_{x,y}$ with the true data distribution $Q_{x,y}$ and corresponding density $q(x, y) = q(x)q(y|x)$, $f(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R}^m$ is the input-output mapping with parameter θ, and L is the single-data loss function. We mainly focus on the negative log-probability loss function

$$L(y, f(x, \theta)) = -\log p(y|f(x, \theta)), \tag{2}$$

where $p(y|f(x, \theta))$ is the density function of y conditioning on $f(x, \theta)$. However, our analysis can be applied to other loss functions as long as the required conditions and assumptions are satisfied. The connection between several loss functions L and its corresponding conditional distribution are established in [1], e.g., the square loss and standard Gaussian distribution, cross-entropy loss and the multinomial distribution. The equivalence between the negative log probability loss and Kullback-Leibler divergence is shown in [1] as well.

A. Literature review

Problem (1) is ubiquitous in deep learning [2], [3], reinforcement learning [4], signal processing [5], [6], and quantum physics/chemistry [7], [8]. For a multi-layer feed-forward neural network, $f(x, \theta)$ is the output of the last layer with respect to the input x and the network parameters θ. For the loss L, it is usually set to the cross-entropy loss for the image classification task. Various algorithms have been developed to solve (1). The first-order methods include the stochastic gradient method [9], the stochastic variance-reduced gradient method [10], SAGA [11] and adaptive gradient methods [12], [13]. We refer to the book [14] for more details. By exploiting the log-probability structure of the loss function, an efficient natural gradient method (NGM) using the information geometry of the parameter space is initially proposed in [15]. A Fisher's method of scoring based on the full-batch gradient and the exact Fisher information matrix is also presented in [16]. Later, it is shown in [2], [17]–[21] that the natural gradient-type method can outperform the stochastic gradient-type methods when tackling large-scale learning problems. Approximate Newton and quasi-Newton methods [22]–[26] have been developed to achieve faster convergence than stochastic gradient-type methods. Compared with these methods, the natural gradient-type methods are more suitable to solve large-scale learning problems in terms of computation and efficiency, especially when the Kronecker factored approximations are used.

The convergence of the stochastic gradient-type methods are extensively studied in [27]–[34]. In the nonconvex case, the Lipschitz continuity of ∇h and bounded variance are standard assumptions for the almost sure convergence of the gradient norm. However, the theoretical bound in their analysis with these assumptions suggests slower convergence than the empirical performance. The Polyak-Łojasiewicz condition proposed in [35] is utilized [36] to prove the linear convergence rate of the stochastic gradient method. Recently, the Kurdyka-Łojasiewicz inequality [37] is also investigated to derive the convergence of the iteration sequence, as well as the convergence rate. As to the natural gradient method, the linear convergence from random initialization has been shown in [38] for an over-parameterized neural network model under an additional stable Jacobian condition. It is well-known that Newton-type methods, such as the Gauss-Newton

Jiayuan Wu is with College of Engineering, Peking University, Beijing, China. Email: 1901110043@pku.edu.cn.

Jiang Hu is with Massachusetts General Hospital and Harvard Medical School, Harvard University, Boston, MA 02114. Email: hujiangopt@gmail.com.

Hongchao Zhang is with Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918. Email: hozhang@math.lsu.edu. Hongchao Zhang is supported in part by NSF DMS 2110722, 2309549.

Zaiwen Wen is with Beijing International Center for Mathematical Research, Center for Machine Learning Research and College of Engineering, Peking University, Beijing, China. Email: wenzw@pku.edu.cn. Zaiwen Wen is supported in part by the NSFC grant 11831002.

method and the Levenberg–Marquardt method, enjoy locally superlinear or quadratic convergence rate [39, Subsection 10.3] for deterministic optimization.

B. Contribution

Although the practicability of the NGM has been verified in a wide range of applications, its theoretical property has not been well understood. The goal of this paper is to derive global convergence of the NGM through an adaptive stochastic trust region framework and establish its fast local convergence by utilizing the regularity conditions of h and the connections between the Fisher information matrix (FIM) and the Hessian of the objective function. We first review several definitions of the FIMs and establish their connections to the generalized Gauss-Newton (GGN) matrix. Both mathematical and computational comparisons between the FIMs and the GGN matrix reveal the popularity of the NGM. Furthermore, the contributions of this paper are summarized as follows.

- A strategy of adaptive regularization in the stochastic trust region framework is proposed to ensure the global convergence of NGM. Our main contributions in the analysis lie in the generalization of the results of stochastic trust region method in [40] from the trust-region constraint to the adaptive regularization for solving the optimization problem (1). Besides, with the assumption of sufficiently probabilistic accurate estimations on the objective function values, we have weakened the condition of the acceptance of the iterates, i.e., removed the dependency on the gradient norms.
- We investigate the local error bound condition instead of the locally strong convexity in the analysis of the locally quadratic convergence rate of NGM. With two stochastic conditions on the estimates of function values and gradients, we prove the locally quadratic convergence of NGM. The key tools exploited here are the equivalence between the FIM and GGN matrices, and the perturbation analysis used in the eigenvalue and singular value decompositions under the local error bound condition. The quadratic convergence rate of the iterates achieved in this work is significantly stronger than the linear convergence rate of the outputs reported in [38, Theorem 2].

Notation. For any $n \in \mathbb{N}$, we use the abbreviation $[n] := \{1, \ldots, n\}$. For a vector $x \in \mathbb{R}^n$, we use $\|x\|$ to denote its ℓ_2 norm. For a matrix $X \in \mathbb{R}^{m\times n}$, $\|X\|$ and $\|X\|_F$ are defined as the spectral norm and the Frobenius norm, respectively. The notations $X \succeq 0$ and $X \succ 0$ denote the sets of positive semidefinite and positive definite matrices, respectively.

Organization. In Section II, we give the definitions of FIMs and the GGN, and clarify their connections. A globalized NGM together with almost sure convergence in the gradient norm is presented in Section III. In Section IV, we show the local quadratic convergence rate of the iterates. Numerical experiments on the logistic regression problem are reported in Section V.

II. Relations between the Hessian, the GGN matrix, and the FIM

The goal of this section is to review when the FIM may serve as a good approximation of the Hessian matrix of h. The key is to establish the connection with the GGN matrix. As the GGN matrix is computationally expensive in the large-scale optimization setting, the NGM based on FIM gains much attention due to its tractable computation. We assume in the following that h is sufficiently regular, which holds if the loss function L and the input-output mapping function f and $\log p$ are sufficiently regular, such that the corresponding Hessian and the FIM are well-defined.

A. Computation of the Hessian matrix and the GGN matrix

Let $\nabla_v h$ and $\nabla_v^2 h$ denote the gradient and Hessian of a real-valued function h with respect to a variable v, respectively. The j-th component of $f(x,\theta) \in \mathbb{R}^m$ is expressed as $f_j(x,\theta)$. The Jacobian matrix $J_{f(x,\theta)}(\theta)$ of $f(x,\theta)$ with respect to θ is defined as

$$J_{f(x,\theta)}(\theta) = \nabla_\theta f(x,\theta)^\top = \begin{bmatrix} \nabla_\theta f_1(x,\theta)^\top \\ \cdots \\ \nabla_\theta f_m(x,\theta)^\top \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

For $\{(x_i,y_i)\}_{i=1}^N$, we denote

$$J(\theta) = \begin{bmatrix} J_{f(x_1,\theta)}(\theta) \\ \vdots \\ J_{f(x_N,\theta)}(\theta) \end{bmatrix} \in \mathbb{R}^{mN\times n}$$

and

$$\mathcal{G}(\theta) = \begin{bmatrix} \nabla_z L(y_1,z)|_{z=f(x_1,\theta)} \\ \vdots \\ \nabla_z L(y_N,z)|_{z=f(x_N,\theta)} \end{bmatrix} \in \mathbb{R}^{mN}.$$

We also define two block diagonal matrices:

$$\mathcal{H}_L(\theta) = \begin{bmatrix} H_1(\theta) & & \\ & \ddots & \\ & & H_N(\theta) \end{bmatrix} \in \mathbb{R}^{mN\times mN}$$

and

$$\bar{\mathcal{H}}_L(\theta) = \begin{bmatrix} \bar{H}_1(\theta) & & \\ & \ddots & \\ & & \bar{H}_N(\theta) \end{bmatrix} \in \mathbb{R}^{mN\times mN},$$

where $H_i(\theta) = \nabla_z^2 L(y_i,z)|_{z=f(x_i,\theta)} \in \mathbb{R}^{m\times m}$ and $\bar{H}_i(\theta) = \nabla_z L(y_i,z)|_{z=f(x_i,\theta)} \nabla_z L(y_i,z)^\top|_{z=f(x_i,\theta)} \in \mathbb{R}^{m\times m}$.

By the chain rule, the Hessian matrix of h is given by

$$\nabla_\theta^2 h(\theta) = \underbrace{\frac{1}{N}\sum_{i=1}^N \left[J_{f(x_i,\theta)}^\top(\theta)\ \nabla_z^2 L(y_i,z)\big|_{z=f(x_i,\theta)} J_{f(x_i,\theta)}(\theta)\right]}_{:=H^{\mathcal{GN}}(\theta)\in\mathbb{R}^{n\times n}} + \frac{1}{N}\sum_{i=1}^N \left[\sum_{j=1}^m \nabla_\theta^2[f_j(x_i,\theta)]\ \nabla_{z_j} L(y_i,z)\big|_{z=f(x_i,\theta)}\right]. \tag{3}$$

Here, $H^{\mathcal{GN}}(\theta)$ is called the GGN matrix [41]. Based on the previous notations, we have

$$H^{\mathcal{GN}}(\theta) = \frac{1}{N} J(\theta)^\top \mathcal{H}_L(\theta) J(\theta). \tag{4}$$

When θ^* is a local minimum with $\nabla_z L(y_i, z)|_{z=f(x_i,\theta^*)} \approx 0$ for each (x_i, y_i) (which happens if each training pair (x_i, y_i) is fitted accurately by the network f), the GGN matrix can serve as a good approximation of $\nabla^2_\theta h(\theta)$.

B. Computation of various FIMs

By [42], the FIM of h defined (1) with the loss function (2) is defined as

$$F(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{y\sim p_{y|x_i}(\theta)} \left[\nabla_\theta \log p(y|x_i,\theta) \nabla_\theta \log p(y|x_i,\theta)^\top\right] \tag{5}$$

where $p_{y|x_i}(\theta) := p(y|f(x_i,\theta))$ and $p(y|x_i,\theta) = p(y|f(x_i,\theta))$. Then, by chain rule we have

$$\nabla_\theta \log p(y|x_i,\theta) = \nabla_\theta f(x_i,\theta) \nabla_z \log p(y|z)|_{z=f(x_i,\theta)} \in \mathbb{R}^n. \tag{6}$$

Using $L(y,z) = -\log p(y|z)$ by (2) and changing the order of taking the integral and the derivatives, one has

$$\begin{aligned} &\mathbb{E}_{y\sim p_{y|x_i}(\theta)} \left[\nabla_z \log p(y|z)|_{z=f(x_i,\theta)} \nabla_z \log p(y|z)^\top|_{z=f(x_i,\theta)}\right] \\ =&\mathbb{E}_{y\sim p_{y|x_i}(\theta)} \left[\tilde{H}_i\right] =: \hat{H}_i(\theta), \end{aligned} \tag{7}$$

where $\tilde{H}_i = \nabla^2_z L(y,z)|_{z=f(x_i,\theta)}$ and the first equality holds because

$$\begin{aligned} &\mathbb{E}_{y\sim p(y|z)}[-\nabla^2_z \log p(y|z)] \\ =&\int \frac{\nabla p(y|z) \nabla p(y|z)^\top - p(y|z)\nabla^2_z p(y|z)}{p^2(y|z)} p(y|z)\mathrm{d}z \\ =&\mathbb{E}_{y\sim p(y|z)}[\nabla_z \log p(y|z) \nabla_z \log p(y|z)^\top] - \nabla^2_z \int p(y|z)\mathrm{d}z \\ =&\mathbb{E}_{y\sim p(y|z)}[\nabla_z \log p(y|z) \nabla_z \log p(y|z)^\top]. \end{aligned}$$

Then, by plugging (6) and (7) into (5), an equivalent formulation of F is given by

$$\begin{aligned} F(\theta) &= \frac{1}{N}\sum_{i=1}^{N} J^\top_{f(x_i,\theta)}(\theta) \hat{H}_i(\theta) J_{f(x_i,\theta)}(\theta) \\ &= \frac{1}{N} J(\theta)^\top \mathcal{H}(\theta) J(\theta), \end{aligned} \tag{8}$$

where $\mathcal{H}(\theta)$ is a block diagonal matrix with i-th block being $\hat{H}_i(\theta)$.

When the conditional expectation $\mathbb{E}_{y\sim p_{y|x_i}(\theta)}$ does not have an explicit form, a sampling approach can be used to approximate this expectation. Specifically, for each x_i, we can sample y from the density $P_{y|x_i}(\theta)$ multiple times to get $y_i^1, \cdots, y_i^{N^y}$ with $N^y \in \mathbb{N}$. In addition, a minibatch $\mathcal{B}^F \subset [N]$ can also be sampled to further reduce computation. Thus, we can obtain the minibatch FIM given by

$$\tilde{F}(\theta) = \frac{1}{b^F}\sum_{i\in\mathcal{B}^F}\sum_{j=1}^{N^y} \nabla_\theta \log p(y_i^j, f(x_i,\theta)) \nabla_\theta \log p(y_i^j, f(x_i,\theta))^\top \tag{9}$$

with $b^F = |\mathcal{B}^F| \cdot N^y$ and $|\mathcal{B}^F|$ being the cardinality of $\mathcal{B}^F$.

Note that the minibatch FIM $\tilde{F}$ involves resampling to approximate $\mathbb{E}_{y\sim P_{y|x_i}(\theta)}$. By utilizing the observed data points $\{(x_i, y_i)\}_{i=1}^N$, the Empirical FIM (EFIM) is proposed in [41] for practical purpose, namely,

$$\begin{aligned} \bar{F}(\theta) =& \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \log p(y_i, f(x_i,\theta)) \nabla_\theta \log p(y_i, f(x_i,\theta))^\top \\ =&\frac{1}{N}\sum_{i=1}^{N} J^\top_{f(x_i,\theta)}(\theta) \bar{H}_i(\theta) J_{f(x_i,\theta)}(\theta) = \frac{1}{N} J(\theta)^\top \bar{\mathcal{H}}_L(\theta) J(\theta). \end{aligned} \tag{10}$$

Different from the FIM, the definition of EFIM does not rely on the exact expression of p. Hence, the EFIM can also be defined for loss functions L, which do not obey the form of negative log probability [1, Section 11.1]. Analogous to the minibatch FIM, the minibatch sampling of $[N]$ in (10) can also be utilized to approximate $\bar{F}(\theta)$ and reduce the computations.

C. Connections between FIMs and the GGN matrix

From (4), (8), and (10), we see the differences between FIMs and the GGN matrix lie in three matrices, $\mathcal{H}(\theta)$, $\bar{\mathcal{H}}_L(\theta)$ and $\mathcal{H}_L(\theta)$. It follows the definition of $H_i(\theta)$ that if $\nabla^2_z L(y,z)$ does not depend on y, we will have $\hat{H}_i(\theta) = \nabla^2_z L(y_i, z)|_{z=f(x_i,\theta)}$, and in this case, $F(\theta) = H^{\mathcal{GN}}(\theta)$. It is noted in [1] that this independence condition will be satisfied by the loss L from the standard Gaussian distribution and the multinomial distribution. For more general distributions, one can refer to [1, Section 9.2]. As pointed out in [43], since y_i may not be sampled from the predictive distribution $p(y|x_i,\theta)$, the EFIM is not a Monte Carlo estimate of the FIM and the equivalence between the EFIM and the FIM relies on strong assumptions, e.g., a correct model f and enough data relative to model capacity. In the case that $y_i \approx f(x_i,\theta)$, the EFIM goes to zero while the FIM and the GGN matrix approach the Hessian. They also explain the practical success of the EFIM based methods from the perspective of variance adaptation.

Although the FIM coincides with the GGN matrix mathematically under the above-mentioned independence condition, the computation of the FIM only involves the gradient of h and the expectation, which can be obtained without formulating the Jacobian. In particular, for f from the deep neural network applications, the explicit storage of the Jacobian is costly and not available in pytorch and tensorflow. These tools often provide the access to the Jacobian vector products but the cost is still expensive when the batch size is large. In the construction of mini-batch FIM and EFIM, we only need to compute the gradients on the resampled or observed data points, which can be efficiently calculated through the back propagation. These comparisons are summarized in Table I, while the connections between different FIMs and the GGN matrix are presented in Figure 1. Due to tractability of the computation, the minibatch FIM and the EFIM are two popular approximations widely used in the literature for solving deep learning problems, see, e.g., [2], [3], [44]. However, the convergence properties of FIM based stochastic methods are not well explored. In the following of the paper, we would

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2024.3398496

TABLE I: Comparisons on computations among different FIMs and the GGN matrix.

	Computation ingredients and storage
The GGN matrix	$J(\theta)$, $\mathcal{H}_L(\theta)$
Minibatch FIM	$\{\nabla_\theta L(y, f(x,\theta))\}_{x\in\{x_1,\ldots x_n\}\times y\sim p_{y\mid x}(\theta)}$
EFIM	$\{\nabla_\theta L(y_i, f(x_i,\theta))\}_{i=1}^N$

derive the global and local convergence of NGMs through the connections between the FIMs and the GGN matrix, which could help us to better understand their practical efficiency.

III. Global convergence of an adaptively regularized NGM

Note that the FIMs are always positive semidefinite. By adding a regularization term, the commonly used iterative scheme of the NGM [2], [3], [17], [19] is given by

$$\theta_{k+1} = \theta_k + d_k, \tag{11}$$

where d_k is obtained by the solution of

$$\min_{d\in\mathbb{R}^n} m_k(d) := g_k^\top d + \frac{1}{2} d^\top (F_k + \lambda_k I) d. \tag{12}$$

Here, $F_k \succeq 0$ is the minibatch EFIM or minibatch FIM approximation of $F(\theta_k)$, $\lambda_k > 0$ is a regularization scalar, and $g_k = g_k(\theta_k)$ is a mini-batch approximation of the gradient $\nabla_\theta h(\theta_k)$, where

$$\begin{aligned} g_k(\theta_k) \;&:=\; \frac{1}{b_k^g} \sum_{i\in\mathcal{B}_k^g} \nabla_\theta L(y_i, f(x_i, \theta_k)) &(13)\\ &=\; \frac{1}{b_k^g} \sum_{i\in\mathcal{B}_k^g} J_{f(x_i,\theta)}^\top(\theta_k) \nabla_z L(y_i, z)|_{z=f(x_i,\theta_k)} \\ &=\; \frac{1}{b_k^g} J_k(\theta_k)^\top \mathcal{G}_k(\theta_k) &(14)\end{aligned}$$

with $\mathcal{B}_k^g = \{i_{k,1}, \ldots, i_{k,b_k^g}\} \subset [N]$, $b_k^g = |\mathcal{B}_k^g|$,

$$\begin{aligned} J_k(\theta) &= \begin{bmatrix} J_{f(x_{i_{k,1}},\theta)}(\theta) \\ \vdots \\ J_{f(x_{i_{k,b_k^g}},\theta)}(\theta) \end{bmatrix}, \\ \mathcal{G}_k(\theta) &= \begin{bmatrix} \nabla_z L(y_{i_{k,1}}, z)|_{z=f(x_{i_{k,1}},\theta)} \\ \vdots \\ \nabla_z L(y_{i_{k,b_k^g}}, z)|_{z=f(x_{i_{k,b_k^g}},\theta)} \end{bmatrix}. \end{aligned} \tag{15}$$

Since $F_k + \lambda_k I \succ 0$, the solution of (12), $(F_k + \lambda_k I)^{-1} g_k$, always exists uniquely. In general, the update (11) may not lead to convergence. The behaviour highly relies on the specific choice of the step size α_k (which is 1 in (11)). One of our goals is to ensure global convergence by adaptively updating the regularization scalar λ_k. Of course, another possible strategy is to adapt a diminishing step size α_k by using the backtracking line search as in [45], [46], and to update $\theta_{k+1} = \theta_k + \alpha_k d_k$.

A. An adaptively regularized NGM

Denoting $\mathcal{G}_k = \mathcal{G}_k(\theta_k)$, our adaptively regularized NGM sets the regularization scalar

$$\lambda_k = \frac{\sigma_k}{\sqrt{b_k^g}} \|\mathcal{G}_k\|, \tag{16}$$

where the regularization parameter σ_k is adaptively updated in the algorithm. In addition to the mini-batch approximations in (13), at the k-th iteration, we define mini-batch approximations of $h_k^0 \approx h(\theta_k)$ and $h_k^d \approx h(\theta_k + d_k)$, respectively, by

$$\begin{aligned} h_k^0 &= \frac{1}{b_k^h} \sum_{i\in\mathcal{B}_k^h} L(y_i, f(x_i, \theta_k)), \\ h_k^d &= \frac{1}{b_k^h} \sum_{i\in\mathcal{B}_k^h} L(y_i, f(x_i, \theta_k + d_k)). \end{aligned} \tag{17}$$

For evaluation of the step d_k, we introduce

$$\rho_k = \frac{h_k^0 - h_k^d}{m_k(0) - m_k(d_k)} = \frac{h_k^0 - h_k^d}{-m_k(d_k)} \tag{18}$$

and use the update rule

$$\theta_{k+1} = \begin{cases} \theta_k + d_k, & \text{if } \rho_k \geq \eta_1, \\ \theta_k, & \text{otherwise}, \end{cases} \tag{19}$$

with some parameter $\eta_1 \in (0,1)$. Due to the stochastic inexactness of the sample average estimates h_k^0, h_k^d, g_k, the classic updating rule [39] for σ_k only depending on ρ_k does not guarantee the convergence. Here, we use an updating rule for regularization parameter σ_k in spirit similar to the strategy used in [47], [48],

$$\sigma_{k+1} = \begin{cases} \max\{\gamma\sigma_k, \sigma_{\min}\}, & \text{if } \rho_k \geq \eta_1 \text{ and } \|g_k\| \geq \frac{\eta_2}{\sigma_k}, \\ \frac{1}{\gamma}\sigma_k, & \text{otherwise}, \end{cases} \tag{20}$$

where $0 < \gamma < 1$, $\eta_2 > 0$ and $\sigma_{\min} > 0$ are constants. Our adaptively regularized NGM is presented in Algorithm 1. In contrast to the algorithms in [32] and [33] that use the Gauss-Newton matrix-based Levenberg–Marquardt method, Algorithm 1 is a FIM-based natural gradient method for solving problem (1) and its convergence is more complicated due to the use of inexact function evaluations. We note that a stochastic trust-region algorithm is presented in [40]. In comparison to their method, where $\theta_{k+1} = \theta_k + d_k$ if both $\rho_k \geq \eta_1$ and $\|g_k\| \geq \frac{\eta_2}{\sigma_k}$, Algorithm 1 utilizes adaptive regularization and has a less strict acceptance rule as defined in equation (19), i.e., removing the dependency on the gradient norms. In addition, the subproblem (12) can be efficiently solved if the cost of the inverse of the matrix $F_k + \lambda_k I$ and a vector multiplication is low. In [2], the authors employ the multi-layer structure of the neural network and give a Kronecker-factored approximation of F_k to reduce the large scale matrix into the Kronecker product of two smaller matrices. In addition, a block diagonal approximation is investigated to further reduce the computations of the inverses.

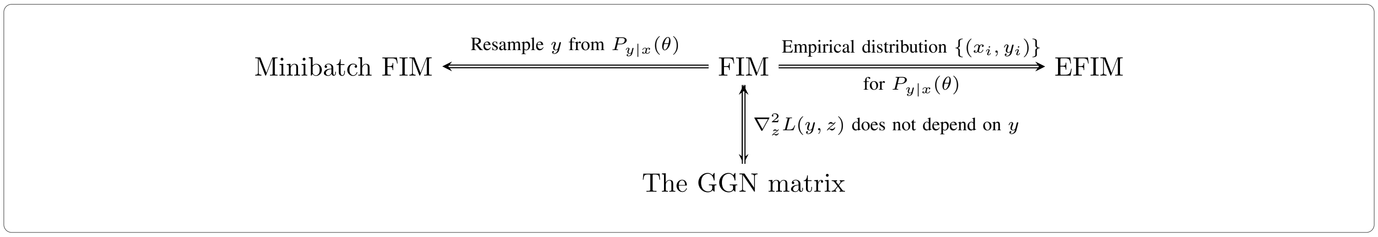

Minibatch FIM
Resample y from $P_{y|x}(\theta)$
FIM
Empirical distribution $\{(x_i, y_i)\}$
for $P_{y|x}(\theta)$
EFIM
$\nabla_z^2 L(y,z)$ does not depend on y
The GGN matrix

Fig. 1: Implications between different concepts of the FIMs and the GGN matrix.

Algorithm 1 An adaptively regularized NGM.

```
Input: Parameters $\theta_0$, $\sigma_0$ and constants $\eta_1 \in (0,1)$, $\gamma \in (0,1)$, $\eta_2 > 0$, $\sigma_{\min} > 0$. Set $k = 0$.
while stopping conditions not met do
    Choose mini-batch index set $\mathcal{B}_k^g \subset [N]$, and compute $g_k$ and $\mathcal{G}_k$.
    Compute $F_k$ and set $\lambda_k = \frac{\sigma_k}{\sqrt{b_k^g}}\|\mathcal{G}_k\|$.
    Compute $d_k$ as the solution of (12).
    Choose mini-batch index set $\mathcal{B}_k^h \subset [N]$, compute $h_k^0$ and $h_k^d$.
    Calculate ratio $\rho_k$ by (18).
    Update $\theta_{k+1}$ by (19) and $\sigma_{k+1}$ by (20).
    $k = k + 1$.
end while
```

B. Global convergence

Before presenting the convergence, let us start with some necessary notations. We use M_k to denote the random model in the k-th iteration and $m_k = M_k(\omega)$ to denote its realization. Consequently, the iterates Θ_k, the regularization parameters Λ_k, Σ_k, and the iteration steps D_k are also random quantities. Let $\theta_k = \Theta_k(\omega)$, $\lambda_k = \Lambda_k(\omega)$, $\sigma_k = \Sigma_k(\omega)$, and $d_k = D_k(\omega)$ be their realizations. Besides, we use $\{O_k^0, O_k^d\}$ to denote the stochastic estimates of $\{h(\Theta_k), h(\Theta_k + D_k)\}$. Their realizations are represented by $h_k^0 = O_k^0(\omega)$ and $h_k^d = O_k^d(\omega)$. Let G_k be the stochastic estimate of $\nabla h(\Theta)$ with $g_k = G_k(\omega)$ being its realization. Hence, a stochastic process $\{\Theta_k, G_k, M_k, D_k, O_k^0, O_k^d, \Sigma_k, \Lambda_k\}$ is generated by Algorithm 1.

For the simplicity in emphasizing the dependency of random quantities, we define $\mathcal{F}_{k-1}^{M,O}$ as the σ-algebra generated by $M_0, \ldots, M_{k-1}$, and $O_1^0, O_1^d, \ldots, O_{k-1}^0, O_{k-1}^d$. Let $\mathcal{F}_{k-1/2}^{M,O}$ be the σ-algebra generated by $M_0, \ldots, M_k$ and $O_1^0, O_1^d, \ldots, O_{k-1}^0, O_{k-1}^d$. Let $\mathcal{F}_{k-1}^{M}$ be the σ-algebra generated by $M_0, \ldots, M_{k-1}$. To derive global convergence, we need the following assumption.

Assumption 1. *Define $\mathcal{L}(\theta_0) = \{\theta \in \mathbb{R}^n : h(\theta) \leq h(\theta_0)\}$.*

(a) The function h is twice continuously differentiable and bounded from below. Its gradient is Lipschitz continuous with modulus κ_h, i.e., for any θ_1, θ_2,

$$\|\nabla h(\theta_1) - \nabla h(\theta_2)\| \leq \kappa_h \|\theta_1 - \theta_2\|. \tag{21}$$

In addition, the Jacobian $J_{f(x_i,\theta)}(\theta)$ and $\nabla_z L(y_i, z)|_{z=f(x_i,\theta)}$ are bounded on $\mathcal{L}(\theta_0)$ with constants κ_J and $\kappa_{\mathcal{G}}$, i.e., for all $i \in [N]$ and any $\theta \in \mathcal{L}(\theta_0)$, $\|J_{f(x_i,\theta)}\| \leq \kappa_J$, $\|\nabla_z L(y_i, z)|_{z=f(x_i,\theta)}\| \leq \kappa_{\mathcal{G}}$.

(b) The approximate FIM F_k is bounded from above for all k, i.e., there exists κ_{fim} such that

$$\|F_k\| \leq \kappa_{\mathrm{fim}}, \quad \forall k = 1, 2, \ldots \tag{22}$$

(c) The mini-batch index set $\mathcal{B}_k^g$ is chosen such that the sequence of random gradients $\{G_k\}$ is at least $\frac{1}{2}$-probabilistically $\frac{\kappa_g}{\sigma_k}$-first-order accurate, i.e.,

$$P(E_{k,G} | \mathcal{F}_{k-1}^{M,O}) \geq \frac{1}{2}, \tag{23}$$

where $E_{k,G} := \left\{\|G_k - \nabla h(\theta_k)\| \leq \frac{\kappa_g}{\sigma_k}\right\}$ with $\kappa_g > 0$.

(d) The mini-batch index set $\mathcal{B}_k^h$ is chosen such that the sequence of random function values $\{O_k^0, O_k^d\}$ is at least $(1-\tau_k)$-probabilistically ϵ_O^k-zero-order accurate, i.e.

$$P(E_{k,O} | \mathcal{F}_{k-1/2}^{M,O}) \geq 1 - \tau_k \tag{24}$$

with a sequence $\{\tau_k\}$ such that $\tau_k \in [0,1)$, $\sum_{k=1}^{\infty} \tau_k < \infty$ and the event

$$E_{k,O} = \left\{|O_k^0 - h(\theta_k)| \leq \epsilon_O^k,\ |O_k^d - h(\theta_k + d_k)| \leq \epsilon_O^k\right\}, \tag{25}$$

where $\epsilon_O^k := \min\left\{\frac{\eta_1\|g_k\|^2}{8(\|F_k\| + \sigma_k/\sqrt{b_k^g}\|\mathcal{G}_k\|)}, \frac{\kappa_O}{\sigma_k^2}\right\}$ and $\kappa_O > 0$ is a constant.

Remark 1. *The Assumptions 1 (a) and (b) are standard in the analysis of optimization methods [31], [39]. In addition, the Assumptions (c) and (d) can be satisfied if the batch sizes for evaluating gradient and function estimations are large enough. We refer to [40], [48] for similar assumptions.*

The global convergence proof can be split into the following steps. Firstly, we show σ_k will go to infinity almost surely. Secondly, we show that the trial step $\theta_k + d_k$ will be accepted as θ_{k+1} for sufficiently large σ_k if the gradient and function value estimates are sufficiently accurate. Finally, by the martingale theorem [49, Exercise 5.3.1], we show the almost sure convergence of the gradient norms. Let us start with the sufficient function value reduction under $E_{k,O}$.

Lemma 1. *Suppose that Assumption 1 (b) holds. For any successful update at θ_k (i.e., $\rho_k \geq \eta_1$), when the event $E_{k,O}$ happens, we have*

$$h(\theta_{k+1}) \leq h(\theta_k) - \frac{\eta_1 \|g_k\|^2}{4\left(\kappa_{\mathrm{fim}} + (\sigma_k/\sqrt{b_k^g})\|\mathcal{G}_k\|\right)}, \tag{26}$$

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2024.3398496

where $E_{k,O}$ is defined in (25).

Proof. By (12) and $F_k \succeq 0$, it holds that

$$-m_k(d_k) \geq \frac{1}{2}\frac{\|g_k\|^2}{\|F_k + \lambda_k I\|} = \frac{\|g_k\|^2}{2\left(\|F_k\| + (\sigma_k/\sqrt{b_k^g})\|\mathcal{G}_k\|\right)}. \tag{27}$$

So, for a successful update $\theta_{k+1} = \theta_k + d_k$, by (19) we have

$$h_k^0 - h_k^d \geq -\eta_1 m_k(d_k) \geq \frac{\eta_1\|g_k\|^2}{2\left(\|F_k\| + (\sigma_k/\sqrt{b_k^g})\|\mathcal{G}_k\|\right)}.$$

Under event $E_{k,O}$, we have h_k^0 and h_k^d are ϵ_O^k-accurate. Hence, by Assumption 1 (b) and (25), we have

$$\begin{aligned} h(\theta_k) - h(\theta_{k+1}) &= h(\theta_k) - h_k^0 + h_k^0 - h_k^d + h_k^d - h(\theta_{k+1}) \\ &\geq \frac{\eta_1\|g_k\|^2}{4\left(\|F_k\|+(\sigma_k/\sqrt{b_k^g})\|\mathcal{G}_k\|\right)} \geq \frac{\eta_1\|g_k\|^2}{4\left(\kappa_{\mathrm{fim}}+(\sigma_k/\sqrt{b_k^g})\|\mathcal{G}_k\|\right)}. \end{aligned}$$

□

Lemma 2. *Suppose that Assumption 1 (a), (b), and (d) hold. Let $\{\Sigma_k\}$ be generated by Algorithm 1. Then, it holds almost surely that*

$$\lim_{k\to\infty} \Sigma_k = +\infty. \tag{28}$$

Proof. First, for any $\epsilon \in (0,1)$, it follows from $\tau_k \in [0,1)$ and $\sum_{k=1}^{\infty}\tau_k < \infty$ in Assumption 1 (d) that there exists a $K > 0$ such that

$$\sum_{k=K}^{\infty}\tau_k \leq -\frac{\ln(1-\epsilon)}{2} \quad \text{and} \quad \tau_k \in [0, 1/2) \text{ for all } k \geq K,$$

which implies that

$$\prod_{k=K}^{\infty}(1-\tau_k) \geq \exp\left(-2\sum_{k=K}^{\infty}\tau_k\right) \geq 1-\epsilon.$$

Hence, by Assumption 1 (d), we have

$$P\Big(E_{k,O} \text{ happens for all } k \geq K\Big) \geq \prod_{k=K}^{\infty}(1-\tau_k) \geq 1-\epsilon. \tag{29}$$

In the following, conditioning on the event that $E_{k,O}$ happens for all $k \geq K$, we show by contradiction as in [48, Lemma 2.5] that $\lim_{k\to\infty}\sigma_k = \infty$, where the sequence $\{\sigma_k\}$ is any realization of $\{\Sigma_k\}$.

Suppose that σ_k does not go to ∞. Then, there exist $\tilde{\sigma}$ such that the set $S_1 = \{k : \sigma_k < \tilde{\sigma}\}$ is infinite. (Otherwise, if such $\tilde{\sigma}$ does not exist, σ_k goes to ∞.) Due to $0 < \gamma < 1$, the set $S_2 := \{k : \sigma_k < \tilde{\sigma}/\gamma\}$ is also infinite. Consider the set

$$S_3 := \{k \in S_2 : \sigma_{k+1} \leq \sigma_k\}. \tag{30}$$

We claim that S_3 is also infinite. If not, there exists a constant $N_0 \in S_2$ such that $\sigma_{k+1} > \sigma_k$ for all $k \geq N_0$ and $k \in S_2$. Since $\sigma_k \geq \tilde{\sigma}/\gamma$ for all $k \notin S_2$, by the updating rule (20) of σ_k, there exists a $N_1 > N_0$ such that $\sigma_k \geq \tilde{\sigma}$ for all $k \geq N_1$. This conflicts to the infiniteness of S_1. Hence, S_3 is infinite. Now, from the update rule (20) and the definition of S_3, it holds that

$$\|g_k\| \geq \frac{\eta_2}{\sigma_k}, \ \rho_k \geq \eta_1, \ \text{and } \sigma_k < \frac{\tilde{\sigma}}{\gamma} \quad \forall k \in S_3.$$

Since $E_{k,O}$ happens for all $k \geq K$, we have from the updating rule (19) of θ_k that the sequence $\{h(\theta_k)\}_{k=K}^{\infty}$ is monotonically nonincreasing, and from Lemma 1 that

$$\begin{aligned} h(\theta_k) - h(\theta_{k+1}) &\geq \frac{\eta_1\|g_k\|^2}{4\left(\kappa_{\mathrm{fim}} + (\sigma_k/\sqrt{b_k^g})\|\mathcal{G}_k\|\right)} \\ &\geq \frac{\eta_1\eta_2^2}{4\sigma_k^2(\kappa_{\mathrm{fim}} + \sigma_k\kappa_{\mathcal{G}})} \geq \frac{\eta_1\eta_2^2\gamma^3}{4\tilde{\sigma}^2(\gamma\kappa_{\mathrm{fim}} + \tilde{\sigma}\kappa_{\mathcal{G}})} \end{aligned}$$

for all $k \in S_3$ and $k \geq K$, where the second inequality follows from Assumption 1 (a) and $\|\mathcal{G}_k\| \leq \kappa_{\mathcal{G}}\sqrt{b_k^g}$. Since the sequence $\{h(\theta_k)\}_{k=K}^{\infty}$ is nonincreasing, for all $\ell \geq K$ we have

$$\begin{aligned} h(\theta_K) - h(\theta_{\ell+1}) &\geq \sum_{j\in S_3,\ K\leq j\leq \ell} h(\theta_j) - h(\theta_{j+1}) \\ &\geq \sum_{j\in S_3,\ K\leq j\leq \ell} \frac{\eta_1\eta_2^2\gamma^3}{4\tilde{\sigma}^2(\gamma\kappa_{\mathrm{fim}} + \tilde{\sigma}\kappa_{\mathcal{G}})}. \end{aligned}$$

Taking $\ell \to +\infty$ and noticing the infiniteness of S_3, the above inequality contradicts with the bounded below assumption of h.

Hence, for any $0 < \epsilon < 1$ we have

$$P\left(\lim_{k\to\infty}\Sigma_k = \infty\right) \geq P\left(E_{k,O} \text{ happens for all } k \geq K\right) \geq 1-\epsilon.$$

Then, by the arbitrary choice of $0 < \epsilon < 1$, $\lim_{k\to\infty}\Sigma_k = \infty$ almost surely. □

The following lemma reveals that when both events $E_{k,G}$ and $E_{k,O}$ happen and σ_k is sufficiently large, we will have the update $\theta_{k+1} = \theta_k + d_k$.

Lemma 3. *Suppose that Assumption 1 (a) and (b) hold. When the events $E_{k,G}$ and $E_{k,O}$ happen and*

$$\sigma_k \geq \max\left\{\frac{\kappa_{\mathrm{fim}}}{\kappa_{\mathcal{G}}}, \frac{4\kappa_{\mathcal{G}}\kappa_J^2(\kappa_g + \kappa_h) + 8\kappa_{\mathcal{G}}\kappa_O}{(1-\eta_1)\|g_k\|^2}\right\}, \tag{31}$$

it holds that $\rho_k \geq \eta_1$ and $\theta_{k+1} = \theta_k + d_k$.

Proof. It follows from Assumption 1 (a) and the definition of g_k in (13) that

$$\begin{aligned} \|g_k\| &= \left\|\frac{1}{b_k^g}\sum_{i\in\mathcal{B}_k^g} J_{f(x_i,\theta)}^{\top}(\theta)\nabla_z L(y_i,z)|_{z=f(x_i,\theta)}\right\| \\ &\leq \frac{1}{b_k^g}\sum_{i\in\mathcal{B}_k^g}\left\|J_{f(x_i,\theta)}^{\top}(\theta)\nabla_z L(y_i,z)|_{z=f(x_i,\theta)}\right\| \\ &\leq \frac{\kappa_J}{b_k^g}\sqrt{b_k^g}\|\mathcal{G}_k\| = \frac{\kappa_J}{\sqrt{b_k^g}}\|\mathcal{G}_k\|, \end{aligned} \tag{32}$$

where the second inequality is due to $\|J_{f(x_i,\theta)}\| \leq \kappa_J$ and $\mathcal{G}_k = \mathcal{G}(\theta_k)$ is defined in (13).

Since $\sigma_k \geq \kappa_{\mathrm{fim}}/\kappa_{\mathcal{G}}$ and the assumption $\|F_k\| \leq \kappa_{\mathrm{fim}}$, recalling $\|\mathcal{G}_k\| \leq \kappa_{\mathcal{G}}\sqrt{b_k^g}$, (27) and $F_k \succeq 0$, we have

$$\begin{aligned} -m_k(d_k) &\geq \frac{\|g_k\|^2}{4\sigma_k\kappa_{\mathcal{G}}}, \\ \|d_k\| &= \left\|\left(F_k + (\sigma_k/\sqrt{b_k^g})\|\mathcal{G}_k\|I\right)^{-1} g_k\right\| \leq \frac{\kappa_J}{\sigma_k}, \end{aligned}$$

where the second inequality uses (32). Then, by the Lipschitz continuity of ∇h in Assumption 1 (a), when the event $E_{k,G}$ happens, we have

$$\begin{aligned} & h(\theta_k+d_k)-h(\theta_k)-m_k(d_k) & (33)\\ \le\ & d_k^\top(\nabla h(\theta_k)-g_k)+\frac{\kappa_h}{2}\|d_k\|^2 & (34)\\ \le\ & \frac{\kappa_g}{\sigma_k}\|d_k\|+\kappa_h\|d_k\|^2 \le \frac{\kappa_J^2(\kappa_g+\kappa_h)}{\sigma_k^2}. & \end{aligned}$$

Hence, when the event $E_{k,O}$ also happens, we have

$$\begin{aligned} 1-\rho_k &= \frac{-m_k(d_k)-h_k^0+h_k^d}{-m_k(d_k)}\\ &= \frac{-m_k(d_k)-h(\theta_k)+h(\theta_k+d_k)+h(\theta_k)-h_k^0-h(\theta_k+d_k)+h_k^d}{-m_k(d_k)}\\ &\le \frac{\kappa_J^2(\kappa_g+\kappa_h)/\sigma_k^2+2\kappa_O/\sigma_k^2}{\|g_k\|^2/(4\sigma_k\kappa_{\mathcal{G}})} = \frac{4\kappa_{\mathcal{G}}\kappa_J^2(\kappa_g+\kappa_h)+8\kappa_{\mathcal{G}}\kappa_O}{\sigma_k\|g_k\|^2}\\ &\le 1-\eta_1, \end{aligned}$$

which implies $\rho_k \ge \eta_1$. Then, $\theta_{k+1}=\theta_k+d_k$ follows from the updating rule (19), $\rho_k\ge\eta_1$. We complete the proof. □

From Lemmas 1, 2, and 3, we are going to show the almost sure convergence of Algorithm 1. To proceed it, we need the following result on the convergence of submartingales.

Proposition 1. *[49, Exercise 5.3.1] Let Q_k be a submartingale, i.e., a sequence of random variables which, for every k, $\mathbb{E}\left[Q_k|\mathcal{F}_{k-1}^Q\right] \ge Q_{k-1}$, where $\mathcal{F}_{k-1}^Q=\sigma(Q_0,\dots,Q_{k-1})$ is the σ-algebra generated by $Q_0,\dots,Q_{k-1}$, and $\mathbb{E}\left[Q_k|\mathcal{F}_{k-1}^Q\right]$ denotes the conditional expectation of Q_k given the past history of events $\mathcal{F}_{k-1}^Q$. If $Q_k-Q_{k-1}\le C<\infty$, for every k, then,*

$$P\left(\left\{\lim_{k\to\infty} Q_k<\infty\right\}\cup\left\{\limsup_{k\to\infty} Q_k=\infty\right\}\right)=1.$$

With the above proposition, there is at least a subsequence generated by Algorithm 1 converging to a stationary point almost surely.

Theorem 1. *Suppose that Assumption 1 holds. Let $\{\Theta_k\}$ be the random iterates generated by Algorithm 1. Then, it holds almost surely that*

$$\liminf_{k\to\infty}\|\nabla h(\Theta_k)\|=0. \qquad (35)$$

Proof. We prove (35) by way of contradiction. If (35) does not hold, then there exist $\tau>0$, $\xi>0$, and an integer $K>0$ such that

$$P\Big(\|\nabla h(\Theta_k)\|\ge\xi \text{ for all } k\ge K\Big)\ge 3\tau. \qquad (36)$$

By (36) and the inequality (29) with $\epsilon=\tau$, when K is sufficiently large, we have

$$P\Big(\|\nabla h(\Theta_k)\|\ge\xi \text{ and } E_{k,O} \text{ happens for all } k\ge K\Big)\ge 2\tau. \qquad (37)$$

Let us define a random variable

$$Z_k=\log_{\gamma^{-1}}(\Sigma_k^{-1}) \qquad (38)$$

and z_k be its realization. By Lemma 2, $\lim_{k\to\infty}\Sigma_k=\infty$ almost surely. So, when K is sufficiently large, we have

$$P\Big(\text{inequality (40) holds for all } k\ge K\Big)\ge 1-\tau, \qquad (39)$$

where

$$\Sigma_k\ge\bar{\Sigma}(\xi):= \max\left\{\frac{2\kappa_g}{\xi},\frac{2\eta_2}{\xi},\frac{\kappa_{\text{fim}}}{\kappa_{\mathcal{G}}},\frac{16\kappa_{\mathcal{G}}\kappa_J^2(\kappa_g+\kappa_h)+32\kappa_{\mathcal{G}}\kappa_O}{(1-\eta_1)\xi^2},\frac{\sigma_{\min}}{\gamma}\right\}. \qquad (40)$$

Combing (37) and (39), it gives $P(\hat{E})\ge\tau$, where the event $\hat{E}$ is defined as

$$\hat{E}:=\Big\{\|\nabla h(\Theta_k)\|\ge\xi, E_{k,O} \text{ happens and inequality (40) holds for all } k\ge K\Big\}.$$

In the following, let us consider the stochastic process generated by Algorithm 1 conditioning on the event $\hat{E}$. Without loss of generality, let us simply assume $K=0$ in the rest of the proof.

First, let the event $\hat{E}_{k,G}$ be the event $E_{k,G}$ conditioning on $\hat{E}$. Then, by Assumption 1 (c) and the definition (23) of $E_{k,G}$, we have

$$P(\hat{E}_{j,G}|\mathcal{F}_{j-1}^M)=P(E_{j,G}|\mathcal{F}_{j-1}^M,\hat{E})=P(E_{j,G}|\mathcal{F}_{j-1}^M)\ge 1/2. \qquad (41)$$

Let $W_j=\sum_{k=0}^j(2\cdot 1_{\hat{E}_{k,G}}-1)$ with $1_{\hat{E}_{k,G}}$ being the characteristic function of $\hat{E}_{k,G}$, i.e., $1_{\hat{E}_{k,G}}$ is 1 if $\hat{E}_{k,G}$ happens and 0 otherwise. Then,

$$W_j=\begin{cases} W_{j-1}+1, & \text{if } 1_{\hat{E}_{j,G}}=1,\\ W_{j-1}-1, & \text{otherwise,}\end{cases} \qquad (42)$$

which gives

$$|W_j-W_{j-1}|=1. \qquad (43)$$

Using (41) and (42), the conditional expectation satisfies

$$\begin{aligned}\mathbb{E}(W_j|\mathcal{F}_{j-1}^M) &= \mathbb{E}(W_{j-1}|\mathcal{F}_{j-1}^M)+\mathbb{E}(2\cdot 1_{\hat{E}_{j,G}}-1|\mathcal{F}_{j-1}^M)\\ &= W_{j-1}+2P(\hat{E}_{j,G}|\mathcal{F}_{j-1}^M)-1\\ &\ge W_{j-1}+2\cdot\frac{1}{2}-1\ge W_{j-1},\end{aligned}$$

which implies W_j is a submartingale. By (43) and Proposition 1, we have

$$P\left(\limsup_{j\to\infty} W_j=\infty\right)=1. \qquad (44)$$

Conditioning on $\hat{E}$, let us consider two cases at the k-th iteration: $\hat{E}_{k,G}$ happens and $\hat{E}_{k,G}$ does not happen.

- $\hat{E}_{k,G}$ happens (i.e. $E_{k,G}$ happens conditioning on $\hat{E}$): In this case, we have $\|g_k-\nabla h(\theta_k)\|\le\frac{\kappa_g}{\sigma_k}\le\frac{\xi}{2}$. Then,

  $$\|g_k\|\ge\|\nabla h(\theta_k)\|-\|\nabla h(\theta_k)-g_k\|\ge\frac{\xi}{2}.$$

  It follows from (40) and Lemma 3 that $\|g_k\|\ge\eta_2/\sigma_k$ and $\rho_k\ge\eta_1$ for all $k\ge K$. Then, it follows from the σ_k updating rule (20) and $\sigma_k\ge\sigma_{\min}/\gamma$ that $\sigma_{k+1}=\gamma\sigma_k$. Therefore,

  $$z_{k+1}=\log_{\gamma^{-1}}(\sigma_{k+1}^{-1})=\log_{\gamma^{-1}}(\gamma^{-1}\sigma_k^{-1})=z_k+1.$$

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2024.3398496

Let w_k be the realization of W_k. Then, since $1_{\hat{E}_{k,G}} = 1$, by (42) it holds that

$$z_{k+1} - z_k = w_{k+1} - w_k = 1. \tag{45}$$

- $E_{k,G}$ does not happen (i.e. $E_{k,G}$ conditioning on $\hat{E}$ does not happen): In this case, by the σ_k updating rule (20), we always have $\sigma_{k+1} \leq (1/\gamma)\sigma_k$. Then,

  $$z_{k+1} = \log_{\gamma^{-1}}(\sigma_{k+1}^{-1}) \geq \log_{\gamma^{-1}}(\gamma\sigma_k^{-1}) = z_k - 1.$$

  Then, since $1_{\hat{E}_{k,G}} = 0$, by (42) it holds that

  $$z_{k+1} - z_k \geq w_{k+1} - w_k = -1. \tag{46}$$

Combining (45) and (46), we have $z_k \geq w_k - w_0 + z_0$. Hence, by (44), one has

$$P\left(\limsup_{k\to\infty} Z_k = \infty \;\middle|\; \hat{E}\right) = 1. \tag{47}$$

Since $P(\hat{E}) \geq \tau > 0$, we have

$$P\left(\limsup_{k\to\infty} Z_k = \infty\right) \geq \tau > 0. \tag{48}$$

By the definition of Z_k in (38) and $\lim_{k\to\infty} \Sigma_k = \infty$ almost surely by Lemma 2, we have $\lim_{k\to\infty} Z_k = -\infty$ almost surely, which contradicts with (48). So, (35) holds almost surely, and we complete the proof. □

Remark 2. *Note that the upper boundedness of $\|F_k\|$ is sufficient for the global convergence. This allows using more practical approximations of FIMs to further reduce the computation, e.g., the Kronecker-factored approximation in [2].*

We also note that the result of Theorem 1 can be improved to the full sequence $\lim_{k\to\infty} \|\nabla h(\Theta_k)\| = 0$ by a similar proof in [40, Theorem 4.18]. Moreover, the explicit complexity bound on the expected number of iterations required to achieve ϵ-accuracy could be obtained by similar approaches in [50], which presents the iteration complexity for a class of trust region based stochastic optimization methods.

The requirement of gradually increasing accuracy on the estimates of the objective function and gradient is a common assumption in the context of probabilistic model based algorithms, including [40]. However, these analyses do not rely on the boundedness of the variance of the estimates, which is often used in analyzing the stochastic gradient methods. Additionally, we are able to establish the almost sure convergence of the gradient norms. This is more reliable than the existing the convergence with expectation or high probability of stochastic gradient type methods. Although we may need large batch sizes to obtain accurate gradients and functional evaluations given by Assumption 1, it will not affect the effectiveness of the natural gradient methods as large batch training [51]–[53] is commonly used and could improve numerical performance.

Remark 3. *We note that the prox-linear method in [54] for solving our problem (1) reduces to a stochastic Gauss-Newton method. This is different from our FIM-based natural gradient method, where the Jacobian approximation in [54] is not needed. In addition, since our method is based on the probabilistic model and the trust-region like adaptive strategy, both the algorithmic framework and the convergence analysis are quite different with their method.*

Algorithm 2 Local NGM

```
Input: Choose an initial parameter $\theta_0$. Set $k = 0$.
while stopping conditions not met do
    Choose $\mathcal{B}_k^F = \mathcal{B}_k^g =: \mathcal{B}_k \subset [N]$ and $N_k^y \in \mathbb{N}$ .
    Compute $g_k$ and $F_k$ and set $\lambda_k = \|\mathcal{G}_k\|/\sqrt{|\mathcal{B}_k|}$.
    Compute $d_k$ as the solution of (12).
    Set $\theta_{k+1} = \theta_k + d_k$.
    $k = k + 1$.
end while
```

In the following section, we examine the local convergence speed of a NGM type Algorithm 2.

IV. Local convergence analysis of the NGM

In the previous section III, we have applied an adaptive regularization and a trust region type technique to ensure global convergence. This results the regularization parameter Σ_k approaches to infinity almost surely (see Lemma 2). In fact, from the proof of Theorem 1 we can see that given any $\xi > 0$, as long as $\Sigma_k > \bar{\Sigma}(\xi)$ holds almost surely for all k sufficiently large, where $\bar{\Sigma}(\xi)$ is defined in (40), we will have $\liminf_{k\to\infty} \|\nabla h(\Theta_k)\| \leq \xi$ almost surely. Hence, in the practical application of NGM Algorithm 1, we can set up a sufficiently large upper bound $\hat{\Sigma}$ of Σ_k. When Σ_k reaches this upper bound $\hat{\Sigma}$ and $\|g_k\|$ does not get reduced sufficiently often, we may consider switching to the local NGM Algorithm 2 to accelerate the convergence. In Algorithm 2, we update $\theta_{k+1} = \theta_k + d_k$ at each iteration, where d_k is the solution of (12), and simply set $\sigma_k = 1$ for convenience of local analysis. However, setting σ_k to be any positive constant, for instance setting $\sigma_k = \hat{\Sigma}$ for all k, will not affect the analysis of local convergence speed. In addition, at each iteration of Algorithm 2, we set the minibatch sampling sets $\mathcal{B}_k^F = \mathcal{B}_k^g =: \mathcal{B}_k \subset [N]$ with $|\mathcal{B}_k| = b_k$.

For local convergence analysis, we consider the loss functions $L(y, z)$, which are twice continuously differentiable with respect to z and $\nabla_z^2 L(y, z)$ does not depend on y. In this case, by (7), we have $\hat{H}_i(\theta) = \nabla_z^2 L(y, z)|_{z=f(x_i,\theta)} = H_i(\theta)$. Since no samplings on y is needed, we have $N_k^y = 1$. Consequently, in this case, the minibatch EFIM defined in (9) will be the same as the minibatch FIM. Furthermore, we assume the matrix F_k in the quadratic model (12) can be theoretically written in the form of

$$F_k = \frac{1}{b_k} J_k(\theta_k)^\top \mathcal{H}_k(\theta_k) J_k(\theta_k), \tag{49}$$

where $\mathcal{H}_k(\theta) = \mathrm{Diag}(H_i(\theta) : i \in \mathcal{B}_k)$ is a block-diagonal submatrix of $\mathcal{H}(\theta)$.

Remark 4. *Note that the Hessian of L with respect to z, i.e., $\nabla_z^2 L(y, z)$, does not depend on y in many practical applications in machine learning. For example, it holds for the following commonly used loss functions:*

- *For the square loss function $L(y, z) = \|y - z\|_2^2$, it holds that $\nabla_z^2 L(y, z) = I$.*

- *Consider the cross-entropy loss function* $L(y,z) = -y\log\tilde z-(1-y)\log(1-\tilde z)$ *with* $\tilde z = \mathrm{Sigmoid}(z)$. *Then, we obtain* $\nabla_z^2 L(y,z) = \mathrm{diag}\big(\tilde z \odot (1-\tilde z)\big)$.
- *For the cross-entropy loss* $L(y,z) = -\sum_i [y]_i \log[\tilde z]_i$ *with* $\tilde z = \mathrm{Softmax}(z)$, *we obtain* $\nabla_z^2 L(y,z) = \mathrm{diag}(\tilde z) - \tilde z\tilde z^\top$.

Let Θ^* be the set of local minimums of (1), which is a closed set, and let us define

$$\mathrm{dist}(\theta,\Theta^*) := \|\theta-\bar\theta\| \text{ with } \bar\theta \in \arg\min_{\tilde\theta\in\Theta^*}\|\tilde\theta-\theta\|.$$

To establish local convergence speed in this section, we also need the following assumptions.

Assumption 2. *Define* $B(\Theta^*,b) = \{\theta : \mathrm{dist}(\theta,\Theta^*) < b\}$, *where* $b>0$ *is some constant.*

(a) *For all* $i\in[N]$, *the Jacobian* $J_{f(x_i,\theta)}(\theta)$, *the gradient* $\nabla_z L(y_i,z)|_{z=f(x_i,\theta)}$, *and the Hessian* $H_i(\theta) = \nabla_z^2 L(y_i,z)|_{z=f(x_i,\theta)}$ *are Lipschitz continuous over* $B(\Theta^*,b)$ *with respect to* θ. *There exist positive constants* $L_{\mathcal G}$ *and* L_J *such that for any* $\theta_1,\theta_2\in B(\Theta^*,b)$ *and any* $i\in[N]$,

$$\begin{aligned}&\|\nabla_z L(y_i,z)|_{z=f(x_i,\theta_1)} - \nabla_z L(y_i,z)|_{z=f(x_i,\theta_2)}\| \\ \le& L_{\mathcal G}\|\theta_1-\theta_2\|,\end{aligned} \tag{50}$$

and

$$\|J_{f(x_i,\theta)}(\theta_1) - J_{f(x_i,\theta)}(\theta_2)\|_F \le L_J\|\theta_1-\theta_2\|. \tag{51}$$

In addition, for any $i\in[N]$ *and* $\theta\in B(\Theta^*,b)$, *we assume*

$$\alpha_1 I \preceq H_i(\theta) = \nabla_z^2 L(y_i,z)|_{z=f(x_i,\theta)} \preceq \alpha_2 I, \tag{52}$$

where $0<\alpha_1<\alpha_2<\infty$ *are two constants, and for any* $\theta^*\in\Theta^*$,

$$\sigma_+\left(F(\theta^*)\right) > \bar\sigma > 0, \tag{53}$$

where $\sigma_+(F(\theta^*))$ *is the smallest positive eigenvalue of a* $F(\theta^*)\succeq 0$ *and* $\bar\sigma>0$ *is some constant.*

(b) $\mathcal G(\theta)$ *has zero residue on* Θ^*, *i.e.,* $\|\mathcal G(\bar\theta)\| = 0$ *for any* $\bar\theta\in\Theta^*$.

(c) *A local error bound condition holds for* $\mathcal G(\theta)$, *that is there exists* $\tau>0$ *such that for all* $\theta\in B(\Theta^*,b)$, *it has*

$$\frac{1}{\sqrt N}\|\mathcal G(\theta)\| \ge \tau\|\theta-\bar\theta\|. \tag{54}$$

(d) *The batch index sets* $\mathcal B_k^F = \mathcal B_k^g = \mathcal B_k \subset [N]$ *with* $|\mathcal B_k| = b_k$ *are chosen such that for any* $\theta_k\in B(\Theta^*,b)$,

$$P\left(E_k|\mathcal F_{k-1}^M\right) \ge (1-\delta_k),$$

with a constant $\delta_k\in[0,1)$ *and the event*

$$\begin{aligned}E_k = \Big\{&\lambda_k(\theta_k) \ge \rho\left(\frac{1}{\sqrt N}\|\mathcal G(\theta_k)\|\right)\\ &\text{and } \ \|F_k - F(\theta_k)\| \le L_F\lambda_k^2(\theta_k)\Big\},\end{aligned}$$

where $\rho>0$ *and* $L_F>0$ *are two constants,* $\lambda_k(\theta_k) = \frac{1}{\sqrt{b_k}}\|\mathcal G_k(\theta_k)\|$, $F(\theta)$ *and* $\mathcal G_k(\theta)$ *are defined in* (8) *and* (15), *respectively.*

Remark 5. *The smoothness conditions* (50) *and* (51) *in Assumption* 2 *(a) are satisifed if* f *and* H *are twice continuously differentiable. Remark* 4 *states that the boundedness of* H_i *in* (52) *is satisfied by the commonly used loss functions. The condition* (53) *holds if* Θ^* *is connected and compact. Assumption* 2 *(b) is satisfied if* $y_i = f(x_i,\theta)$, *which is the case when the number of parameters* m *is larger than the number of examples* N.

Remark 6. *Note that Assumption* 2 *(c) is weaker than assuming the positive definiteness of* $\nabla_\theta^2 h(\theta^*)$. *Moreover, we can see from* (49), *the definition of* $F(\theta)$ *in* (8) *and the definition of* $\mathcal G_k(\theta)$ *in* (15) *that for any* $\delta_k\in[0,1)$, *Assumption* 2 *(d) will hold by choosing the batch size* $|\mathcal B_k|$ *sufficiently large. Assumptions* 2 *(b) and (c) are used in* [48], [55] *for the Levenberg-Marquardt method. Assumption* 2 *(d) on the accuracies of the estimates of the gradient and the FIM is crucial in controlling the stochasticity in our NGM.*

Remark 7. *Due to the possibility of the FIM not being positive definite, the authors* [38] *explore a deterministic natural gradient method using the generalized inverse of the FIM and a step size. To ensure the well-posedness of the method, they require that the Jacobian* $J(\theta_k)$ *is full row-rank and that the loss function* L *is strongly convex and with Lipschitz continuous gradient. By assuming a stable-Jacobian condition, they obtained a linear convergence rate on the output* $u^k := f(x,\theta^k)$. *This is a first-order type convergence analysis on the output space, and it does not take into account the relationship between the FIM and the Hessian matrix. In addition, the convergence rate of* θ^k *cannot be determined based on their assumptions. In this work, we will show the quadratic convergence rate of the natural gradient method, Algorithm* 2, *by using the error bound condition. We note that the stable-Jacobian condition and the local error bound are independent of each other.*

Remark 8. *From the Lipschitz continuity of* $H_i(\theta)$ *in Assumption* 2 *(a) and* (62), $H_i^{1/2}(\theta)$ *is also Lipschitz continuous over the bounded set* $B(\Theta^*,b)$ *due to the differentiability of the square root of positive definite matrices. Furthermore, by the Lipschitz continuity of* $J_{f(x_i,\theta)}(\theta)$ *in Assumption* 2 *(a), one has the Lipschitz continuity of* $H_i^{\frac12}(\theta)J_{f(x_i,\theta)}(\theta)$ *and* $H_i(\theta)J_{f(x_i,\theta)}(\theta)$ *over* $B(\Theta^*,b)$, *namely, there exists positive constants* $L_{H^{1/2}J}$ *and* L_{HJ} *such that for any* $\theta_1,\theta_2\in B(\Theta^*,b)$ *and any* $i\in[N]$,

$$\begin{aligned}&\left\|H_i(\theta_1)^{1/2}J_{f(x_i,\theta)}(\theta_1) - H_i(\theta_2)^{1/2}J_{f(x_i,\theta)}(\theta_2)\right\| \\ \le& L_{H^{1/2}J}\,\|\theta_1-\theta_2\|,\end{aligned} \tag{55}$$

and

$$\left\|H_i(\theta_1)J_{f(x_i,\theta)}(\theta_1) - H_i(\theta_2)J_{f(x_i,\theta)}(\theta_2)\right\| \le L_{HJ}\,\|\theta_1-\theta_2\|. \tag{56}$$

Moreover, from (55), (56) *and the definitions of* $\mathcal H(\theta)$ *in* (8) *and* $\mathcal H_k$ *in* (49), *for any* $\theta_1,\theta_2\in B(\Theta^*,b)$ *and any* $i\in[N]$, *we can derive*

$$\left\|\mathcal H^{\frac12}(\theta_1)J(\theta_1) - \mathcal H^{\frac12}(\theta_2)J(\theta_2)\right\| \le L_{H^{1/2}J}\sqrt N\|\theta_1-\theta_2\|, \tag{57}$$

$$\left\|\mathcal H_k^{\frac12}(\theta_1)J_k(\theta_1) - \mathcal H_k^{\frac12}(\theta_2)J_k(\theta_2)\right\| \le L_{H^{1/2}J}\sqrt{b_k}\|\theta_1-\theta_2\|, \tag{58}$$

and

$$\|\mathcal{H}(\theta_1)J(\theta_1)-\mathcal{H}(\theta_2)J(\theta_2)\| \le L_{HJ}\sqrt{N}\|\theta_1-\theta_2\|, \tag{59}$$

$$\|\mathcal{H}_k(\theta_1)J_k(\theta_1)-\mathcal{H}_k(\theta_2)J_k(\theta_2)\| \le L_{HJ}\sqrt{b_k}\|\theta_1-\theta_2\|. \tag{60}$$

Furthermore, from (50) in Assumption 2 (a) and the definition of $\mathcal{G}_k(\theta)$ in (15) that

$$\|\mathcal{G}_k(\theta_k)-\mathcal{G}_k(\bar{\theta}_k)\| \le \sqrt{b_k}L_{\mathcal{G}}\|\theta_k-\bar{\theta}_k\|. \tag{61}$$

Finally, for any $\theta \in B(\Theta^*, b)$, by the definition of $\mathcal{H}_k$ in (49) and (52), we have

$$\alpha_1 I \preceq \mathcal{H}_k(\theta) \preceq \alpha_2 I, \tag{62}$$

where $0<\alpha_1<\alpha_2<\infty$ are given in Assumption 2 (a).

Firstly, let us validate the Assumption 2 (c) through the following proposition.

Proposition 2. *Consider the optimization problem*

$$\min_{\theta\in\mathbb{R}^n} \frac{1}{2N}\sum_{i=1}^{N}(y_i-f(x_i,\theta))^2. \tag{63}$$

(i) Assume that $y_i=f(x_i,\theta^*)$ for all (x_i,y_i), where $\theta^*\in\Theta^*$, and

$$\mathbb{E}_{x\sim Q_x}\left[J_{f(x,\theta)}^\top(\theta^*)J_{f(x,\theta)}(\theta^*)\right] \succeq \alpha I \tag{64}$$

for some $\alpha>0$. Then, for any $\delta\in(0,1)$, Assumption 2 (c) is satisfied with $c_1=\sqrt{\frac{\alpha}{2}}$ and probability $1-\delta$ if $N\in\mathbb{N}$ is sufficiently large.

(ii) In particular, for the choice

$$f(x,\theta)=a^\top\phi(W^\top x)=\sum_{j=1}^{d_1}a_j\phi(x^\top w_j) \tag{65}$$

with $a\in\mathbb{R}^{d_1}$, $W=[w_1,\dots,w_{d_1}]\in\mathbb{R}^{d\times d_1}$, $x\in\mathbb{R}^d$, $\theta=[w_1^\top,\dots,w_{d_1}^\top]^\top$, and ϕ is a smooth activation function (e.g., GELU [56]), suppose that the matrix $G(\theta^*)\in\mathbb{R}^{(dd_1)\times(dd_1)}$ with its (i,j)-th d-by-d block given by

$$[G(\theta^*)]_{ij}=\mathbb{E}_{x\sim Q_x}\left[a_ia_j\phi'(x^\top w_i^*)\phi'(x^\top w_j^*)xx^\top\right]$$

satisfies $G(\theta^*)\succeq\alpha I$ for some $\alpha>0$, where $\theta^*=[(w_1^*)^\top,\dots,(w_{d_1}^*)^\top]^\top$. Then, (64) holds. Hence, for any $\delta\in(0,1)$, Assumption 2 (c) is satisfied with $c_1=\sqrt{\frac{\alpha}{2}}$ and probability $1-\delta$ if $N\in\mathbb{N}$ is sufficiently large.

Proof. We first note that for the problem (63), we have

$$\begin{aligned}\mathcal{G}(\theta) &= \Big((f(x_1,\theta)-y_1)^\top,\dots,(f(x_N,\theta)-y_N)^\top\Big)^\top\\ &= \Big((f(x_1,\theta)-f(x_1,\theta^*))^\top,\dots,(f(x_N,\theta)-f(x_N,\theta^*))^\top\Big)^\top\\ &= \Big(\big(J_{f(x_1,\theta)}(\tilde{\theta})(\theta-\theta^*)\big)^\top,\dots,\big(J_{f(x_N,\theta)}(\tilde{\theta})(\theta-\theta^*)\big)^\top\Big)^\top\end{aligned}$$

with $\tilde{\theta}=\xi\theta+(1-\xi)\theta^*$ for some $\xi\in[0,1]$. Then, we obtain

$$\begin{aligned}&\frac{1}{N}\|\mathcal{G}(\theta)\|^2\\ =&(\theta-\theta^*)^\top\left(\frac{1}{N}\sum_{i=0}^{N}J_{f(x_i,\theta)}^\top(\tilde{\theta})J_{f(x_i,\theta)}(\tilde{\theta})\right)(\theta-\theta^*).\end{aligned}$$

By (64), there exists $\bar{N}\in\mathbb{N}$ and $b\in(0,1)$ such that if $N\ge\bar{N}$, then

$$\frac{1}{N}\sum_{i=0}^{N}J_{f(x_i,\theta)}^\top(\tilde{\theta})J_{f(x_i,\theta)}(\tilde{\theta})\succeq\frac{\alpha}{2}I$$

holds for all $\theta\in B(\theta^*,b)$ with probability $1-\delta$. This yields

$$\frac{1}{\sqrt{N}}\|\mathcal{G}(\theta)\|\ge\sqrt{\frac{\alpha}{2}}\|\theta-\theta^*\| \tag{66}$$

with probability $1-\delta$. Hence, Proposition 2 (i) holds.

Now, we show that (64) holds for the choice of f in (65). The derivative of f is $\nabla_{w_j}f(x,\theta)=a_j\phi'(x^\top w_j)x$. Hence, the 1-by-$dd_1$ Jacobian of $f(x,\theta)$ with respect to θ is

$$\begin{aligned}J_{f(x,\theta)}(\theta) &= [a_1\phi'(x^\top w_1)x^\top,\dots,a_{d_1}\phi'(x^\top w_{d_1})x^\top]\\ &= \hat{a}^\top\otimes x^\top,\end{aligned}$$

where $\otimes$ is the Kronecker product and $\hat{a}=\big(a_1\phi'(x^\top w_1),\dots,a_{d_1}\phi'(x^\top w_{d_1})\big)$. So, by direct calculation, we have

$$\begin{aligned}\mathbb{E}_{x\sim Q_x}\left[J_{f(x,\theta)}^\top(\theta^*)J_{f(x,\theta)}(\theta^*)\right] &= \mathbb{E}_{x\sim Q_x}\left[(\hat{a}\hat{a}^\top)\otimes xx^\top\right]\\ &= G(\theta^*)\succeq\alpha I.\end{aligned}$$

Then, Proposition 2 (ii) follows from Proposition 2 (i). □

Remark 9. *The distribution Q_x in Proposition 2 can be either discrete or continuous. For the finite datasets $(x_1,y_1),\dots,(x_S,y_S)$, taking Q_x as the uniform distribution over $\{x_1,\dots,x_S\}$, Assumption 2 (c) will hold for large N $(\le S)$ if $y_i=f(x_i,\theta^*)$ for all $i=1,\dots,S$ and $\frac{1}{S}\sum_{i=1}^{S}J_{f(x_i,\theta)}^\top(\theta^*)J_{f(x_i,\theta)}(\theta^*)\succeq\alpha I$. The positive definiteness condition in (64) basically corresponds to the strong convexity of a population-form of (63). Hence, the error bound condition used in Assumption 2 can be seen as a generalization of the strong convexity to a nonconvex problem.*

In the following, for notation simplicity, we let $\mathcal{H}_k=\mathcal{H}_k(\theta_k)$, $\lambda_k=\lambda_k(\theta_k)$ and $J_k=J_k(\theta_k)$ and define the function

$$\varphi_k(d):=\left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k+\mathcal{H}_k^{\frac{1}{2}}J_kd\right\|^2+b_k\lambda_k\|d\|^2. \tag{67}$$

Then, we can observe from (13) and (49) that the quadratic model defined in (12) can be rewritten as

$$m_k(d)=\frac{1}{2b_k}\varphi_k(d)-\frac{1}{2b_k}\left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\|^2$$

and

$$d_k=\arg\min_{d\in\mathbb{R}^n}m_k(d)=\arg\min_{d\in\mathbb{R}^n}\varphi_k(d).$$

The local quadratic convergence rate of Algorithm 2 can be shown in the following two steps: Firstly, by utilizing Assumptions 2 (a), (b) and (c), we show that the projections of generalized residual $\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k$ to the left singular vector space of $\mathcal{H}_k^{\frac{1}{2}}J_k$ can be controlled by $\|\theta_k-\bar{\theta}_k\|$ and $\|\theta_k-\bar{\theta}_k\|^2$. This further gives the bound of the direction d_k and ensures the iterations staying in the neighborhood $B(\Theta^*,b)$. Secondly, with the local error bound condition given in Assumption 2 (c), the distance to the optimal solution set $\|\theta_{k+1}-\bar{\theta}_{k+1}\|$ is controlled by the residual $\mathcal{G}(\theta_{k+1})$, which can be further related to the generalized residual $\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k$ by the Assumption 2 (a). Combining with the established estimation for the generalized residual and Assumption 2 (d), the quadratic convergence rate is then obtained.

We now derive some inequalities and set up some notations which will be used in our following proof of theorems. First, for any $\theta_k \in B(\Theta^*, b)$, we do the following eigenvalue decomposition of $F(\bar{\theta}_k) \succeq 0$:

$$F(\bar{\theta}_k) = \frac{1}{N} J(\bar{\theta}_k)^\top \mathcal{H}(\bar{\theta}_k) J(\bar{\theta}_k) = \bar{V}_k \bar{\Lambda}_k \bar{V}_k^\top,$$

where $\bar{V}_k^\top \bar{V}_k = I$ and $\bar{\Lambda}_k$ has the format

$$\bar{\Lambda}_k = \begin{pmatrix} \bar{\Lambda}_{k,1} & \\ & 0 \end{pmatrix}, \qquad \bar{\Lambda}_{k,1} = \mathrm{Diag}\left(\bar{\lambda}_{k,1}, \ldots, \bar{\lambda}_{k,r_k}\right) \succ 0,$$

and r_k is the rank of $F(\bar{\theta}_k)$. And here, by (53) and $\bar{\theta}_k \in \Theta^*$, we can assume that

$$\bar{\lambda}_{k,1} \geq \bar{\lambda}_{k,2} \geq \ldots \geq \bar{\lambda}_{k,r_k} \geq \bar{\sigma}. \tag{68}$$

Suppose that θ_k is sufficiently close to Θ^*. Then, from the Lipschitz continuity assumption in Assumption 2 (a), we can have the eigenvalue decomposition of $F(\theta_k) \succ 0$:

$$F(\theta_k) = \frac{1}{N} J(\theta_k)^\top \mathcal{H}(\theta_k) J(\theta_k) = \widetilde{V}_k \widetilde{\Lambda}_k \widetilde{V}_k^\top,$$

where $\widetilde{V}_k^\top \widetilde{V}_k = I$ and $\widetilde{\Lambda}_k$ has the format

$$\widetilde{\Lambda}_k = \begin{pmatrix} \widetilde{\Lambda}_{k,1} & \\ & \widetilde{\Lambda}_{k,2} \end{pmatrix}$$

with $\widetilde{\Lambda}_{k,1} = \mathrm{diag}\left(\widetilde{\lambda}_{k,1}, \ldots, \widetilde{\lambda}_{k,r_k}\right) \succ 0,\ \widetilde{\Lambda}_{k,2} \succeq 0$. Since it holds that

$$F(\bar{\theta}_k) = \left(\frac{1}{\sqrt{N}} \mathcal{H}^{\frac{1}{2}}(\bar{\theta}_k) J(\bar{\theta}_k)\right)^\top \left(\frac{1}{\sqrt{N}} \mathcal{H}^{\frac{1}{2}}(\bar{\theta}_k) J(\bar{\theta}_k)\right),$$
$$F(\theta_k) = \left(\frac{1}{\sqrt{N}} \mathcal{H}^{\frac{1}{2}}(\theta_k) J(\theta_k)\right)^\top \left(\frac{1}{\sqrt{N}} \mathcal{H}^{\frac{1}{2}}(\theta_k) J(\theta_k)\right),$$

by the relationships between the eigenvalue decomposition and singular value decomposition (SVD), we have the following SVDs:

$$\frac{1}{\sqrt{N}} \mathcal{H}^{\frac{1}{2}}(\bar{\theta}_k) J(\bar{\theta}_k) = \bar{W}_k \begin{pmatrix} (\bar{\Lambda}_{k,1})^{\frac{1}{2}} & \\ & 0 \end{pmatrix} \bar{V}_k^\top$$

and

$$\frac{1}{\sqrt{N}} \mathcal{H}^{\frac{1}{2}}(\theta_k) J(\theta_k) = \widetilde{W}_k \begin{pmatrix} \left(\widetilde{\Lambda}_{k,1}\right)^{\frac{1}{2}} & & \\ & \left(\widetilde{\Lambda}_{k,2}\right)^{\frac{1}{2}} & \\ & & 0 \end{pmatrix} (\widetilde{V}_k)^\top,$$

where $\bar{W}_k^\top \bar{W}_k = I$ and $\widetilde{W}_k^\top \widetilde{W}_k = I$. Here, we assume $\bar{\Lambda}_{k,1}$ has the same size with $\widetilde{\Lambda}_{k,1}$. Then, by the continuation of matrix singular values and (57), we have

$$\left\| \begin{pmatrix} \left(\widetilde{\Lambda}_{k,1}\right)^{\frac{1}{2}} - (\bar{\Lambda}_{k,1})^{\frac{1}{2}} & & \\ & \left(\widetilde{\Lambda}_{k,2}\right)^{\frac{1}{2}} & \\ & & 0 \end{pmatrix} \right\|$$
$$\leq \frac{1}{\sqrt{N}} \left\| \mathcal{H}^{\frac{1}{2}}(\bar{\theta}_k) J(\bar{\theta}_k) - \mathcal{H}^{\frac{1}{2}}(\theta_k) J(\theta_k) \right\|$$
$$\leq \frac{1}{\sqrt{N}} \sqrt{N} \cdot L_{H^{1/2}J} \|\theta_k - \bar{\theta}_k\| = L_{H^{1/2}J} \|\theta_k - \bar{\theta}_k\|.$$

So, for θ_k sufficiently close to Θ^* and (68), we have

$$\widetilde{\Lambda}_{k,1} \succeq \frac{3}{4} \bar{\Lambda}_{k,1} \succeq \frac{3\bar{\sigma}}{4} I \quad \text{and} \quad \left\| \widetilde{\Lambda}_{k,2} \right\| \leq L_{H^{1/2}J}^2 \|\theta_k - \bar{\theta}_k\|^2. \tag{69}$$

On the other hand, we also have the eigenvalue decomposition of $F_k = \frac{1}{b_k} J_k^\top \mathcal{H}_k J_k = V_k \Lambda_k V_k^\top$, where $V_k^\top V_k = I$,

$$\Lambda_k = \begin{pmatrix} \Lambda_{k,1} & \\ & \Lambda_{k,2} \end{pmatrix}$$

and the following SVD:

$$\frac{1}{\sqrt{b_k}} \mathcal{H}_k^{\frac{1}{2}} J_k = W_k \begin{pmatrix} \Lambda_{k,1}^{\frac{1}{2}} & & \\ & \Lambda_{k,2}^{\frac{1}{2}} & \\ & & 0 \end{pmatrix} V_k^\top, \tag{70}$$

where $W_k^\top W_k = I$. Here, we assume $\Lambda_{k,1}$ and $\Lambda_{k,2}$ have the same size with $\widetilde{\Lambda}_{k,1}$ and $\widetilde{\Lambda}_{k,2}$, respectively. By Assumption 2 (b), we have $\mathcal{G}_k(\bar{\theta}_k) = 0$ since $\bar{\theta}_k \in \Theta^*$. If the event E_k in Assumption 2 (d) happens, by (61), we have

$$\|F_k - F(\theta_k)\| \leq L_F \lambda_k^2(\theta_k) = L_F \cdot \frac{1}{b_k} \|\mathcal{G}_k(\theta_k) - \mathcal{G}_k(\bar{\theta}_k)\|^2$$
$$\leq L_F L_{\mathcal{G}}^2 \|\theta_k - \bar{\theta}_k\|^2.$$

Therefore, by the matrix eigenvalue perturbation theory [57], when the event E_k in Assumption 2 (d) happens, we have

$$\left\| \begin{pmatrix} \Lambda_{k,1} - \widetilde{\Lambda}_{k,1} & \\ & \Lambda_{k,2} - \widetilde{\Lambda}_{k,2} \end{pmatrix} \right\| \leq \|F_k - F(\theta_k)\|$$
$$\leq L_F L_{\mathcal{G}}^2 \|\theta_k - \bar{\theta}_k\|^2.$$

Hence, by (69), for θ_k sufficiently close to θ^*, we have

$$\Lambda_{k,1} \succeq \frac{1}{2} \bar{\Lambda}_{k,1} \succeq \frac{\bar{\sigma}}{2} I \quad \text{and} \tag{71}$$
$$\|\Lambda_{k,2}\| \leq \left\| \widetilde{\Lambda}_{k,2} \right\| + \|F_k - F(\theta_k)\| \tag{72}$$
$$\leq \left(L_{H^{1/2}J}^2 + L_F L_{\mathcal{G}}^2\right) \|\theta_k - \bar{\theta}_k\|^2.$$

Let $\Sigma_{k,1} = \sqrt{b_k} \Lambda_{k,1}^{\frac{1}{2}}$ and $\Sigma_{k,2} = \sqrt{b_k} \Lambda_{k,2}^{\frac{1}{2}}$. Denote $W_k = (W_{k,1}, W_{k,2}, W_{k,3})$ and $V_k = (V_{k,1}, V_{k,2}, V_{k,3})$. Then, we can rewrite the SVD (70) as

$$\mathcal{H}_k^{\frac{1}{2}} J_k = (W_{k,1}, W_{k,2}, W_{k,3}) \begin{pmatrix} \Sigma_{k,1} & & \\ & \Sigma_{k,2} & \\ & & 0 \end{pmatrix} \begin{pmatrix} V_{k,1}^\top \\ V_{k,2}^\top \\ V_{k,3}^\top \end{pmatrix}. \tag{73}$$

Now, let $b^* \in (0, b]$ be sufficiently small such that if $\theta_k \in B(\Theta^*, b^*)$ and the event E_k in Assumption 2 (d) happens, then the inequalities (69) and (71) hold for any $\theta_k \in B(\theta^*, b^*)$. Then, based on the above preliminary analysis, we have the following lemmas on the control of the projections of generalized residuals $\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k$.

Lemma 4. *Suppose that Assumption 2 (a), (b) and (c) hold, and $\theta_k \in B(\Theta^*, b^*)$. Then, we have*

$$\left\|W_{k,1}W_{k,1}^\top \mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\| \le \alpha_1^{-\frac{1}{2}} L_{\mathcal{G}}\sqrt{b_k}\cdot\|\theta_k - \bar\theta_k\|, \tag{74}$$

$$\left\|W_{k,3}W_{k,3}^\top \mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\| \le \alpha_1^{-\frac{1}{2}} L_{HJ}\sqrt{b_k}\cdot\|\theta_k - \bar\theta_k\|^2. \tag{75}$$

If, in addition, the event E_k in Assumption 2 (d) happens, we have

$$\begin{aligned}&\left\|W_{k,2}W_{k,2}^\top \mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\| \\ \le\ & \left(\alpha_1^{-\frac{1}{2}}L_{HJ} + L_F^{\frac{1}{2}}L_{\mathcal{G}} + L_{H^{\frac{1}{2}}J}\right)\sqrt{b_k}\|\theta_k - \bar\theta_k\|^2\end{aligned} \tag{76}$$

Proof. For the first inequality (74), we have from $W_{k,1}^\top W_{k,1} = I$, $\mathcal{G}_k(\bar\theta_k) = 0$, (62) and (61) that

$$\begin{aligned}&\left\|W_{k,1}W_{k,1}^\top \mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\| \le \left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\| \\ \le&\alpha_1^{-\frac{1}{2}}\|\mathcal{G}_k(\theta_k) - \mathcal{G}_k(\bar\theta_k)\| \le \alpha_1^{-\frac{1}{2}}L_{\mathcal{G}}\sqrt{b_k}\cdot\|\theta_k - \bar\theta_k\|.\end{aligned}$$

Defining $s_k = \arg\min_{s\in\mathbb{R}^n}\left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k s\right\|$, for the second inequality (75), we have from the SVD of $\mathcal{H}_k^{\frac{1}{2}}J_k$ in (73) that

$$\begin{aligned}&\left\|W_{k,3}W_{k,3}^\top \mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\| \le \left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k s_k\right\| \\ \le& \left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k(\bar\theta_k - \theta_k)\right\| \\ \le& \left\|\mathcal{H}_k^{-\frac{1}{2}}\right\|\|\mathcal{G}_k + \mathcal{H}_k J_k(\bar\theta_k - \theta_k)\| \\ \le&\alpha_1^{-\frac{1}{2}}L_{HJ}\sqrt{b_k}\cdot\|\theta_k - \bar\theta_k\|^2,\end{aligned}$$

where the last inequality is from (62) and (56).

Finally, we prove the third inequality (76). To this end, we define

$$\tilde s_k = \arg\min_{s\in\mathbb{R}^n}\left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \left(W_{k,1}\Sigma_{k,1}V_{k,1}^\top\right)s\right\|.$$

Then, due to $W_k^\top W_k = I$, it holds that

$$\begin{aligned}&\left\|W_{k,2}W_{k,2}^\top \mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k\right\| = \left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \left(W_{k,1}\Sigma_{k,1}V_{k,1}^\top\right)\tilde s_k\right\| \\ \le& \left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \left(W_{k,1}\Sigma_{k,1}V_{k,1}^\top\right)(\bar\theta_k - \theta_k)\right\| \\ \le& \left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k(\bar\theta_k - \theta_k)\right\| + \left\|\left(W_{k,2}\Sigma_{k,2}V_{k,2}^\top\right)(\bar\theta_k - \theta_k)\right\| \\ \le&\alpha_1^{-\frac{1}{2}}\sqrt{b_k}L_{HJ}\|\theta_k - \bar\theta_k\|^2 + \|\Sigma_{k,2}\|\cdot\|\theta_k - \bar\theta_k\|.\end{aligned} \tag{77}$$

If the event E_k in Assumption 2 (d) happens, by the choice of $\Sigma_{k,2}$ and (71), we have

$$\|\Sigma_{k,2}\| = \sqrt{b_k}\left\|\Lambda_{k,2}^{\frac{1}{2}}\right\| \le \sqrt{b_k}\left(L_{H^{1/2}J} + L_F^{\frac{1}{2}}L_{\mathcal{G}}\right)\|\theta_k - \bar\theta_k\|,$$

which together with (77) implies (76) holds. □

Then, we can bound the direction d_k by the distance of θ_k to the optimal solution set, i.e., $\|\theta_k - \bar\theta_k\|$. This ensures that θ_{k+1} remains within the neighborhood of Assumption 2 if θ_k is close enough to Θ^*.

Lemma 5. *Suppose that Assumption 2 (a), (b) and (c) hold, and $\theta_k \in B(\Theta^*, b^*)$. If the event E_k in Assumption 2 (d) happens, we have*

$$\|d_k\|^2 \le \frac{L_{HJ}^2}{\rho\tau\alpha_1}\|\theta_k - \bar\theta_k\|^3 + \|\theta_k - \bar\theta_k\|^2. \tag{78}$$

Moreover, if $\theta_k \in B(\Theta^*, \bar b)$, we have

$$\|d_k\| \le 2\|\theta_k - \bar\theta_k\|, \tag{79}$$

where $\bar b = \min\{b^*, \rho\tau\alpha_1/L_{HJ}^2\}$.

Proof. First, we observe from the definition of φ_k in (67) and $d_k = \arg\min_{d\in\mathbb{R}^n}\varphi_k(d)$ that

$$\begin{aligned}\|d_k\|^2 \le&\frac{\varphi_k(d_k)}{b_k\lambda_k} \le \frac{\varphi_k(\bar\theta_k - \theta_k)}{b_k\lambda_k} \\ =&\frac{\left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k(\bar\theta_k - \theta_k)\right\|^2}{b_k\lambda_k} + \|\theta_k - \bar\theta_k\|^2.\end{aligned} \tag{80}$$

Notice that

$$\left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k(\bar\theta_k - \theta_k)\right\|^2 \le b_k\alpha_1^{-1}L_{HJ}^2\|\bar\theta_k - \theta_k\|^4. \tag{81}$$

By Assumption 2 (c), when the event E_k in Assumption 2 (d) happens, we get

$$\lambda_k = \lambda_k(\theta_k) \ge \rho\left(\frac{1}{\sqrt{N}}\|\mathcal{G}(\theta_k)\|\right) \ge \rho\tau\|\theta_k - \bar\theta_k\|. \tag{82}$$

Then, (80)-(82) lead to (78). Finally, (79) follows from (78) and $\theta_k \in B(\Theta^*, \bar b)$ with $\bar b = \min\{b^*, \rho\tau\alpha_1/L_{HJ}^2\}$. □

Combining above and using Assumption 2 (d), we show the local quadratic convergence rate of Algorithm 2.

Theorem 2. *Suppose that Assumption 2 holds and $\theta_k \in B(\Theta^*, \bar b/3)$. Then, we have*

$$\tau\rho\|\theta_{k+1} - \bar\theta_{k+1}\| \le \hat C\ \|\theta_k - \bar\theta_k\|^2 \tag{83}$$

with probability at least $1-\delta_k$, where $\bar b > 0$ is defined in (79), and $\hat C$ is a constant given in (85).

Proof. Since

$$\begin{aligned}d_k = \arg\min_{d\in\mathbb{R}^n}\varphi(d) =& \left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k d\right\|_F^2 + b_k\lambda_k\|d\|^2 \\ =& -V_{k,1}\left(\Sigma_{k,1}^2 + b_k\lambda_k I\right)^{-1}\Sigma_{k,1}W_{k,1}^\top\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k \\ & - V_{k,2}\left(\Sigma_{k,2}^2 + b_k\lambda_k I\right)^{-1}\Sigma_{k,2}W_{k,2}^\top\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k,\end{aligned}$$

we obtain

$$\begin{aligned}&\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}}J_k d_k \\ =&\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k - W_{k,1}\Sigma_{k,1}\left(\Sigma_{k,1}^2 + b_k\lambda_k I\right)^{-1}\Sigma_{k,1}W_{k,1}^\top\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k \\ & - W_{k,2}\Sigma_{k,2}\left(\Sigma_{k,2}^2 + b_k\lambda_k I\right)^{-1}\Sigma_{k,2}W_{k,2}^\top\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k \\ =&W_{k,3}W_{k,3}^\top\mathcal{H}_k^{-1}\mathcal{G}_k + b_k\lambda_k W_{k,1}\left(\Sigma_{k,1}^2 + b_k\lambda_k I\right)^{-1}W_{k,1}^\top\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k \\ & + b_k\lambda_k W_{k,2}\left(\Sigma_{k,2}^2 + b_k\lambda_k I\right)^{-1}W_{k,2}^\top\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k.\end{aligned} \tag{84}$$

In the following, suppose the event E_k in Assumption 2 (d) happens. Then, by (71), we have $\Lambda_{k,1} \succeq \frac{\bar\sigma}{2}I$ holds. So, we have from $\Sigma_{k,1} = \sqrt{b_k}\Lambda_{k,1}^{\frac{1}{2}}$ that

$$\left\|\left(\Sigma_{k,1}^2 + b_k\lambda_k I\right)^{-1}\right\| \le \left\|\Sigma_{k,1}^{-2}\right\| = \frac{1}{b_k}\left\|\Lambda_{k,1}^{-1}\right\| \le \frac{2}{b_k\bar\sigma}$$

and

$$\left\|\left(\Sigma_{k,2}^2 + b_k\lambda_k I\right)^{-1}\right\| \le \frac{1}{b_k\lambda_k},$$

$$\lambda_k = \lambda_k(\theta_k) = \frac{1}{\sqrt{b_k}} \|\mathcal{G}_k(\theta_k)\| \leq L_{\mathcal{G}} \left\|\theta_k - \bar{\theta}_k\right\| .$$

Using the above inequalities, we have from (84) and Lemma 4 that

$$\begin{aligned}
&\left\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}} J_k d_k\right\| \\
\leq&\alpha_1^{-\frac{1}{2}} L_{HJ}\sqrt{b_k}\|\theta_k - \bar{\theta}_k\|^2 + \frac{2L_{\mathcal{G}}^2\alpha_1^{-\frac{1}{2}}}{\bar{\sigma}}\sqrt{b_k}\|\theta_k - \bar{\theta}_k\|^2 \\
&+ \left(\alpha_1^{-\frac{1}{2}} L_{HJ} + L_F^{\frac{1}{2}} L_{\mathcal{G}} + L_{H^{\frac{1}{2}}J}\right)\sqrt{b_k}\|\theta_k - \bar{\theta}_k\|^2.
\end{aligned}$$

Since $\theta_{k+1} = \theta_k + d_k$, when the event E_k in Assumption 2 (d) happens, we have from (79) that

$$\begin{aligned}
\|\theta_{k+1} - \bar{\theta}_{k+1}\| &\leq \|\theta_k - \bar{\theta}_{k+1}\| + \|d_k\| \\
&\leq \|\theta_k - \bar{\theta}_{k+1}\| + 2\|\theta_k - \bar{\theta}_k\| \leq 3\|\theta_k - \bar{\theta}_k\|,
\end{aligned}$$

which, by $\theta_k \in B(\Theta^*, \bar{b}/3)$ and $\bar{\theta}_{k+1} \in \Theta^*$, implies $\theta_{k+1} \in B(\Theta^*, \bar{b})$. Then, under the event E_k in Assumption 2 (d), it follows from Assumption 2 (c), $\|\mathcal{H}_k^{\frac{1}{2}}\| \leq \alpha_2^{\frac{1}{2}}$ by (62), and (79) that

$$\begin{aligned}
&\tau\rho\|\theta_{k+1} - \bar{\theta}_{k+1}\| \leq \frac{\rho}{\sqrt{N}}\|\mathcal{G}(\theta_{k+1})\| \leq \frac{1}{\sqrt{b_k}}\|\mathcal{G}_k(\theta_k + d_k)\| \\
\leq&\frac{1}{\sqrt{b_k}}\|\mathcal{G}_k + \mathcal{H}_k J_k d_k\| + L_{HJ}\|d_k\|^2 \\
\leq&\frac{\alpha_2^{\frac{1}{2}}}{\sqrt{b_k}}\|\mathcal{H}_k^{-\frac{1}{2}}\mathcal{G}_k + \mathcal{H}_k^{\frac{1}{2}} J_k d_k\| + L_{HJ}\|d_k\|^2 \leq \hat{C}\|\theta_k - \bar{\theta}_k\|^2,
\end{aligned}$$

where

$$\hat{C} = 2\sqrt{\frac{\alpha_2}{\alpha_1}}\left(L_{HJ} + L_{\mathcal{G}}^2/\bar{\sigma}\right) + \sqrt{\alpha_2}\left(L_F^{\frac{1}{2}} L_{\mathcal{G}} + L_{H^{\frac{1}{2}}J}\right) + 4L_{HJ}. \tag{85}$$

By Assumption 2 (d), the event E_k happens with at least probability $1 - \delta_k$. Hence, the inequality (83) holds with at least probability $1 - \delta_k$. This completes the proof. □

V. Numerical Experiments

In this section, we would like to perform some simple tests on examining both global and local convergence properties of our proposed NGM on the following logistic regression problem:

$$\min_{\theta \in \mathbb{R}^n} h(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \exp\left(-b_i\left(a_i^\top\theta\right)\right)\right) + \frac{\nu}{2}\|\theta\|^2, \tag{86}$$

where $\{(a_i, b_i)\}_{i=1}^N$, where $a_i \in \mathbb{R}^n$ and $b \in \{-1, 1\}$, is the dataset and $\nu > 0$ is the regularization parameter. In the numerical tests, ν is set to 0.01. Let $g_k = \frac{1}{|\mathcal{B}_k^g|}\sum_{i\in\mathcal{B}_k^g}\nabla_\theta \log\left(1 + \exp\left(-b_i\left(a_i^\top\theta\right)\right)\right) + \nu\theta$ and V_g be an upper bound of the variance of g_k. For comparisons, we also present the results of the probabilistic model based first-order method, STORM given by [40]. Default algorithmic parameters are used except for the initial trust-region radius Δ_0. We set $\Delta_0 = 0.8$ as it will give better performance than 1. By the Chebyshev's inequality [49, Exercise 4.1.2], we have for any $v > 0$,

$$P\left[\|g_k - \nabla h\left(\theta_k\right)\| \geq v \mid \mathcal{F}_{k-1}^{M,H}\right] \leq \frac{V_g}{|\mathcal{B}_k^g| v^2}.$$

TABLE II: A description of binary datasets. The integers N and n denote the number of the samples and the dimension of data, respectively.

Dataset	N	n	Reference
news20	19996	1355191	[58]
rcv1	20242	47236	[59]
SUSY	5000000	18	[60]
kdd	19264097	748401	[61]

Then, the condition (23) holds as long as $|B_k^g| \geq \frac{2V_g\sigma_k^2}{\kappa_g^2}$. We test two settings on the sample size for two algorithms, linearly increasing sample size $|B_k^g| = \lceil\min\{N, \max\{100k + 1000, \sigma_k^2\}\}\rceil$ and the exponentially increasing sample size $|B_k^g| = \lceil\min\{N, \max\{500 \cdot 1.8^k, \sigma_k^2\}\}\rceil$. To test the local quadratic convergence, we adopt a more accurate estimation on the objective function values, gradients, and FIMs by setting

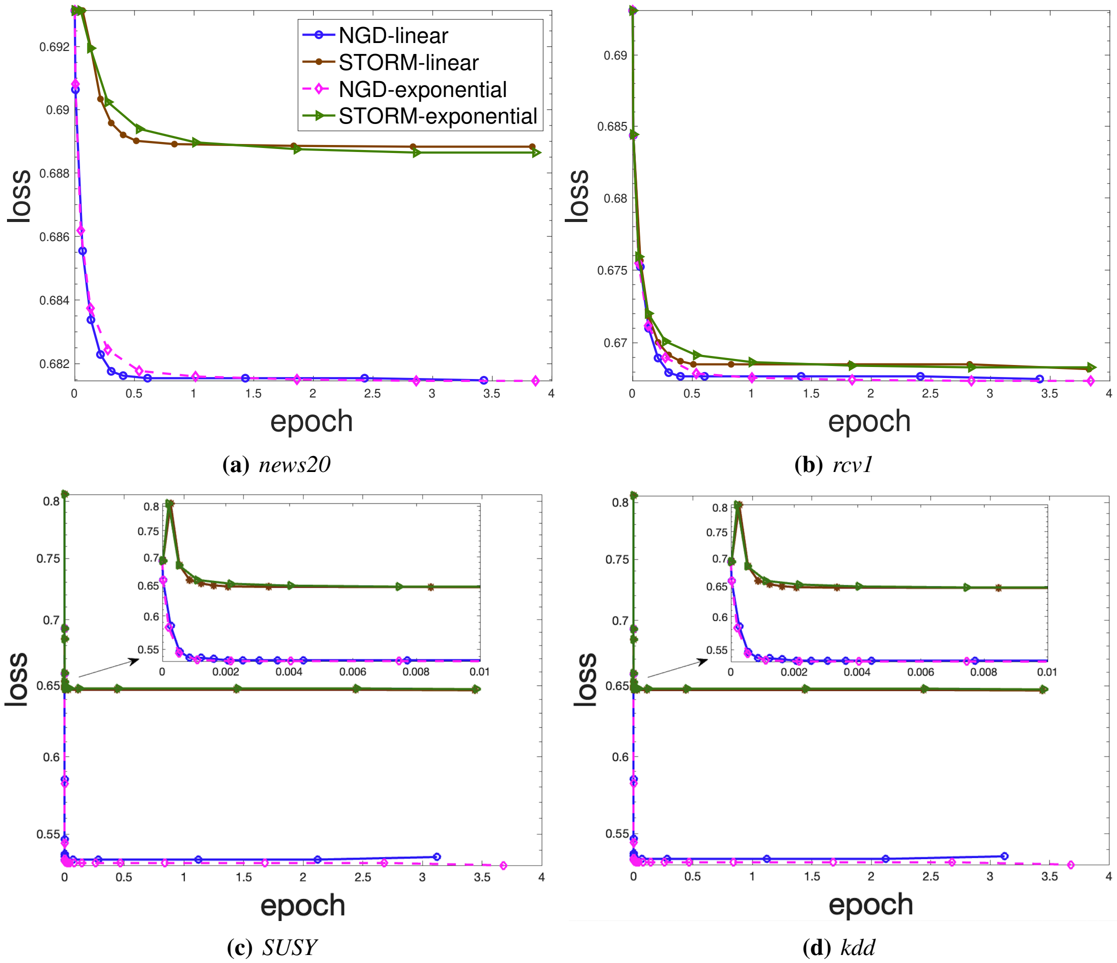
$$|B_k^g| = \lceil\min\{N, \max\{200, \lceil 1/\lambda_k^2\rceil\}\}\rceil, \tag{87}$$

where λ_k is set as $0.01\|\mathcal{G}_k\|$. When the sample size is given, we randomly draw the samples from $\{1, 2, \ldots, N\}$ without replacement. The datasets used with descriptions are presented in Table II.

For the implementation, we set $\eta_1 = 0.1, \eta_2 = 0.001, \gamma = 2$, $\sigma_{\min} = 10^{-8}$. The initial σ_0 is chosen as 1 and 0.01 to test the global and local convergence, respectively. We first run Algorithm 1 and then transit to Algorithm 2 when $\|g_k\| \leq 10^{-4}$. To ensure the boundedness of σ_k in the local phase, we add an extra projection, $\sigma_k = \min(\sigma_k, 10^{10})$. In this way, the resulting algorithm could enjoy both global convergence and fast local convergence as presented in our previous analysis. With both linearly and exponentially increasing sample sizes, we present the results in Figure 2. We can see our adaptively regularized NGM converges in both cases, while the exponentially sample strategy often converges to a point with lower objective function value. Compared with STORM, our NGD returns a point with a lower function value in much smaller epochs, which indicates the advantage of using Fisher information. We note that the per-epoch computational cost of NGD is higher than that of STORM, as computing the natural gradient direction involves solving a linear equation. For the sample size (87), Figure 3 clearly shows the quadratic convergence of the norms of gradients.

VI. Conclusion

Due to the computational efficiency of the FIMs, the natural gradient method (NGM) attracts much attention recently, while its global convergence and local convergence properties were not fully studied in the literature. We propose a trust region based adaptive regularization technique for ensuring global convergence of NGM under the assumption that the gradient and function evaluations are probabilistically sufficiently accurate. By utilizing the connections between the FIM and the GGN matrix and exploiting the properties of eigenvalue and singular value decompositions under the local error bound conditions, we show the quadratic convergence rate of our proposed method. Our numerical experiment on the logistic regression problems verifies the global and local convergence analysis results given in the paper.



(a) *news20* **(b)** *rcv1* **(c)** *SUSY* **(d)** *kdd*

Fig. 2: Numerical tests for linearly and exponentially increasing batchsizes, i.e., $|B_k^g| = \lceil \min\{N, \max\{100k + 1000, \sigma_k^2\}\}\rceil$ and $|B_k^g| = \lceil \min\{N, \max\{500 \cdot 1.8^k, \sigma_k^2\}\}\rceil$.

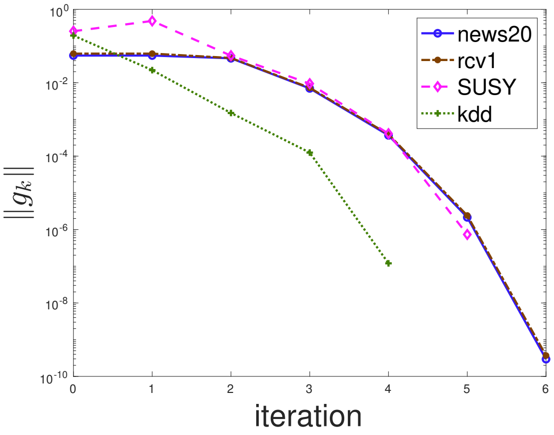


Fig. 3: Numerical results for $|B_k^g| = \lceil \min\{N, \max\{200, \lceil 1/\lambda_k^2 \rceil\}\}\rceil$. Local quadratic convergence rate is observed for all the datasets.

The effectiveness of our proposed NGM relies on the low computational cost of solving (12). As the FIM is high-dimensional, it is worthwhile to investigate more practical approximations with inexpensive inverses, such as the Kronecker-factored approximation [2].

Acknowledgement

The authors thank Dr. Johann Michael Schmitt for the discussion on the NGM and preparation of an early draft.

References

[1] J. Martens, "New insights and perspectives on the natural gradient method," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5776–5851, 2020.

[2] J. Martens and R. Grosse, "Optimizing neural networks with Kronecker-factored approximate curvature," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2408–2417.

[3] R. Grosse and J. Martens, "A Kronecker-factored approximate Fisher matrix for convolution layers," in *International Conference on Machine Learning*. PMLR, 2016, pp. 573–582.

[4] M. Zhao, Y. Li, and Z. Wen, "A stochastic trust-region framework for policy optimization," *Journal of Computational Mathematics*, vol. 40, no. 6, pp. 1004–1030, 2022.

[5] A. Mokhtari and A. Ribeiro, "RES: Regularized stochastic BFGS algorithm," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6089–6104, 2014.

[6] R. Zhao, W. B. Haskell, and V. Y. Tan, "Stochastic L-BFGS: Improved convergence rates and practical acceleration strategies," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1155–1169, 2017.

[7] D. Pfau, J. S. Spencer, A. G. Matthews, and W. M. C. Foulkes, "Ab initio solution of the many-electron schrödinger equation with deep neural networks," *Physical Review Research*, vol. 2, no. 3, p. 033429, 2020.

[8] F. Vicentini, D. Hofmann, A. Szabó, D. Wu, C. Roth, C. Giuliani, G. Pescia, J. Nys, V. Vargas-Calderón, N. Astrakhantsev *et al.*, "Netket 3: machine learning toolbox for many-body quantum systems," *SciPost Physics Codebases*, p. 007, 2022.

[9] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[10] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[11] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in neural information processing systems*, vol. 27, 2014.

[12] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of Machine Learning Research*, vol. 12, no. 7, 2011.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] S.-i. Amari, "Neural learning in structured parameter spaces-natural Riemannian gradient," in *Advances in Neural Information Processing Systems*, vol. 9, 1996.

[16] M. R. Osborne, "Fisher's method of scoring," *International Statistical Review/Revue Internationale de Statistique*, pp. 99–117, 1992.

[17] M. Yang, D. Xu, Z. Wen, M. Chen, and P. Xu, "Sketch-based empirical natural gradient methods for deep learning," *Journal of Scientific Computing*, vol. 92, no. 3, pp. 1–29, 2022.

[18] R. Anil, V. Gupta, T. Koren, K. Regan, and Y. Singer, "Scalable second order optimization for deep learning," *arXiv:2002.09018*, 2020.

[19] M. Yang, D. Xu, Q. Cui, Z. Wen, and P. Xu, "An efficient Fisher matrix approximation method for large-scale neural network optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[20] A. Bahamou, D. Goldfarb, and Y. Ren, "A mini-block natural gradient method for deep neural networks," *arXiv:2202.04124*, 2022.

[21] L. Nurbekyan, W. Lei, and Y. Yang, "Efficient natural gradient descent methods for large-scale optimization problems," *arXiv:2202.06236*, 2022.

[22] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer, "A stochastic quasi-newton method for large-scale optimization," *SIAM Journal on Optimization*, vol. 26, no. 2, pp. 1008–1031, 2016.

[23] M. Pilanci and M. J. Wainwright, "Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 205–245, 2017.

[24] F. Roosta-Khorasani and M. W. Mahoney, "Sub-sampled newton methods," *Mathematical Programming*, vol. 174, pp. 293–326, 2019.

[25] D. Goldfarb, Y. Ren, and A. Bahamou, "Practical quasi-newton methods for training deep neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2386–2396, 2020.

[26] M. Yang, D. Xu, H. Chen, Z. Wen, and M. Chen, "Enhance curvature information by structured stochastic quasi-newton methods," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 654–10 663.

[27] J.-C. Culioli and G. Cohen, "Decomposition/coordination algorithms in stochastic optimization," *SIAM Journal on Control and Optimization*, vol. 28, no. 6, pp. 1372–1403, 1990.

[28] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[29] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *International Conference on Machine Learning*, 2013.

[30] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *International Conference on Machine learning*, 2004.

[31] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[32] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," *arXiv:1904.09237*, 2019.

[33] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, "SGD: General analysis and improved rates," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5200–5209.

[34] X. Li and A. Milzarek, "A unified convergence theorem for stochastic optimization methods," *arXiv:2206.03907*, 2022.

[35] B. T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 3, no. 4, pp. 643–653, 1963.

[36] R. Bassily, M. Belkin, and S. Ma, "On exponential convergence of sgd in non-convex over-parametrized learning," *arXiv:1811.02564*, 2018.

[37] X. Li, A. Milzarek, and J. Qiu, "Convergence of random reshuffling under the Kurdyka-Lojasiewicz inequality," *arXiv:2110.04926*, 2021.

[38] G. Zhang, J. Martens, and R. B. Grosse, "Fast convergence of natural gradient descent for over-parameterized neural networks," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[39] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 1999.

[40] R. Chen, M. Menickelly, and K. Scheinberg, "Stochastic optimization using a trust-region method and random models," *Mathematical Programming*, vol. 169, no. 2, pp. 447–487, 2018.

[41] N. N. Schraudolph, "Fast curvature matrix-vector products for second-order gradient descent," *Neural computation*, vol. 14, no. 7, pp. 1723–1738, 2002.

[42] H. Park, S.-I. Amari, and K. Fukumizu, "Adaptive natural gradient learning algorithms for various stochastic models," *Neural Networks*, vol. 13, no. 7, pp. 755–764, 2000.

[43] F. Kunstner, P. Hennig, and L. Balles, "Limitations of the empirical fisher approximation for natural gradient descent," *Advances in neural information processing systems*, vol. 32, 2019.

[44] J. Martens, J. Ba, and M. Johnson, "Kronecker-factored curvature approximations for recurrent neural networks," in *International Conference on Learning Representations*, 2018.

[45] A. S. Berahas, L. Cao, and K. Scheinberg, "Global convergence rate analysis of a generic line search algorithm with noise," *SIAM Journal on Optimization*, vol. 31, no. 2, pp. 1489–1518, 2021.

[46] V. Roulet, S. Srinivasa, M. Fazel, and Z. Harchaoui, "Complexity bounds of iterative linear quadratic optimization algorithms for discrete time nonlinear control," *arXiv preprint arXiv:2204.02322*, 2022.

[47] R. Zhao and J. Fan, "On a new updating rule of the Levenberg–Marquardt parameter," *Journal of Scientific Computing*, vol. 74, no. 2, pp. 1146–1162, 2018.

[48] ——, "Levenberg–Marquardt method based on probabilistic Jacobian models for nonlinear equations," *Computational Optimization and Applications*, pp. 1–21, 2022.

[49] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2019.

[50] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg, "Convergence rate analysis of a stochastic trust-region method via supermartingales," *INFORMS Journal on Optimization*, vol. 1, no. 2, pp. 92–119, 2019.

[51] L. Ma, G. Montague, J. Ye, Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney, "Inefficiency of k-fac for large batch size training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5053–5060.

[52] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[53] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," in *International Conference on Learning Representations*, 2018.

[54] J. Zhang and L. Xiao, "Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization," *Mathematical Programming*, pp. 1–43, 2021.

[55] E. H. Bergou, Y. Diouane, and V. Kungurtsev, "Convergence and complexity analysis of a levenberg–marquardt algorithm for inverse problems," *Journal of Optimization Theory and Applications*, vol. 185, pp. 927–944, 2020.

[56] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[57] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*. San Diego: Academic Press, 1990.

[58] S. S. Keerthi, D. DeCoste, and T. Joachims, "A modified finite Newton method for fast solution of large scale linear SVMs." *Journal of Machine Learning Research*, vol. 6, no. 3, 2005.

[59] D. D. Lewis, Y. Yang, T. Russell-Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, no. Apr, pp. 361–397, 2004.

[60] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, vol. 5, no. 1, p. 4308, 2014.

[61] Y. Juan, Y. Zhuang, W.-S. Chin, and C.-J. Lin, "Field-aware factorization machines for ctr prediction," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 43–50.