A fast and efficient randomized quasi-Newton method

Danny Duan BDUAN5@WISC.EDU
Hanbaek Lyu HLYU@MATH.WISC.EDU
Department of Mathematics, University of Wisconsin - Madison, USA

Abstract

We propose a novel randomized quasi-Newton method that scales well with problem dimension by leveraging a recent randomized low-rank Hessian approximation technique. Our algorithm achieves the seemingly exclusive benefits of the first-order and second-order methods. The iteration cost of our algorithm scales linearly with the problem dimension, as with the first-order methods. For non-convex smooth objectives, our algorithm globally converges to a stationary point with convergence rate $O(n^{-1/2})$, matching that of the standard gradient descent with an improved implicit constant. When the Hessian of the objective near a local minimum has a good low-rank approximation, our algorithm can leverage such local structure and achieve a linear local convergence with a rate superior to that of standard gradient descent. If the Hessian is actually low-rank, our algorithm achieves superlinear local convergence. We verify our theoretical results with various numerical experiments.

1. Introduction

Consider the nonconvex smooth minimization problem $\boldsymbol{\theta}_* \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N} f(\boldsymbol{\theta})$, where $f : \mathbb{R}^N \to \mathbb{R}$ is a twice continuously differentiable nonconvex function that is bounded from below. Many classical optimization algorithms for this problem take the following form of 'preconditioned gradient descent':

$$\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n - \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n), \tag{1}$$

where $\mathbf{B}_n \in \mathbb{R}^{N\times N}$ is a suitable preconditioning matrix. If $\mathbf{B}_n = \alpha^{-1}\mathbf{I}_N$, where $\alpha > 0$ is a fixed stepsize and I_N is the identity matrix, then (1) becomes the standard gradient descent (GD). Assuming a gradient oracle, GD has $O(N)$ per-iteration complexity. It has sublinear global convergence for general nonconvex objectives. Near a local minimizer, GD converges at a linear rate depending on the condition number of the local landscape (see, e.g., [1, 20]). This can be very slow when the Hessian of the objective is ill-conditioned. When $\mathbf{B}_n = \nabla^2 f(\boldsymbol{\theta}_n)$, provided that the Hessian is invertible, then (1) becomes the classical Newton's method (see, e.g., [21, 23]). It has quadratic local convergence near a local minimum with positive definite Hessian [15, 33]. However, it has a high per-iteration complexity of $O(N^3)$ for inverting the Hessian, and the Newton step has infeasibility and instability issues when the Hessian is singular or ill-conditioned. In this work, we use the term *Quasi-Newton methods* to mean a broad class of optimization algorithms (1) where $\mathbf{B}_n$ is some approximation of the Hessian with a suitable regularization. These methods aim to remedy the drawbacks of Newton's methods while maintaining the advantages of using second-order information. For example, Levenberg-Marquardt (LM) regularization uses the Newton step with proximal regularization $\frac{\tau_n}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\|^2$, which reduces to the iterate (1) with $\mathbf{B}_n = \nabla^2 f(\boldsymbol{\theta}_n)+\tau_n\mathbf{I}_N$, where $\tau_n \geq 0$ is large enough so that the regularized Hessian is positive definite. The celebrated cubic-regularized quasi-Newton by Nestrov and Polyak [12] instead uses the cubic proximal term

$\frac{\tau_n}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^3$. Besides, trust-region methods [7] proceeds by computing $\boldsymbol{\theta}_n$ via minimizing approximate second-order Taylor expansion of the objective within a 'trust-region' $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \le \Delta_n$ with adaptively chosen radii Δ_n [7]. These quasi-Newton methods enjoy global convergence properties and superlinear local convergence (see Table 1). However, they have a high iteration cost of at least $O(N^2)$. Other quasi-Newton methods such as BFGS [3, 9, 10, 30] and its limited-memory variant (L-BFGS) [17] have lower per-iteration complexity but they do not enjoy global convergence to first-order optimal points or superlinear local convergence.

Randomized Low-rank Quasi-Newton. In this work, we propose a new quasi-Newton algorithm based on a randomized low-rank matrix approximation, *Randomly Pivoted Cholesky (RPC) decomposition*, developed recently by Chen, Epperly, Tropp, and Webber [6]. Given a positive semi-definite matrix $\mathbf{A}$ and a rank parameter k, RPC obtains a low-rank *random* Cholesky decomposition $\mathbf{F}\mathbf{F}^T \approx \mathbf{A}$, where $\mathbf{F}$ is a rank-k random Cholesky factor while only using the diagonal entries and at most k columns of $\mathbf{A}$. Our key idea is to use RPC to obtain a low-rank approximation of (regularized) Hessian within the LM framework by only computing $O(kN)$ entries in the Hessian. This method can be seen as an efficient randomized variant of the quasi-Newton method of Bräuninger [2] that uses full Cholesky decomposition of the regularized Hessian. A high-level description of our algorithm, *Randomized Low-rank Quasi-Newton* (RLQN), is shown below (see Sec. 2 for details):

$$\textbf{(RLQN)} \quad \begin{cases} \mathbf{F}_n\mathbf{F}_n^T \approx \nabla^2 f(\boldsymbol{\theta}_n) + \tau_n \mathbf{I} & (\triangleright \textit{ Randomized low-rank Hessian approx.}) \\ \mathbf{B}_n \leftarrow \mathbf{F}_n\mathbf{F}_n^T + \delta_n \mathbf{I}_N & (\triangleright \textit{ Preconditioning matrix}) \\ \boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n - \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) & (\triangleright \textit{ Parameter update}). \end{cases} \tag{2}$$

Here, τ_n, δ_ns are the LM regularization coefficients found adaptively by Alg. 1 and Alg. 3.

Notation. For real symmetric matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ and an integer $1 \le r \le N$, let $[\![\mathbf{A}]\!]_r$ denote the best rank-r approximation of $\mathbf{A}$: namely if $\mathbf{A} = \sum_{i=1}^N \lambda_i \mathbf{u}_i\mathbf{u}_i^T$ is the spectral decomposition of $\mathbf{A}$ with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_N$, then $[\![\mathbf{A}]\!]_r = \sum_{i=1}^r \lambda_i \mathbf{u}_i\mathbf{u}_i^T$. Denote $\log_+(x) = \max\{0, \log x\}$.

Related works and Contribution. An extensive comparison between our method and various existing methods is given in Table 1.

Methods	*Global conv.*	*Local convergence*		*per-iter cost*	*iteration complexity*
		Property	*Rate*		
GD[1, 20]	✓	strongly convex	Linear $O((1-\frac{\mu}{L})^n)$	$O(N)$	$O(\epsilon^{-2})$
Newton [15]	✗	pos. def.	quadratic	$O(N^3)$	NA
Trust Region [11, 26]	✓	pos. def.	quadratic	$O(N^3)$	$O(\varepsilon^{-2})$
			superlinear	$O(N^2)$	
LM [8]	✓	Error bound (6)	quadratic	$O(N^3)$	$O(\varepsilon^{-2})$
Cubic regularization [4, 5, 19]	✓	PL (6)	quadratic	$O(N^2 j)$	$O(\varepsilon^{-3/2})$
BFGS [14, 28, 29]	✗	pos. def.	superlinear	$O(N^2)$	NA
L-BFGS [17, 22]	✗	pos. def.	linear	$O(Nm)$	NA
RLQN (Ours)	✓	low-rank	superlinear	$O(k^2N)$	$O(\epsilon^{-2})$ (see Thm. 1)
		approximate low-rank	Linear $O((1-\frac{\mu}{L^{\text{eff}}})^n)$		

Table 1: Comparison between the proposed Randomized Low-rank Quasi-Newton (RLQN) method and various benchmark methods. j in the cubic regularization row stands for the number of iterations used in the Lanczos algorithm. L^{eff} denotes the 'effective' smoothness parameter that can be much less than L, see Thm. 2.

In this work, we show that our RLQN has the following theoretical properties:

- *Nearly first-order per-iteration complexity* $O(k^2N)$

- *Global convergence to stationary points with convergence rate that of gradient descent* $O(n^{-1/2})$ *with a possible improvement of the implicit constant.* (Thm. 1, Rmk. 13)
- *Improved local linear convergence with approximately low-rank local landscape* in expectation (Thm. 2) and high-probability (Cor. 3)
- *Superlinear local convergence with low-rank local landscape* (Thm. 4).

To the best of our knowledge, our RLQN is the first quasi-Newton method in the literature that achieves these desirable properties all at once. We also demonstrate the superior numerical performance of RLQN for solving regularized matrix factorization against standard gradient descent and L-BFGS, and for solving large linear systems against state-of-the-art methods such as Randomized Kaczmarz [31] and conjugate gradient [13].

2. Statement of the algorithm

Below we give details on our RLQN algorithm in (2). An auxiliary algorithm for randomized low-rank Hessian approximation (Alg. 3) is given in Appendix A.

Algorithm 1 Randomized Low-rank Quasi-Newton (RLQN)

Input: $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ (initial estimate)
Parameters: M (number of iterations); k (lower rank parameter); L (Lipchitz constant of the gradient); L_H (Lipchitz constant of the Hessian)
  for $n = 0, 1, \cdots, M$ **do**
    $\mathbf{F}_n, R_n \leftarrow$ output of Algorithm 3 with input $(\boldsymbol{\theta}_n, k, L)$ (▷ *RP Cholesky [6] with LM-reg.*)
    $\delta_n \leftarrow \min\{L, \max\{R_n, \sqrt{L_H \|\nabla f(\boldsymbol{\theta}_n)\|}\}\}$
    $p_n \leftarrow -\left(\mathbf{F}_n \mathbf{F}_n^T + \delta_n \mathbf{I}\right)^{-1} \nabla f(\boldsymbol{\theta}_n)$
      $= \delta_n^{-1} \nabla f(\boldsymbol{\theta}_n) - \delta_n^{-1} \mathbf{F} (\delta_n \mathbf{I}_k + \mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \nabla f(\boldsymbol{\theta}_n)$ (▷ *Woodbury identity [32]; Cost* $O(k^2 N)$)
    set new iterate $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n + p_n$
output $\boldsymbol{\theta}_{M+1}$

3. Main results

We provide both global and local convergence analysis of our RLQN algorithm (2). For both analysis, we assume that the objective has Lipschitz continuous Hessian as below:

A1 (Smooth objective) *The objective* $f : \mathbb{R}^N \to \mathbb{R}$ *is uniformly lower bounded and is twice continuously differentiable, and for some* $L, L_H > 0$, ∇f, $\nabla^2 f$ *satisfy*

$$\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\| \le L \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \quad \|\nabla^2 f(\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}')\| \le L_H \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \tag{3}$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{L}$, *where* $\mathcal{L}$ *is a convex set containing the sublevel set* $\{\boldsymbol{\theta} : f(\boldsymbol{\theta}) \le f(\boldsymbol{\theta}_0)\}$.

Under this mild assumption, we establish the following first-order global asymptotic convergence and convergence rate of the general preconditioned gradient descent (1) in Thm 1 below.

Theorem 1 (Global convergence for preconditioned GD 1) *Suppose A1 holds. Let* $(\boldsymbol{\theta}_n)_{n \ge 0}$ *be outputs generated by* (1), *with* $\mathbf{B}_n \succcurlyeq \nabla^2 f(\boldsymbol{\theta}_n)$, $\lambda_{\min}(\mathbf{B}_n) \ge \min\{L, \sqrt{C L_H \|\nabla f(\boldsymbol{\theta}_n)\|}\}$ *for some* $C > \frac{1}{3}$, *and* $\sup_{n \ge 0} \|\mathbf{B}_n\| < \infty$. *Then almost surely* $\min_{0 \le k \le n-1} \|\nabla f(\boldsymbol{\theta}_k)\| \le \left(\frac{(f(\boldsymbol{\theta}_0) - \inf f)}{A \sum_{k=0}^{k-1} \|\mathbf{B}_k\|^{-1}} \right)^{1/2}$, *where* $A = \frac{1}{2} - \frac{1}{6C} > 0$. *And asymptotically* $\lim_{n \to \infty} \|\nabla f(\boldsymbol{\theta}_n)\| = 0$.

The hypothesis in Theorem 1 is satisfied for the standard GD with constant stepsize ($\mathbf{B}_n = \alpha \mathbf{I}$, $\alpha \ge L$) and GD with diminishing step size ($\mathbf{B}_n = \sqrt{n} \mathbf{I}$ when n is sufficiently large). It is also suitable for algorithms with line search since one can consider the stepsize to be a constant multiplied by $\mathbf{B}_n$. Most importantly, our RLQN chooses the preconditioning matrices $\mathbf{B}_n$ that satisfy the

hypothesis of Theorem 1. Hence RLQN has asymptotic convergence to stationary points and iteration complexity as good as the standard GD, but can exhibit much faster convergence if our adaptive choice of $\mathbf{B}_n$ yields a large sum of the inverse spectral norms. See Rmk.13 for more discussion.

Next, we will see that local convergence rates of RLQN have a more direct improvement by randomized Hessian approximation. Classical analysis of quasi-Newton methods concerns linear convergence toward a local minimizer within a strongly convex local landscape. In our analysis, we extend this to 'rank-deficient flat' minima, where Hessian near a local minimizer can be rank-deficient (A2) and satisfy a local curvature condition through the PL inequality (A3). (See Appendix B for further discussion on our assumptions.)

A2 (Approximately low-rank local landscape) *Suppose $f : \mathbb{R}^N \to \mathbb{R}$ satisfies A1. For a local minimizer $\boldsymbol{\theta}_*$, there is an open ball U centered at $\boldsymbol{\theta}_*$ and constants $\mu > 0$, $\rho \in (0, 1)$, and $1 \le r \le N$ such that for all $\boldsymbol{\theta} \in U$, $\nabla^2 f(\boldsymbol{\theta}) \succeq \mu \mathbf{I}_N$, and $\operatorname{tr}(\nabla^2 f(\boldsymbol{\theta}) - [\![\nabla^2 f(\boldsymbol{\theta})]\!]_r) \le \rho\, \lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}))$.*

A3 (Second-order accumulated flat local minima) *Suppose A1 holds. Fix a local minimizer $\boldsymbol{\theta}_*$ of f and denote $S(\boldsymbol{\theta}_*) := \{\boldsymbol{\theta}'_* : f(\boldsymbol{\theta}'_*) = f(\boldsymbol{\theta}_*)$ and $\boldsymbol{\theta}'_*$ is a local minimizer of $f\}$. Then there exists constants $\mu > 0$, $k \le N$, and an open ball U centered at $\boldsymbol{\theta}_*$ with the following properties:*

(i) *(Rank-deficiency) For all $\boldsymbol{\theta} \in U$, $\nabla^2 f(\boldsymbol{\theta})$ is PSD and has rank at most k.*

(ii) *(Local-curvature) For any $\boldsymbol{\theta}'_* \in U \cap S(\boldsymbol{\theta}_*)$, $\boldsymbol{\theta}'_*$ is μ-PL [25], namely $f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}'_*) \le \frac{1}{2\mu}\|\nabla f(\boldsymbol{\theta})\|^2$ for all $\boldsymbol{\theta} \in U$.*

A key benefit of having a low-rank approximation of the Hessian is that the condition number of the local landscape automatically improves depending on the accuracy of the low-rank Hessian approximation. We establish this improved local convergence result in Theorem 2 below.

Theorem 2 (Linear local convergence with improved condition number) *Suppose A1 holds. and $\boldsymbol{\theta}_0$ is sufficiently close to a local minimizer $\boldsymbol{\theta}_*$ satisfying A2. Let L, μ, ρ, r be constants in A1 and A2. Let $\varepsilon_0 > 0$ be the smallest so that $k \ge r\left(\frac{1}{\varepsilon_0} + 1 + \log_+\left(\frac{2r}{\varepsilon_0}\right)\right)$, where k is the number of columns of the Hessian sampled in RLQN (see Alg. 2). Then for all $n \ge 1$,*

$$\mathbb{E}[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_*\|] \le \left(1 - \frac{\mu}{4\mu + 8(1+\varepsilon_0)\rho L}\right)^n \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|, \tag{4}$$

where the expectation is taken with respect to all the random choice of pivot columns in the algorithm up to step n. Furthermore, $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_*$ almost surely.

The contraction constant in (4) should be compared with that for GD with fixed stepsize $1/L$ for μ-strongly convex and L-smooth objectives, for which we have $\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_*\| \le (1 - \mu/L)^{n/2}\, \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$. Thus our linear local convergence rate in (4) is significantly faster than that for GD for ill-conditioned ($\mu \ll L$) local landscape with approximately low-rank Hessian as in A2 ($r \ll N$ and $\rho \ll 1$).

From Theorem 2, we can also obtain a high-probability local convergence guarantee that we can reach a local minimizer within distance ε in $O(\log \varepsilon^{-1})$ iterations with better implied constant.

Corollary 3 (Local iteration complexity) *Suppose A1 and A2 hold, and $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$ is sufficiently small. Let ε_0 be the same as in theorem 2, then there is some $n \le 5(\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu} \log \varepsilon^{-1})$ so that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \le \varepsilon$ with probability at least $1 - \exp\left(-\frac{1}{40\mu}(\log \varepsilon^{-1} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|)\,(\mu + 2(1+\varepsilon_0)\rho L)\right)$.*

If the local landscape is in fact low-rank and if the rank parameter k in our RLQN is sufficiently large, then we obtain superlinear local convergence, as stated in the following result.

Theorem 4 (Superlinear local convergence with low-rank local landscape) *Suppose $\boldsymbol{\theta}_0$ is sufficiently close to a local minimum $\boldsymbol{\theta}_*$ of f satisfying A3, then $\|\nabla f(\boldsymbol{\theta}_n)\|$ converges to 0 superlinearly in n, in particular $\|\nabla f(\boldsymbol{\theta}_{n+1})\| = O\left(\mu^{-1}\|\nabla f(\boldsymbol{\theta}_n)\|^{3/2}\right)$ with μ as in A3.*

4. Numerical Experiments: Matrix factorization

We test our algorithm on solving an ℓ_2 regularized matrix factorization problem. We fixed a matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$, and try to find its best rank-r approximation $\mathbf{X} \approx \mathbf{WH}$ in terms of Frobenius norm, where $\mathbf{W} \in \mathbb{R}^{p\times r}$, $\mathbf{H} \in \mathbb{R}^{r\times n}$. This can be formalized in minimizing the nonconvex objective $f(\mathbf{W},\mathbf{H}) = \|\mathbf{X} - \mathbf{WH}\|_F^2 + \frac{\lambda}{2}(\|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2)$, where $\lambda > 0$ is the ℓ_2-regularization penalty. In our experiments $\mathbf{X} \in \mathbb{R}^{1000\times 10}$ is generated by sampling i.i.d. standard normal entries and we applied rank-2 factorization. Hyperparameters for GD and L-BFGS are chosen by cross-validation so that they perform competitively.

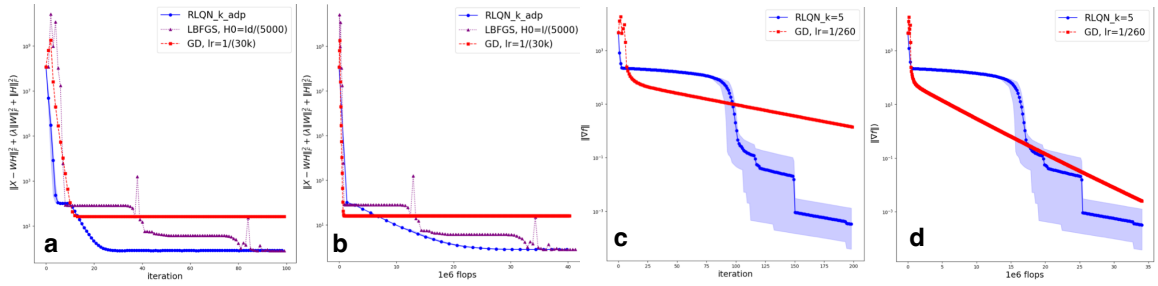

Figure 1: Comparison of RLQN against L-BFGS and GD for matrix factorization. (**a**)-(**b**) Objective value vs. iterations and flops with $\lambda = 10^{-3}$; RLQN uses low-rank ($k = 5$) Hessian approximation until iteration 5 and more precise Hessian approximation ($k = r\min\{p, n\} + 1 = 21$) after that. (**c**)-(**d**) Gradient norm v.s. iterations and flops, respectively, with $\lambda = 5$. RLQN uses constant low-rank Hessian approximation ($k = 5$). Shaded blow regions indicate one standard deviation over five random initializations of RLQN.

In Figure 1 we compare our algorithm (RLQN) with respect to L-BFGS and gradient descent (GD). We find that a practical implementation of RLQN is to use small k for the first few iterations so that it converges to a stationary point with low per-iteration cost (comparable to GD) and then later use large k so that we leverage improved local convergence rate of RLQN by good Hessian approximation (Thm. 4). This two-phase strategy is demonstrated in panels **a**-**b**. Even after GD flattens out in terms of the objective value, RLQN remains a steady linear convergence. L-BFGS makes good progress in general, but its performance features instability: periods of stagnation and fluctuation, presumably due to the restarting of the Hessian approximation using limited memory.

In panels **c**-**d**, we numerically verify the improved iteration complexity of RLQN over GD in terms of the gradient norm (Thm. 1), which does not require low-rank local landscape assumption (A2). For this, we used large ℓ_2-regularization ($\lambda = 5$) to make the problem is relatively well-conditioned. This favors GD as we see a steady linear convergence. One can clearly see from panel **c** that RLQN shows a much-improved iteration complexity over GD. The improvement is still significant but a bit degraded if we measure the progress of the gradient norm w.r.t. flops. This is due to the relatively higher per-iteration computational cost of RLQN over GD.

5. Numerical experimental: Large linear inverse problems

Here we test our algorithm on solving linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{p\times N}$ and $\mathbf{b} \in \mathbb{R}^p$ with $p \gg N$. This can be reformulated as $\min_{\mathbf{x}\in\mathbb{R}^N}\left[f(\mathbf{x}) := \frac{1}{2}\|\mathbf{Ax} - \mathbf{b}\|^2\right]$. While this problem is convex with constant Hessian, it has a wide range of applications especially in scientific computing.

Also, since we can control the Hessian of the objective through the singular value profile of $\mathbf{A}$, it provides a nice class of problems to test our improved local linear convergence rate (Thm. 4), which operates under the assumption of approximate low-rank local landscape (A2). We present experimental results with both synthetic and real data. We compare our algorithm with two widely used linear problem solvers of randomized Kaczmarz [31] and conjugate gradient least squares (CGLS) [13]. We plot the least square error as a function of flops. Since the randomized Kacmarz and our RLQN are stochastic, we repeated the experiments 10 times and plotted the mean and shaded the region in between the running maximum and minimum.

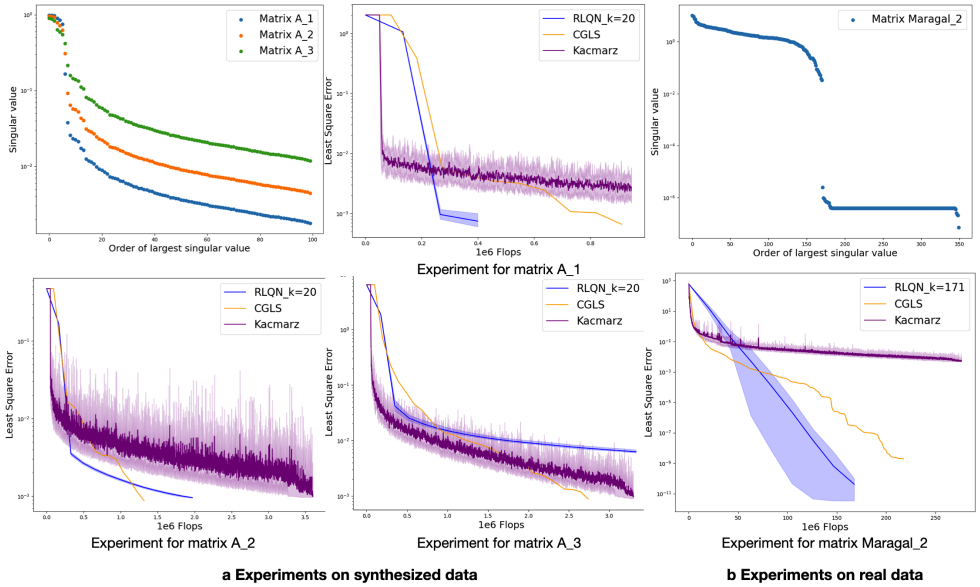

Figure 2: Plots of singular values distributions of four matrices and the result of solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ is plotted as least square error vs. flops used by the algorithms. Part **a**, the left hand side, are the experiments on synthesized data. Part **b**, the right hand side, is the experiment on matrix `Maragal2` in the SuiteSparse Matrix Collection [16].

In Figure 2 part **a**, we present three synthetic experiments with matrices $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \in \mathbb{R}^{500\times 100}$. We first sample $\mathbf{A} \in \mathbb{R}^{500\times 100}$, $\mathbf{b} \in \mathbb{R}^{500}$ from i.i.d. standard normal. Then we perform a singular value decomposition on matrix $\mathbf{A}$ and get $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. We perturb the singular values in $\mathbf{\Sigma}$ to create $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ with singular value profiles depicted in Figure 2 top left. Going from $\mathbf{A}_1$ to $\mathbf{A}_3$, the singular values decay slower and hence the parameter ρ in A2 becomes larger. And we use rank parameter $k = 20$ in this set of experiments.

In Figure 2 part **b** we also test our algorithm on matrix `Maragal2` in the SuiteSparse Matrix Collection [16]. This matrix has 555 rows, 350 columns, and has significant rank deficiency as seen in its singular value distribution in Figure 2 top right. We used $k = 171$ for this experiment. Although $k = 171$ is a relatively large rank, capturing that part of the curvature structure helps our algorithm to converge very fast. It only takes 5-10 iterations for our algorithm to reach 10^{-10} error. We observe a superior convergence rate with our RLQN algorithm in comparison to the other benchmark methods. This highlights that our RLQN algorithm leverages low-rank structure in the landscape of the problem and yields faster convergence than other methods that do not leverage such geometric information.

References

[1] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Jürgen Bräuninger. Eine quasi-newton-methode mit cholesky-faktorisierung. *Computing*, 25: 155–162, 1980.

[3] Charles G Broyden. The convergence of single-rank quasi-newton methods. *Mathematics of Computation*, 24(110):365–382, 1970.

[4] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[5] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii: worst-case function-and derivative-evaluation complexity. *Mathematical programming*, 130(2):295–319, 2011.

[6] Yifan Chen, Ethan N Epperly, Joel A Tropp, and Robert J Webber. Randomly pivoted cholesky: Practical approximation of a kernel matrix with few entry evaluations. *arXiv preprint arXiv:2207.06503*, 2022.

[7] Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

[8] Jin-yan Fan and Ya-xiang Yuan. On the quadratic convergence of the Levenberg-Marquardt method without nonsingularity assumption. *Computing*, 74:23–39, 2005.

[9] Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3): 317–322, 1970.

[10] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26, 1970.

[11] Geovani N Grapiglia, Jinyun Yuan, and Ya-xiang Yuan. On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization. *Mathematical Programming*, 152:491–520, 2015.

[12] Andreas Griewank. The modification of newton's method for unconstrained optimization by bounding cubic terms. Technical report, Technical report NA/12, 1981.

[13] Magnus Rudolph Hestenes, Eduard Stiefel, et al. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

[14] Qiujiang Jin, Ruichen Jiang, and Aryan Mokhtari. Non-asymptotic global convergence rates of bfgs with exact line search. *arXiv preprint arXiv:2404.01267*, 2024.

[15] Leonid V Kantorovich. On newton's method for functional equations. In *Dokl. Akad. Nauk SSSR*, volume 59, pages 1237–1240, 1948.

[16] Scott P Kolodziej, Mohsen Aznaveh, Matthew Bullock, Jarrett David, Timothy A Davis, Matthew Henderson, Yifan Hu, and Read Sandstrom. The suitesparse matrix collection website interface. *Journal of Open Source Software*, 4(35):1244, 2019.

[17] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[18] Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291, 2013.

[19] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.

[20] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[21] Isaac Newton. *Philosophiae naturalis principia mathematica*. Jussu Societatis Regiae ac Typis Josephi Streater. Prostat Venales apud Sam . . . , 2022.

[22] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[23] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[24] Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[25] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[26] Michael James David Powell. Convergence properties of a class of minimization algorithms. In *Nonlinear programming 2*, pages 1–27. Elsevier, 1975.

[27] Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent conditions for c^2 functions. *arXiv preprint arXiv:2303.00096*, 2023.

[28] Anton Rodomanov and Yurii Nesterov. New results on superlinear convergence of classical quasi-newton methods. *Journal of optimization theory and applications*, 188:744–769, 2021.

[29] Anton Rodomanov and Yurii Nesterov. Rates of superlinear convergence for classical quasi-newton methods. *Mathematical Programming*, pages 1–32, 2021.

[30] David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[31] Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[32] Max A Woodbury. *Inverting modified matrices*. Department of Statistics, Princeton University, 1950.

[33] Stephen J Wright. Numerical optimization, 2006.

[34] Fuzhen Zhang. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

A fast and efficient randomized quasi-Newton method

Supplementary Material

Appendix A. Details on the Algorithm

In this section, we give details for the randomized Hessian approximation algorithm that we use as part of Algorithm 1.

Algorithm 2 Random Pivoted Cholesky Factorization (RPC) [6]

Input: $\mathbf{A} \in \mathbb{R}^{N\times N}$; $k \leq N$ (number of columns to be sampled);
  Set $\mathbf{F} \leftarrow O \in \mathbb{R}^{N\times k}$, $\mathbf{d} \leftarrow$ diagonals and selected columns of $\mathbf{A}$;
  for $n = 1, \cdots, k$ **do**
    pick $s \in \{1, ..., N\}$ according to distribution $|\mathbf{d}|/\|\mathbf{d}\|_1$
    $\mathbf{col} \leftarrow \mathbf{A}[:, s]$
    $\mathbf{g} \leftarrow \mathbf{col} - \mathbf{F}[:, 1 : n-1]\mathbf{F}[s, 1 : n-1]^T$
    if $\mathbf{g}[s] \leq 0$
      output None (▷ *Algorithm failed!*)
    else
      $\mathbf{F}[:, n] \leftarrow \mathbf{g}/\sqrt{\mathbf{g}[s]}$
      $\mathbf{d} \leftarrow \mathbf{d} - \mathbf{F}[:, n] * \mathbf{F}[:, n]$
  $R \leftarrow \|\mathbf{d}\|_1$ (▷ *Trace norm approx. error*)
  $\lambda \leftarrow \lambda_{\max}(\mathbf{F}^T\mathbf{F})$ (▷ *Same as the largest eigenvalue of* $\mathbf{F}\mathbf{F}^T$*, cost* $O(k^3)$)
  output $\mathbf{F}$, R, λ

Remark 5 *We list the matrix* $\mathbf{A} \in \mathbb{R}^{N\times N}$ *for the ease of notation. In fact, this randomized factorization only requires the reading of* $(k+1)N - k$ *entries of* $\mathbf{A}$ *coming from* k *column of* $\mathbf{A}$ *and its diagonal. When the input* $\mathbf{A}$ *is PSD, the output* λ *is at most* $\|\mathbf{A}\|$ *(see Thm 5.3 in [34]). But when input* $\mathbf{A}$ *is not PSD, the factorization might fail to go through; or the low rank factorization might go through despite the fact that* $\mathbf{A}$ *is not PSD, which could result in* $\lambda > \|\mathbf{A}\|$*. Since our objective function* $f(\boldsymbol{\theta})$ *could be nonconvex, the input matrices* $\mathbf{A} = \nabla^2 f(\boldsymbol{\theta})$ *are not necessarily PSD. And we deal with this subtlety using the next algorithm.*

Algorithm 3 RPC of LM-regularized Hessian

Input: $\boldsymbol{\theta}$ (parameter state); $k \leq N$ (number of columns for RPC); L (smoothness parameter)
  $\mathbf{F}, R, \lambda \leftarrow$ output of Alg. 2 with input ($\mathbf{A} =$ diagonals and selected columns of $\nabla^2 f(\boldsymbol{\theta})$, k)
  if $(\mathbf{F}, R, \lambda)$ is None **or** $\lambda > L$: (▷ *Encounter* $\nabla^2 f(\boldsymbol{\theta})$ *that is not PSD*)
    $\mathbf{F}, R, \lambda \leftarrow$ output of Alg. 2 with input ($\mathbf{A} = \nabla^2 f(\boldsymbol{\theta}) + L\mathbf{I}$, k)
  output $\mathbf{F}$, R

Remark 6 *The key implementation detail is to 'try' RPC (Alg. 2) with no regularization; if RPC fails to go through or if RPC gives an output* $\mathbf{F}$ *with* $\lambda_{\max}(\mathbf{F}\mathbf{F}^T) > L$ *(line 3 of Alg.3), then we know the current Hessian is not PSD. In this case we follow up with another RPC factorization with the Hessian regularized by the smoothness parameter* L*, namely* $\nabla^2 f(\boldsymbol{\theta}) + L\mathbf{I}$*, where* $\boldsymbol{\theta}$ *is the current iterate. The second factorization will certainly go though because* $\nabla^2 f(\boldsymbol{\theta}) + L\mathbf{I}$ *is PSD. In addition, if we use* $\mathbf{F}$ *to denote the new output of the RPC factorization,* $\lambda_{\max}(\mathbf{F}\mathbf{F}^T) \leq \|\nabla^2 f(\boldsymbol{\theta}) + L\mathbf{I}\| \leq 2L$*. After the factorization, we choose the preconditioning matrix* $\mathbf{B}_n$ *(2) to be the obtained Cholesky decomposition plus additional LM-regularization* $\delta_n \mathbf{I}_N$*, where* δ_n *is chosen adaptively in line 5 of Alg.1. This choice can be easily computed, and guarantees the hypothesis in Thm. 1 as well as* $\lambda_{\max}(\mathbf{B}_n) \leq 3L$*.*

Remark 7 (Matrix-vector product) *In line 7 of Alg. 1, by using Woodbury identity [32], we can use the following trick to calculate the search direction*

$$(\mathbf{F_n}\mathbf{F_n}^T + \delta_n \mathbf{I}_N)^{-1} \nabla f(\boldsymbol{\theta}_n) = \delta_n^{-1} \nabla f(\boldsymbol{\theta}_n) - \delta_n^{-1} \mathbf{F} (\delta_n \mathbf{I}_k + \mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \nabla f(\boldsymbol{\theta}_n). \tag{5}$$

The dominating computation above are $N \times k$ *matrices multiplying vectors of dimension* k*,* $k \times N$ *matrices multiplying vectors of dimension* N*, and inverting a* $k \times k$ *matrix. Therefore the computational cost of each iteration is indeed* $O(k^2 N)$*.*

Appendix B. Discussion on assumptions

(A1) This is a standard assumption on smoothness of the objective.

(A2) Classical convergence results for (quasi-)Newton methods assume that the target local minimizer $\boldsymbol{\theta}_*$ has positive definite Hessian and the algorithm is initialized sufficiently close to $\boldsymbol{\theta}_*$. Our first local convergence analysis (Thm. 2) operates under a similar assumption but additionally exploits the 'approximately low-rank' structure of the Hessian.

(A3) Our local convergence analysis (Thm. 4) concerns the challenging case when the local minima are 'flat', meaning that the Hessian is positive semi-definite with some zero eigenvalues, and that they are accumulated. One simple objective function f that satisfies A3 is a quadratic that only depends on a few coordinates upon a linear transformation. Consider the low rank quadratic example, $f(\boldsymbol{\theta}) = 2^{-1}\boldsymbol{\theta}_T \mathbf{A} \boldsymbol{\theta}$, where matrix $\mathbf{A}$ is symmetric PSD and has rank k. And our assumption A3 generalizes such simple examples: we allow the range of the Hessian of the objective to be non-constant, and we only require a mild curvature condition (μ-PL) at the local minimum. We list some common assumptions to incorporate degenerate Hessian. Three commonly used such assumptions are Polyak-Lojasiewicz (PL), error bound (EB), and quadratic growth (QG). We briefly recall the three definitions below:

$$\begin{aligned} &\mu\text{-PL near } \boldsymbol{\theta}_* : f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*) \leq \mu \|\nabla f(\boldsymbol{\theta})\|^2; \\ &\mu\text{-EB near } \boldsymbol{\theta}_* : \text{dist}(\boldsymbol{\theta}, S(\boldsymbol{\theta}_*)) \leq \|\nabla f(\boldsymbol{\theta})\|; \\ &\mu\text{-QG near } \boldsymbol{\theta}_* : f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*) \geq \frac{\mu}{2} \text{dist}(\boldsymbol{\theta}, S(\boldsymbol{\theta}_*))^2. \end{aligned} \tag{6}$$

Rebjock and Boumal [27] showed that the three conditions above are 'essentially equivalent' if the objective f is C^2. Here, 'essentially' means that the constant and the neighborhood can degrade when one deduces one condition from another.

Appendix C. Preliminary lemmas

The following classical lemma is useful for deriving iteration complexity and global convergence guarantees.

Lemma 8 *Let $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ be sequences of non-negative real numbers such that $\sum_{n=0}^{\infty} a_n b_n < \infty$. Then the following hold.*

(i) $\min_{1\leq k\leq n} b_k \leq \dfrac{\sum_{k=0}^{\infty} a_k b_k}{\sum_{k=0}^{n} a_k} = O\left(\left(\sum_{k=0}^{n} a_k\right)^{-1}\right)$.

(ii) *Further assume $\sum_{n=0}^{\infty} a_n = \infty$ and $|b_{n+1} - b_n| = O(a_n)$. Then $\lim_{n\to\infty} b_n = 0$.*

Proof **(i)** follows from noting that

$$\left(\sum_{k=1}^{n} a_k\right) \min_{1\leq k\leq n} b_k \leq \sum_{k=1}^{n} a_k b_k \leq \sum_{k=1}^{\infty} a_k b_k < \infty. \tag{7}$$

The proof of **(ii)** is omitted and can be found in [18, Lem. A.5]. ■

If the function f exhibits lower rank Hessian near a local minimum, and the number of pivoting columns, k, used in the RPC process exceeds the actual rank, then the lower rank Hessian approximation is exact. This is helpful for establishing the superlinear convergence in Theorem 4.

Lemma 9 *For any symmetric $\mathbf{A} \succcurlyeq 0$, if the rank of $\mathbf{A}$ is at most $k \leq N$, then after k loops of RPC factorization the output $\mathbf{F}$ satisfies $\mathbf{F}\mathbf{F}^T = \mathbf{A}$.*

Proof This is a direct consequence of [6, Lem. 3.4]. ■

Theorem 10 (Error bound for RPC; Thm. 5.1 in [6]) *Let $\mathbf{A}$ be a positive semi-definite matrix. Fix $r \in \mathbb{N}$ and $\varepsilon > 0$. The column Nyström approximation $\hat{\mathbf{A}}^{(k)}$ produced by k steps of RPC (Alg. 2) attains the bound*

$$\mathbb{E}\left[\operatorname{tr}(\mathbf{A} - \hat{\mathbf{A}}^{(k)})\right] \leq (1+\varepsilon)\cdot \operatorname{tr}(\mathbf{A} - [\![\mathbf{A}]\!]_r),$$

provided that the number k of sampled columns satisfies

$$k \geq \frac{r}{\varepsilon} + \min\left\{r\log\left(\frac{1}{\varepsilon\eta}\right), r + r\log_+\left(\frac{2^r}{\varepsilon}\right)\right\}.$$

The relative error η is defined by $\eta := \frac{\operatorname{tr}(\mathbf{A}-[\![\mathbf{A}]\!]_r)}{\operatorname{tr}(\mathbf{A})}$. $\log_+(x) := \max\{\log x, 0\}$ for $x > 0$, and the logarithm has base e.

Appendix D. Global convergence analysis for general preconditioned GD

First let's show that the smallest eigenvalue of $\mathbf{B}_n$ is lower bounded by a linear factor of the parameter difference $r_n := \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|$.

Lemma 11 *For any $n \geq 0$, let $\boldsymbol{\theta}_n$ be the current iterate and $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)$ be the next iterate found by (1), with $\mathbf{B}_n \succcurlyeq \sqrt{CL_H\|\nabla f(\boldsymbol{\theta}_n)\|}$ for some $C > 1/3$. Then almost surely*

$$\lambda_{\min}(\mathbf{B}_n) \geq CL_H r_n. \tag{8}$$

Proof By the requirements on $\mathbf{B}_n$, $\lambda_{\min}(\mathbf{B}_n) \geq \sqrt{CL_H\|\nabla f(\boldsymbol{\theta}_n)\|}$. Thus

$$\begin{aligned} r_n = \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| &= \|\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\| \\ &\leq \frac{1}{\lambda_{\min}(\mathbf{B}_n)}\|\nabla f(\boldsymbol{\theta}_n)\| \\ &\leq \frac{1}{\lambda_{\min}(\mathbf{B}_n)}\frac{\lambda_{\min}(\mathbf{B}_n)^2}{CL_H}. \end{aligned}$$

Rearrange we get $\lambda_{\min}(\mathbf{B}_n) \geq CL_H r_n$. ■

To the end of showing the global convergence, below we prove a lower bound for per-iteration improvement in objective value.

Lemma 12 *Suppose A1 holds. For any $n \geq 0$, let $\boldsymbol{\theta}_n$ be the current iterate and $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)$ be the next iterate generated by (1), with $\lambda_{\min}(\mathbf{B}_n) \geq \min\{L, \sqrt{CL_H\|\nabla f(\boldsymbol{\theta}_n)\|}\}$ and $\mathbf{B}_n \succcurlyeq \nabla^2 f(\boldsymbol{\theta}_n)$ for some $C > 1/3$. Then almost surely*

$$f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) \geq A\left\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle, \tag{9}$$

where $A = \frac{1}{2} - \frac{1}{6C} > 0$

Proof First consider the easier case where $\min\{L, \sqrt{CL_H\|\nabla f(\boldsymbol{\theta}_n)\|}\} = L$. Then we can use the first order Taylor expansion and the update rule $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)$ to lower bound the per-iteration progress by

$$\begin{aligned} & f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) \\ & \geq -\left\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle - \frac{L}{2}r_n^2 \\ & = \left\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle - \frac{L}{2}r_n^2 \\ & = \frac{1}{2}\left\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle + \frac{1}{2}\left\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle - \frac{L}{2}r_n^2 \\ & \geq \frac{1}{2}\left\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle, \end{aligned} \tag{10}$$

where the last line follows from the assumption that $\lambda_{\min}(\mathbf{B}_n) \geq L$. Next consider the other case where $\min\{L, \sqrt{CL_H\|\nabla f(\boldsymbol{\theta}_n)\|}\} = \sqrt{CL_H\|\nabla f(\boldsymbol{\theta}_n)\|}$. Then by a second order Taylor's expansion (see [19] for a detailed derivation), we have

$$\begin{aligned} & f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) \\ & \geq -\left\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle - \frac{1}{2}\left\langle \nabla^2 f(\boldsymbol{\theta}_n)(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\right\rangle - \frac{L_H}{6}r_n^3. \end{aligned}$$

By the update rule $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)$, we can write $-\nabla f(\boldsymbol{\theta}_n) = \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n)$, thus

$$\begin{aligned}
&f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) \\
&= \langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle - \frac{1}{2}\left\langle \nabla^2 f(\boldsymbol{\theta}_n)(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \right\rangle - \frac{L_H}{6} r_n^3 \\
&= \frac{1}{2}\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle + \frac{1}{2}\left\langle (\mathbf{B}_n - \nabla^2 f(\boldsymbol{\theta}_n))(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \right\rangle \\
&\qquad - \frac{L_H}{6} r_n^3 \\
&\geq \frac{1}{2}\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle - \frac{L_H}{6} r_n^3,
\end{aligned}$$

where the last line follows from the assumption that $\mathbf{B}_n \succcurlyeq \nabla^2 f(\boldsymbol{\theta}_n)$. Continuing the computation we have

$$\begin{aligned}
&f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) \\
&\geq \frac{1}{2}\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle - \frac{1}{6} r_n^3 \\
&= (\frac{1}{2} - \frac{1}{6C})\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle + \frac{1}{6C}\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle - \frac{1}{6} r_n^3 \\
&\geq (\frac{1}{2} - \frac{1}{6C})\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle, \qquad (11)
\end{aligned}$$

where the last line is a consequence of (8). Indeed, by (8), $\lambda_{\min}(\mathbf{B}_n) \geq C L_H r_n$, thus

$$\frac{1}{6C}\langle \mathbf{B}_n(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle \geq \frac{1}{6C}\lambda_{\min}(\mathbf{B}_n) r_n^2 \geq \frac{L_H}{6} r_n^3.$$

Since $C > \frac{1}{3}$, $0 < \frac{1}{2} - \frac{1}{6C} < \frac{1}{2}$. Thus, combining the results of two cases (10) and (11), we can conclude (9). ■

Proof [of Thm. 1] Denote $A = \frac{1}{2} - \frac{L_H}{6C} > 0$, by (9)

$$\begin{aligned}
f(\boldsymbol{\theta}_k) - f(\boldsymbol{\theta}_{k+1}) &\geq A\langle \mathbf{B}_n(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle \\
&= A\left\langle \nabla f(\boldsymbol{\theta}_k), \mathbf{B}_k^{-1}\nabla f(\boldsymbol{\theta}_k) \right\rangle \\
&\geq \frac{A\|\nabla f(\boldsymbol{\theta}_k)\|^2}{\|\mathbf{B}_k\|}.
\end{aligned}$$

Sum them up from $k = 0$ to $n - 1$ we have

$$\sum_{k=0}^{n-1} A\|\nabla f(\boldsymbol{\theta}_k)\|^2 \|\mathbf{B}_k\|^{-1} \leq f(\boldsymbol{\theta}_0) - f(\boldsymbol{\theta}_n) \leq f(\boldsymbol{\theta}_0) - \inf f. \qquad (12)$$

Thus

$$\min_{0 \leq k \leq n-1} \|\nabla f(\boldsymbol{\theta}_k)\|^2 \sum_{k=0}^{n-1} \|\mathbf{B}_k\|^{-1} \leq \frac{f(\boldsymbol{\theta}_0) - \inf f}{A}.$$

That is

$$\min_{0\le k\le n-1} \|\nabla f(\boldsymbol{\theta}_k)\| \le \left(\frac{(f(\boldsymbol{\theta}_0)-\inf f)}{A\sum_{k=0}^{n-1}\|\mathbf{B}_k\|^{-1}}\right)^{1/2}. \tag{13}$$

To show that asymptotically the gradient norm converges to zero, we first take n to infinity in (12) and get

$$\sum_{k=0}^{\infty} A\|\nabla f(\boldsymbol{\theta}_k)\|^2\|\mathbf{B}_k\|^{-1} \le f(\boldsymbol{\theta}_0)-\inf f<\infty.$$

Let $a_k = A\|\mathbf{B}_k\|^{-1}$, and $b_k = \|\nabla f(\boldsymbol{\theta}_k)\|^2$. Since $A>0$ and $\|\mathbf{B}_k\|$ is uniformly bounded from above, there is $c>0$ so that $a_k = A\|\mathbf{B}_k\|^{-1}\ge c$ for all k. Thus, $\sum_{k=0}^{\infty} a_k=\infty$. Then by the second part of the Lemma 8, the asymptotic convergence of $\|\nabla f(\boldsymbol{\theta}_n)\|$ to zero would follow from $\|\nabla f(\boldsymbol{\theta}_n)\| = O(1)$. Indeed, as a result of (9), we have $\boldsymbol{\theta}_0\ge\boldsymbol{\theta}_1\ge\ldots\ge\boldsymbol{\theta}_n$, thus the sequence of iterates generated by our algorithm $(\boldsymbol{\theta}_n)_{n=0}^{\infty}\subset\{\boldsymbol{\theta}: f(\boldsymbol{\theta})\le f(\boldsymbol{\theta}_0)\}$. Since f is bounded from below, the sublevel set $\{\boldsymbol{\theta}: f(\boldsymbol{\theta})\le f(\boldsymbol{\theta}_0)\}$ is compact, then $\|\nabla f(\boldsymbol{\theta}_n)\|$ is uniformly bounded for all $n\ge 0$. ■

Remark 13 *In standard GD, the preconditioning matrix $\mathbf{B}_k=\frac{1}{L}\mathbf{I}$ for all $k\ge 0$, in which case we have $\frac{1}{\sum_{k=0}^{n-1}\|\mathbf{B}_k\|^{-1}}=\frac{L}{n}$. So applying our general result (13) to standard GD gives the rate*

$$\min_{0\le k\le n-1}\|\nabla f(\boldsymbol{\theta}_k)\| = O\left(\sqrt{\frac{L(f(\boldsymbol{\theta}_0)-\inf f)}{n}}\right),$$

which recovers exactly the rate of global convergence of GD for non-convex objective function. In particular, in the course of our algorithm RLQN, $\|\mathbf{B}_n\|\le 3L$ from the construction of Alg. 1 and Alg. 3. Thus the global convergence rate RLQN for a nonconvex function matches that of GD. And heuristically $\|\mathbf{B}_k\|\approx\|\nabla^2 f(\boldsymbol{\theta}_k)\|$ should be in general much smaller than L, the uniform upper-bound of the spectrum norm of the Hessian. So it is likely that in general $\frac{1}{\sum_{k=0}^{n-1}\|\mathbf{B}_k\|^{-1}}$ would be significantly smaller than $\frac{L}{n}$.

Appendix E. Proof of local linear convergence with improved rate

In this section, we prove Theorems 2, 4, and Corollary 3. Throughout this section, we use the following notations:

- $\mathbf{A}_n := \mathbf{F}_n\mathbf{F}_n^T$: random Cholesky factorization of the LM-regularized Hessian $\nabla^2 f(\boldsymbol{\theta}_n)+\tau_n\mathbf{I}_N$;
- $\mathbf{E}_n := \nabla^2 f(\boldsymbol{\theta}_n)+\tau_n\mathbf{I}-\mathbf{A}_n$: the error matrix between the input and output of the RPC factorization;
- $\mathbf{B}_n := \mathbf{A}_n+\delta_n\mathbf{I}_N$: the the final preconditioning matrix of the gradient at step n.

Lemma 14 (Local contraction lemma) *Suppose A1 holds. and $\boldsymbol{\theta}_n$ is sufficiently close to a local minimizer $\boldsymbol{\theta}_*$ of f, then almost surely,*

$$\|\boldsymbol{\theta}_{n+1}-\boldsymbol{\theta}_*\| \le \left(1-\frac{\mu-L_H\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_*\|}{\delta_n+\mu}\right)\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_*\|. \tag{14}$$

Above lemma is the key ingredient for local convergence. It states that when $\boldsymbol{\theta}_n$ is sufficiently close to $\boldsymbol{\theta}_*$, the parameter estimation error contracts at rate $\sim 1 - \frac{\mu}{\delta_n + \mu}$, which does not depend on the smoothness parameter L.

Proof Since $\boldsymbol{\theta}_n$ is sufficiently close to $\boldsymbol{\theta}_*$, $\nabla^2 f(\boldsymbol{\theta}_n)$ is PSD. We can then write $\nabla^2 f(\boldsymbol{\theta}_n) = \mathbf{A}_n + \mathbf{E}_n$. Since $\mathbf{B}_n = \mathbf{A}_n + \delta_n \mathbf{I}$, we can write $\mathbf{B}_n = \nabla^2 f(\boldsymbol{\theta}_n) - \mathbf{E}_n + \delta_n \mathbf{I}$. Then we have the following update:

$$
\begin{aligned}
\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_* &= \boldsymbol{\theta}_n - \boldsymbol{\theta}_* - \mathbf{B}_n^{-1} \nabla f(\boldsymbol{\theta}_n) \\
&= \mathbf{B}_n^{-1} \mathbf{B}_n (\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) - \mathbf{B}_n^{-1} \int_0^1 \nabla^2 f(\boldsymbol{\theta}_* + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*))\, dt (\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \\
&= \mathbf{B}_n^{-1} \left(\nabla^2 f(\boldsymbol{\theta}_n) - \mathbf{E}_n + \delta_n \mathbf{I} \right) (\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \\
&\qquad - \mathbf{B}_n^{-1} \int_0^1 \nabla^2 f(\boldsymbol{\theta}_* + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*))\, dt (\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \\
&= \mathbf{B}_n^{-1} (\delta_n \mathbf{I} - \mathbf{E}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \\
&\qquad - \mathbf{B}_n^{-1} \int_0^1 \nabla^2 f(\boldsymbol{\theta}_n) - \nabla^2 f(\boldsymbol{\theta}_* + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*))\, dt (\boldsymbol{\theta}_n - \boldsymbol{\theta}_*).
\end{aligned}
\tag{15}
$$

For the quadratic term we have the 2-norm estimate by the L_H-Lipschitz continuity of the Hessian,

$$
\begin{aligned}
&\left\| \mathbf{B}_n^{-1} \int_0^1 \nabla^2 f(\boldsymbol{\theta}_n) - \nabla^2 f(\boldsymbol{\theta}_* + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)) dt (\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \right\| \\
&\leq \left\| \mathbf{B}_n^{-1} \int_0^1 \nabla^2 f(\boldsymbol{\theta}_n) - \nabla^2 f(\boldsymbol{\theta}_* + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*))\, dt \right\|_{op} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \\
&\leq \frac{L_H}{2\delta_n} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|^2 \\
&\leq \frac{L_H}{\mu + \delta_n} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|^2,
\end{aligned}
$$

where the last line follows from the $\mu \leq \|\mathbf{E}_n\|_{op}$, to see this pick any unit vector $\mathbf{v}$ in the orthogonal complement of span$(\mathbf{A}_n)$, then

$$
\mu \leq \left\langle \mathbf{v}, \nabla^2 f(\boldsymbol{\theta}_n) \mathbf{v} \right\rangle = \langle \mathbf{v}, \mathbf{A}_n \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{E}_n \mathbf{v} \rangle = 0 + \langle \mathbf{v}, \mathbf{E}_n \mathbf{v} \rangle \leq \|\mathbf{E}_n\|_{op}.
$$

For the linear part we have

$$
\begin{aligned}
&\|\mathbf{B}_n^{-1} (\delta_n \mathbf{I} - \mathbf{E}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)\| \\
&= \|\left(\nabla^2 f(\boldsymbol{\theta}_n) + \delta_n \mathbf{I} - \mathbf{E}_n \right)^{-1} (\delta_n \mathbf{I} - \mathbf{E}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)\| \\
&\leq \|\left(\nabla^2 f(\boldsymbol{\theta}_n) + \delta_n \mathbf{I} - \mathbf{E}_n \right)^{-1} (\mu \mathbf{I} + \delta_n \mathbf{I} - \mathbf{E}_n)\|_{op} \cdot \|\left(\mu \mathbf{I} + \delta_n \mathbf{I} - \mathbf{E}_n \right)^{-1} (\delta_n \mathbf{I} - \mathbf{E}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)\|.
\end{aligned}
$$

Notice that $\|\left(\nabla^2 f(\boldsymbol{\theta}_n) + \delta_n \mathbf{I} - \mathbf{E}_n \right)^{-1} (\mu \mathbf{I} + \delta_n \mathbf{I} - \mathbf{E}_n)\|_{op} \leq 1$. To see this, let $\mathbf{X}$ be positive definite and $\mathbf{Y}$ be positive semidefinite, then

$$
\begin{aligned}
\|(\mathbf{X} + \mathbf{Y})^{-1} \mathbf{X}\|_{op} &= \|\left(\mathbf{X}(\mathbf{I} + \mathbf{X}^{-1} \mathbf{Y}) \right)^{-1} \mathbf{X}\|_{op} \\
&= \|\left(\mathbf{I} + \mathbf{X}^{-1} \mathbf{Y} \right)^{-1} \mathbf{X}^{-1} \mathbf{X}\|_{op} \\
&= \|\left(\mathbf{I} + \mathbf{X}^{-1} \mathbf{Y} \right)^{-1}\|_{op} \\
&= \|\left(\mathbf{I} + \mathbf{X}^{-1} \mathbf{Y} \right)\|_{op}^{-1} \leq 1,
\end{aligned}
$$

where the last inequality follows from the fact that $\mathbf{X}^{-1}\mathbf{Y} \succcurlyeq 0$. Then by taking $\mathbf{X} = \mu\mathbf{I} + \delta_n\mathbf{I} - \mathbf{E}_n$ and $\mathbf{Y} = \nabla^2 f(\boldsymbol{\theta}_n) - \mu\mathbf{I}$, we arrive at the claim

$$\|\left(\nabla^2 f(\boldsymbol{\theta}_n) + \delta_n\mathbf{I} - \mathbf{E}_n\right)^{-1}(\mu\mathbf{I} + \delta_n\mathbf{I} - \mathbf{E}_n)\|_{op} \leq 1.$$

Now let $\mathbf{v}$ be any unit eigenvector of $\mathbf{E}_n$ with eigenvalue $\nu \geq 0$, then $\mathbf{v}$ is also an eigenvector of $\delta_n\mathbf{I} - \mathbf{E}_n$ and $((\mu+\delta_n)\mathbf{I} - \mathbf{E}_n)^{-1}$ with eigenvalues $\delta_n - \nu$ and $(\mu + \delta_n - \nu)^{-1}$. Therefore

$$\|\left((\mu+\delta_n)\mathbf{I} - \mathbf{E}_n\right)^{-1}(\delta_n\mathbf{I} - \mathbf{E}_n)\mathbf{v}\| = \frac{\delta_n - \nu}{\mu + \delta_n - \nu} \leq \frac{\delta_n}{\mu + \delta_n},$$

where the last inequality follows from $\delta_n \geq \mathrm{tr}(\mathbf{E}_n) \geq \|\mathbf{E}_n\|_{op} \geq \nu$ (recall that $\mathbf{E}_n$ is PSD). Combining the results above, we have

$$\begin{aligned}
&\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_*\| \\
&\leq \|\mathbf{B}_n^{-1}(\delta_n\mathbf{I} - \mathbf{E}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)\| + \left\|\mathbf{B}_n^{-1}\int_0^1 \nabla^2 f(\boldsymbol{\theta}_n) - \nabla^2 f(\boldsymbol{\theta}_* + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*))\, dt\right\|_{op}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \\
&\leq \frac{\delta_n}{\delta_n + \mu}\|(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)\| + \frac{L_H}{\mu + \delta_n}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|^2 \\
&= \left(1 - \frac{\mu - L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|}{\delta_n + \mu}\right)\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|.
\end{aligned}$$

This shows the assertion. ■

Now we show Theorem 2.

Proof [Proof of Theorem 2] For each $n \geq 0$, let $\mathcal{F}_n$ denote the σ-algebra generated by all randomness up to step n of the algorithm. Let U denote the open ball centered at $\boldsymbol{\theta}_*$ in A2. Suppose that $\boldsymbol{\theta}_0 \in U$ and $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\| \leq \mu/(2L_H)$. Then by Lemma 14, it holds that $\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_*\| \leq \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|$ almost surely. Thus, provided that $\boldsymbol{\theta}_0$ is sufficiently close to $\boldsymbol{\theta}_*$, we have $\nabla^2 f(\boldsymbol{\theta}_n)$ is PSD for all $n \geq 0$, and thus $\mathbf{E}_n = \nabla^2 f(\boldsymbol{\theta}_n) - \mathbf{A}_n$ for all n. Since all $\boldsymbol{\theta}_n$'s are sufficiently close to $\boldsymbol{\theta}_*$, the regularization $\delta_n = \mathrm{tr}(\mathbf{E}_n)$ from the constructions of Alg.1 and Alg.2.

By Theorem 10, for our choice of k, at any $n \geq 0$,

$$\mathbb{E}\left[\mathrm{tr}(\nabla^2 f(\boldsymbol{\theta}_n) - \mathbf{A}_n)\,|\,\mathcal{F}_{n-1}\right] \leq (1+\varepsilon_0)\left(\mathrm{tr}\left(\nabla^2 f(\boldsymbol{\theta}_n) - [\![\nabla^2 f(\boldsymbol{\theta}_n)]\!]_r\right)\right),$$

where the expectation in taken over the random choice of pivots in RPC algorithm at time step n. For any $t > 0$, we say step n is a 't-good step' if

$$\mathrm{tr}(\nabla^2 f(\boldsymbol{\theta}_n) - \mathbf{A}_n) \leq (1+t)(1+\varepsilon_0)\left(\mathrm{tr}\left(\nabla^2 f(\boldsymbol{\theta}_n) - [\![\nabla^2 f(\boldsymbol{\theta}_n)]\!]_r\right)\right).$$

And by Markov's inequality,

$$\mathbb{P}\left\{\mathrm{tr}(\nabla^2 f(\boldsymbol{\theta}_n) - \mathbf{A}_n) \geq (1+\varepsilon_0)(1+t)\left(\mathrm{tr}\left(\nabla^2 f(\boldsymbol{\theta}_n) - [\![\nabla^2 f(\boldsymbol{\theta}_n)]\!]_r\right)\right)\,|\,\mathcal{F}_{n-1}\right\} \leq \frac{1}{1+t}.$$

Therefore at step n, conditional on $\mathcal{F}_{n-1}$, with probability at least $\frac{t}{1+t}$ the step n is a t-good step, and on this event, we have

$$\begin{aligned}
\delta_n &= \mathrm{tr}(\nabla^2 f(\boldsymbol{\theta}_n) - \mathbf{A}_n) \\
&\leq (1+\varepsilon_0)(1+t)\,\mathrm{tr}\left(\nabla^2 f(\boldsymbol{\theta}_n) - [\![\nabla^2 f(\boldsymbol{\theta}_n)]\!]_r\right) \\
&\leq (1+\varepsilon_0)(1+t)\rho L,
\end{aligned}$$

where the first inequality uses the hypothesis A2 that $\nabla^2 f(\boldsymbol{\theta})$ is PSD near $\boldsymbol{\theta}_*$ and the last inequality uses A2 and that $\boldsymbol{\theta}_n \in U$. Then with probability at least $\frac{t}{1+t}$, noting that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \le \mu/(2L_H)$, by Lemma 14,

$$\begin{aligned}\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_*\| &\le \left(1 - \frac{\mu - L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|}{\delta_n + \mu}\right)\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \\ &\le \left(1 - \frac{\mu/2}{\mu + (1+\varepsilon_0)(1+t)\rho L}\right)\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|.\end{aligned} \tag{16}$$

Taking conditional expectation with respect to the filtration $\mathcal{F}_n$ consisting of all randomness up to step n, we get

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_*\| \,|\, \mathcal{F}_n] &\le \frac{t}{1+t}\left(1 - \frac{\mu/2}{\mu + (1+\varepsilon_0)(1+t)\rho L}\right)\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| + \frac{1}{1+t}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \\ &= \left(\frac{t}{1+t}\frac{\mu/2 + (1+\varepsilon_0)(1+t)\rho L}{\mu + (1+\varepsilon_0)(1+t)\rho L} + \frac{1}{1+t}\right)\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|.\end{aligned}$$

By taking the total expectation and recursively using the resulting inequality, we get

$$\mathbb{E}[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|] \le \left(\frac{t}{1+t}\frac{\mu/2 + (1+\varepsilon_0)(1+t)\rho L}{\mu + (1+\varepsilon_0)(1+t)\rho L} + \frac{1}{1+t}\right)^n \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|.$$

We can optimize the contraction factor by minimizing over $t > 0$. Temporarily we denote $u = \mu/2$, and $c = (1+\varepsilon_0)\rho L$, then we wish to find

$$\min_{t>0} \frac{ct^2 + (2c+u)t + 2u + c}{ct^2 + (2c+2u)t + 2u + c}.$$

Denote $g(t) = ct^2 + (2c+u)t + 2u + c$, then we wish to find

$$\min_{t>0} \frac{g(t)}{g(t) + ut}. \tag{17}$$

Take derivative we get

$$\frac{g'(t)(g(t)+ut) - g(t)(g'(t)+u)}{(g(t)+ut)^2} = \frac{g'(t)ut - u}{(g(t)+ut)^2} = \frac{cut^2 - (2u+c)u}{(g(t)+ut)^2}.$$

So the contraction factor achieves it minimum in $t > 0$ at

$$t = \sqrt{\frac{2u+c}{c}} \approx 1,$$

because $c = (1+\varepsilon_0)\rho L \ge \mathrm{tr}(\mathbf{E}_n) \gg \mu = 2u$.

Take $t = 1$ in (17) we have

$$\begin{aligned}\min_{t>0} \frac{g(t)}{g(t)+ut} &\le \frac{3u+4c}{4u+4c} \\ &= 1 - \frac{2u}{8u+8c} \\ &= 1 - \frac{\mu}{4\mu + 8(1+\varepsilon_0)\rho L}.\end{aligned}$$

So we have in expectation

$$\mathbb{E}[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|] \leq \left(1 - \frac{\mu}{4\mu + 8(1+\varepsilon_0)\rho L}\right)^n \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|.$$

If we let M be the smallest integer that

$$\left(1 - \frac{\mu}{4\mu + 8(1+\varepsilon_0)\rho L}\right)^M \leq \frac{1}{2},$$

then we have for all $\ell \geq 1$

$$\mathbb{E}\|\boldsymbol{\theta}_{\ell M} - \boldsymbol{\theta}_*\| \leq \frac{1}{2^\ell}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|.$$

And by Markov's inequality we have

$$\mathbb{P}\left(\|\boldsymbol{\theta}_{\ell M} - \boldsymbol{\theta}_*\| \geq \frac{1}{2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|\right) \leq \frac{1}{2^{\ell-1}},$$

which is summable, which by Borel-Cantelli tells us that with probability 1, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \leq \frac{1}{2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$ for some almost surely finite n. Recursively applying this, we deduce that $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_*$ almost surely. ■

Proof [**Proof of Corollary 3**] In (16) take $t = 1$, we have with probability at least $\frac{1}{2}$, at step $i \geq 1$

$$\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_*\| \leq \left(1 - \frac{1}{2}\frac{\mu}{\mu + 2(1+\varepsilon_0)\rho L}\right)\|\boldsymbol{\theta}_{i-1} - \boldsymbol{\theta}_*\|. \tag{18}$$

Let us call such a step good step. Let X_i denote the random variable with $X_i = 0$ if step i is good, and $X_i = 1$ if the step is bad. Then $S_n = X_1 + ... + X_n$ counts the number of bad steps upto iteration n. Denote

$$p_i = \mathbb{E}[X_i|\mathcal{F}_{i-1}] \leq \frac{1}{2} \text{ almost surely,}$$

the expected value of X_i provided all the information of previous choices of pivots in random factorization. Let $M_n = S_n - p_1 - ... - p_n$ be the centered random variable of S_n. We claim M_n is a martingale. Indeed

$$\begin{aligned}\mathbb{E}[M_n|\mathcal{F}_{n-1}] &= \mathbb{E}[M_{n-1} + X_n - p_n|\mathcal{F}_{n-1}] \\ &= M_{n-1} - p_n + \mathbb{E}[X_n|\mathcal{F}_{n-1}] \\ &= M_{n-1} - p_n + p_n = M_{n-1},\end{aligned}$$

where the second line follows from the fact that M_{n-1} and p_n are measurable with respect to $\mathcal{F}_{n-1}$, and the last line follows from the definition of p_n. Also notice that M_n has bounded increment

$$|M_n - M_{n-1}| = |X_n - p_n| \leq 1.$$

So by Azuma-Hoeffding Inequality for any $\alpha > 0$,

$$\mathbb{P}\left(M_n \geq \alpha\right) \leq \exp\left(-\frac{\alpha^2}{2n}\right)$$

Since $p_i \leq \frac{1}{2}$ for all i, in particular take $\alpha = n/10$ we have

$$\begin{aligned}\mathbb{P}\left(S_n \geq \frac{n}{10} + \frac{n}{2}\right) &\leq \mathbb{P}\left(S_n \geq \frac{n}{10} + p_1 + ... + p_n\right) \\ &= \mathbb{P}\left(M_n \geq \frac{n}{10}\right) \\ &\leq \exp\left(-\frac{n}{200}\right)\end{aligned}$$

So with probability at most $\exp(-n/200)$, the number of bad steps upto iteration n is at most $\frac{3n}{5}$. Or equivalently, with probability at least $1 - \exp(-n/200)$, the number of good steps is more than $\frac{2n}{5}$. Fix a constant $0 < \beta < 1$, for any integer $K > 1$ let M_K denote the number of good steps needed to shrink the parameter distance from $\beta^{K-1}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$ to $\beta^K\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$. Setting

$$\left(1 - \frac{1}{2}\frac{\mu}{\mu + 2(1+\varepsilon_0)\rho L}\right)^{M_1} = \beta,$$

we get

$$M_1 \leq 2|\log\beta|\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu} + 1.$$

After that the shrinkage rate of each good step becomes

$$1 - \frac{\mu - \beta\mu/2}{\mu + 2(1+\varepsilon_0)\rho L} = 1 - \left(1 - \frac{\beta}{2}\right)\frac{\mu}{\mu + 2(1+\varepsilon_0)\rho L}.$$

Then we have

$$M_2 \leq \left(1 - \frac{\beta}{2}\right)^{-1}|\log\beta|\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu} + 1.$$

And in general we have

$$M_K \leq \left(1 - \frac{\beta^{K-1}}{2}\right)^{-1}|\log\beta|\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu} + 1.$$

Therefore to shrink the parameter distance to $\beta^K\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\| \leq \varepsilon\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$, the number of good steps needed is bounded by

$$\begin{aligned}\sum_{k=1}^{K} M_K &= K + |\log\beta|\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu}\sum_{k=1}^{K}\left(1 - \frac{\beta^{K-1}}{2}\right)^{-1} \\ &\leq K + |\log\beta|\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu}\sum_{k=1}^{K}\left(1 + \sum_{m=1}^{\infty}\left(\frac{\beta^{K-1}}{2}\right)^m\right) \\ &\leq K + |\log\beta|\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu}\sum_{k=1}^{K}\left(1 + \beta^{K-1}\right) \\ &= K + |\log\beta|\left(K + \frac{1-\beta^K}{1-\beta}\right)\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu} \\ &= K + \left(|\log\beta^K| + |\log\beta|\frac{1-\beta^K}{1-\beta}\right)\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu} \\ &\leq \frac{|\log\varepsilon|}{|\log\beta|} + \left(|\log\varepsilon| + \frac{|\log\beta|}{1-\beta}\right)\frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu} =: M_\varepsilon.\end{aligned}$$

Provided ε is small, there is an appropriate $\beta \in (0,1)$ so that

$$M_\varepsilon \leq 2(\log \varepsilon^{-1}) \frac{\mu + 2(1+\varepsilon_0)\rho L}{\mu}.$$

Therefore with probability at least

$$1 - \exp\left(- \frac{(\log \varepsilon^{-1})\,(\mu + 2(1+\varepsilon_0)\rho L)}{40\mu} \right),$$

we have for $n = \frac{5}{2} M_\varepsilon = 5(\log \varepsilon^{-1})\,(\mu + 2(1+\varepsilon_0)\rho L)\,/\mu$

$$\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\| \leq \varepsilon \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|.$$

■

Appendix F. Proof of local superlinear convergence

In this section, we prove Theorem 4. Throughout this section, we assume A1 holds and $\boldsymbol{\theta}_*$ will denote a local minimizer of f satisfying A3 and let $S := S(\boldsymbol{\theta}_*)$ as in A3. U will denote the open ball centered at $\boldsymbol{\theta}_*$ in A3. Unless otherwise mentioned, we will assume $\boldsymbol{\theta}_0 \in U$ and is sufficiently close to $\boldsymbol{\theta}_*$.

Lemma 15 *Assume f and a minimum $\boldsymbol{\theta}_*$ satisfy A3, then for any $\boldsymbol{\theta}_n \in U$ sufficiently close to $\boldsymbol{\theta}_*$, there is $\boldsymbol{\theta}_{n,*} \in S \cap U$ so that*

i *$\boldsymbol{\theta}_{n,*}$ is the most efficient reference local minimum to $\boldsymbol{\theta}_n$ in the sense that*

$$\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*} = \mathbf{P}_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}), \tag{19}$$

where the operator $\mathbf{P}_n$ is defined by

$$\mathbf{P}_n := \text{the projection operator onto the range of } \nabla^2 f(\boldsymbol{\theta}_{n,*}), \tag{20}$$

(See Figure 3 for illustration.)

ii *For any $\mathbf{v} \in$ Range$(\nabla^2 f(\boldsymbol{\theta}_{n,*}))$, $\langle \mathbf{v}, \nabla^2 f(\boldsymbol{\theta}_{n,*})\mathbf{v} \rangle \geq \mu \|\mathbf{v}\|^2$, where μ is the PL constant in A3.*

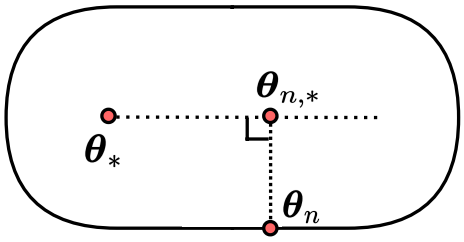

Figure 3: Example of flat local minimum $\boldsymbol{\theta}_*$. Contour represents the level curve of the objective.

Proof By Lemma 1.4, 1.5 and Corollary 2.7 of [27], our hypothesis A3 implies that

(a) The set $\text{proj}_{S\cap U}(\boldsymbol{\theta}_n) := \arg\min_{\boldsymbol{\theta}'_*\in S\cap U} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}'_*\|$ is not empty.

(b) For any $\boldsymbol{\theta}'_* \in \text{proj}_{S\cap U}(\boldsymbol{\theta}_n)$, $\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*} \in (T_{\boldsymbol{\theta}_{n,*}}S)^\perp$, where $T_{\boldsymbol{\theta}_{n,*}}S$ is the tangent space of S at $\boldsymbol{\theta}_{n,*}$.

(c) $\text{Ker}(\nabla^2 f(\boldsymbol{\theta}_{n,*})) = T_{\boldsymbol{\theta}_{n,*}}S$, and for any $\mathbf{v} \in \text{Range}(\nabla^2 f(\boldsymbol{\theta}_{n,*}))$, $\langle \mathbf{v}, \nabla^2 f(\boldsymbol{\theta}_{n,*})\mathbf{v}\rangle \geq \mu\|\mathbf{v}\|^2$.

Thus we can fix one such $\boldsymbol{\theta}'_* \in \text{proj}_{S\cap U}(\boldsymbol{\theta}_n)$ and call it $\boldsymbol{\theta}_{n,*}$. Then by point (b) and (c),

$$\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*} \in (T_{\boldsymbol{\theta}_{n,*}}S)^\perp = \text{Ker}(\nabla^2 f(\boldsymbol{\theta}_{n,*}))^\perp = \text{Range}(\nabla^2 f(\boldsymbol{\theta}_{n,*})),$$

which translates to our first assertion, $\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*} = \mathbf{P}_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*})$. And the second part of point (c) is exactly the second assertion. ■

One difficulty of the local convergence analysis comes from the fact that the gradient at $\boldsymbol{\theta}_n$ does not lie in the range of the Hessian at the local minimum $\boldsymbol{\theta}_*$. We overcome this by choosing a proper reference local minimum $\boldsymbol{\theta}_{n,*}$ given by lemma 15, $\nabla f(\boldsymbol{\theta}_n)$ 'almost' lie in the range of $\nabla^2 f(\boldsymbol{\theta}_{n,*})$, where the exact meaning of 'almost' is spelled in the lemma 16 below.

Lemma 16 *Suppose A1 holds. Let $\boldsymbol{\theta}_n$ be close enough to $\boldsymbol{\theta}_*$ satisfying A3. Let $\boldsymbol{\theta}_{n,*}$ and $\mathbf{P}_n$ be as before (see (19)), and let $\mathbf{Q}_n$ be the projection operator onto Ker$(\nabla^2 f(\boldsymbol{\theta}_{n,*}))$. Then*

$$\|\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n)\| \geq \mu\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\| - \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|^2, \tag{21}$$

$$\|\mathbf{Q}_n\nabla f(\boldsymbol{\theta}_n)\| \leq \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|^2. \tag{22}$$

In particular if $\mu\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\| \geq \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|^2$, then we have

$$\|\mathbf{Q}_n\nabla f(\boldsymbol{\theta}_n)\| \leq \frac{L_H}{2}\left(\mu - \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|\right)^{-2}\|\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n)\|^2, \tag{23}$$

$$\|\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\| \leq \left(1 - \frac{L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|}{\delta_n}\right)^{-1}\left(\frac{1}{\mu+\delta_n} + \frac{C_n\|\mathbf{P}_n(\nabla f(\boldsymbol{\theta}_n))\|}{\delta_n}\right)\|\mathbf{P}_n(\nabla f(\boldsymbol{\theta}_n))\|, \tag{24}$$

where $C_n = \frac{L_H}{2}\left(\mu - \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|\right)^{-2}$.

Proof By Taylor's theorem,

$$\nabla f(\boldsymbol{\theta}_n) = \nabla f(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_{n,*}) = \int_0^1 \nabla^2 f(\boldsymbol{\theta}_{n,*} + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}))dt\ (\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}).$$

Denote

$$\mathbf{R}(n,t) := \nabla^2 f(\boldsymbol{\theta}_{n,*} + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*})) - \nabla^2 f(\boldsymbol{\theta}_{n,*}).$$

Its operator norm has bound as $\|\mathbf{R}(n,t)\|_{op} \leq tL_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|$ by Lipschitz-continuity of the Hessian. Then we have

$$\begin{aligned}\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n) &= \mathbf{P}_n\nabla^2 f(\boldsymbol{\theta}_{n,*})(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}) + \mathbf{P}_n\int_0^1 \mathbf{R}(n,t)dt\ (\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}) \\ &= \nabla^2 f(\boldsymbol{\theta}_{n,*})\mathbf{P}_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}) + \mathbf{P}_n\int_0^1 \mathbf{R}(n,t)dt\ (\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}) \\ &= \nabla^2 f(\boldsymbol{\theta}_{n,*})(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}) + \mathbf{P}_n\int_0^1 \mathbf{R}(n,t)dt\ (\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*})\end{aligned}$$

where the last line holds from the first part of Lemma 15, and the previous line holds because $\nabla^2 f(\boldsymbol{\theta}_{n,*})$ and $\mathbf{P}_n$ commutes since they share the same frame of orthogonal eigenvectors. Now apply the norm on both sides and use the second part of Lemma 15, we get the estimate

$$\|\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n)\| \geq \mu\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\| - \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|^2. \tag{25}$$

Similarly we have

$$\begin{aligned}\|\mathbf{Q}_n\nabla f(\boldsymbol{\theta}_n)\| &= \left\|\nabla^2 f(\boldsymbol{\theta}_{n,*})\mathbf{Q}_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}) + \mathbf{Q}_n\int_0^1 \mathbf{R}(n,t)dt\ (\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*})\right\| \\ &= \left\|\mathbf{Q}_n\int_0^1 \mathbf{R}(n,t)dt\ (\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*})\|\right\| \\ &\leq \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|^2\end{aligned} \tag{26}$$

where the second line follows because $\text{Range}(\mathbf{Q}_n) = \text{Ker}(\nabla^2 f(\boldsymbol{\theta}_{n,*}))$. Since $\mu\|\boldsymbol{\theta}_n - \boldsymbol{\theta}'_*\| \geq \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}'_*\|^2$, squaring both sides of (25) we have

$$\|\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n)\|^2 \geq \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|^2\left(\mu - \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|\right)^2.$$

Plug this into (26) we have

$$\|\mathbf{Q}_n\nabla f(\boldsymbol{\theta}_n)\| \leq \frac{L_H}{2}\left(\mu - \frac{L_H}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|\right)^{-2}\|\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n)\|^2,$$

■

Another difficulty of the local convergence analysis comes from the fact that when $\boldsymbol{\theta}_n$ close to some $\boldsymbol{\theta}_*$ satisfying assumption A3, the preconditioning matrix $\mathbf{B}_n$ is very ill-conditioned. This is because under the assumption A3, the RPC factorization of the Hessian $\nabla^2 f(\boldsymbol{\theta}_n)$ is exact by lemma 9. Therefore, if $\boldsymbol{\theta}_n$ is close to $\boldsymbol{\theta}_*$, $\|\nabla f(\boldsymbol{\theta}_n)\|$ is close to zero, so $\lambda_{\min}(\mathbf{B}_n) = \delta_n = \sqrt{L_H\|\nabla f(\boldsymbol{\theta}_n)\|}$ can be arbitrarily small.

In the next lemma, we show that despite $\lambda_{\min}(\mathbf{B}_n)$ approaching zero, the size of $\|\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\|$ is well controlled thanks to the special geometrical alignment of $\nabla f(\boldsymbol{\theta}_n)$ shown in lemma 16.

Lemma 17 *Suppose A1 holds. Let $\boldsymbol{\theta}_n$ be sufficiently close to $\boldsymbol{\theta}_*$ satisfying A3. Define β_n to be the number so that $\|\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\| = \beta_n\|\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n)\|$, then*

$$\beta_n \leq 2\left(\frac{1}{\mu} + \frac{2\sqrt{L_H\|\mathbf{P}_n(\nabla f(\boldsymbol{\theta}_n))\|}}{\mu^2}\right). \tag{27}$$

In particular, $\beta_n = O(\mu^{-1})$ is uniformly bounded.

Proof Let $\boldsymbol{\theta}_{n,*}$ be the reference local minimum found by lemma 15. As in lemma 16, we denote

$$\mathbf{R}(n,t) := \nabla^2 f(\boldsymbol{\theta}_{n,*} + t(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*})) - \nabla^2 f(\boldsymbol{\theta}_{n,*}).$$

Recall now $\mathbf{B}_n = \nabla^2 f(\boldsymbol{\theta}_n) + \delta_n\mathbf{I}$ due to the exact factorization, we can write

$$\begin{aligned}
&\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\\
&= \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I} - \mathbf{R}(n,1)\right)^{-1}\left(\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n) + \mathbf{Q}_n\nabla f(\boldsymbol{\theta}_n)\right)\\
&= \left(\mathbf{I} - \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\mathbf{R}(n,1)\right)^{-1}\left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\left(\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n) + \mathbf{Q}_n\nabla f(\boldsymbol{\theta}_n)\right)\\
&= \left(\mathbf{I} - \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\mathbf{R}(n,1)\right)^{-1}\left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\mathbf{P}_n\nabla f(\boldsymbol{\theta}_n)\\
&\quad + \left(\mathbf{I} - \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\mathbf{R}(n,1)\right)^{-1}\left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\mathbf{Q}_n\nabla f(\boldsymbol{\theta}_n)
\end{aligned} \tag{28}$$

Now we bound the different parts of (28) individually. First part we have

$$\begin{aligned}
\left\|\left(\mathbf{I} - \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\mathbf{R}(n,1)\right)^{-1}\right\|_{op} &\leq \left(1 - \frac{\|\mathbf{R}(n,1)\|_{op}}{\delta_n}\right)^{-1}\\
&\leq \left(1 - \frac{L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|}{\delta_n}\right)^{-1}.
\end{aligned}$$

By A3, $\boldsymbol{\theta}_{n,*}$ is μ-PL, that is $\|\nabla f(\boldsymbol{\theta}_n)\|^2 \geq 2\mu(f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n,*}))$. By [24], $\boldsymbol{\theta}_{n,*}$ is also μ-QG, that is $f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n,*}) \geq \frac{\mu}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|^2$. Combine the two inequalities above we have

$$\|\nabla f(\boldsymbol{\theta}_n)\| \geq \mu\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|.$$

Recall now the regularization $\delta_n = \sqrt{L_H\|\nabla f(\boldsymbol{\theta}_n)\|}$, in which case we have,

$$\left(1 - \frac{L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|}{\delta_n}\right)^{-1} = \left(1 - \frac{L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|}{\sqrt{L_H\|\nabla f(\boldsymbol{\theta}_n)\|}}\right)^{-1} \leq \left(1 - \sqrt{\frac{L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|}{\mu}}\right)^{-1}.$$

Since $\boldsymbol{\theta}_n$ is sufficiently close to $\boldsymbol{\theta}_*$, and by the construction of $\boldsymbol{\theta}_{n,*}$ in lemma 15, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\| \leq \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|$, we may assume $\left(1 - \frac{L_H\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|}{\delta_n}\right)^{-1} \leq 2$. Then the first part of (27) has upper bound

$$\left\|\left(\mathbf{I} - \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n\mathbf{I}\right)^{-1}\mathbf{R}(n,1)\right)^{-1}\right\|_{op} \leq 2 \tag{29}$$

Using lemma 15, the second part of (27) can be easily bounded by

$$\|(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n \mathbf{I})^{-1} \mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\| \leq \frac{1}{\mu + \delta_n} \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\| \leq \frac{1}{\mu} \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|. \tag{30}$$

The third part is the key part that uses the lemma 16, according to which,

$$\left\| \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n \mathbf{I}\right)^{-1} \mathbf{Q}_n \nabla f(\boldsymbol{\theta}_n) \right\|_2 \leq \frac{C_n}{\delta_n} \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|^2,$$

where $C_n = \frac{L_H}{2} \left(\mu - \frac{L_H}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n,*}\|\right)^{-2}$. Since $\boldsymbol{\theta}_n$ is sufficiently close to $\boldsymbol{\theta}_*$, we may assume $C_n \leq \frac{2L_H}{\mu^2}$, then

$$\frac{C_n}{\delta_n} \leq \frac{2L_H}{\mu^2} \frac{1}{\sqrt{L_H \|\nabla f(\boldsymbol{\theta}_n)\|}} \leq \frac{2\sqrt{L_H}}{\mu^2 \sqrt{\|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|}}.$$

Thus we have

$$\left\| \left(\nabla^2 f(\boldsymbol{\theta}_{n,*}) + \delta_n \mathbf{I}\right)^{-1} \mathbf{Q}_n \nabla f(\boldsymbol{\theta}_n) \right\|_2 \leq \frac{2\sqrt{L_H}}{\mu^2} \|\mathbf{P}_n (\nabla f(\boldsymbol{\theta}_n))\|^{3/2}. \tag{31}$$

Then combine (29), (30), and (31) into (27), we have

$$\|\mathbf{B}_n^{-1} \nabla f(\boldsymbol{\theta}_n)\| \leq 2 \left(\frac{1}{\mu} + \frac{2\sqrt{L_H}}{\mu^2} \sqrt{\|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|} \right) \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|.$$

Since β_n is the number so that $\|\mathbf{B}_n^{-1} \nabla f(\boldsymbol{\theta}_n)\| = \beta_n \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|$, then

$$\beta_n \leq 2 \left(\frac{1}{\mu} + \frac{2\sqrt{L_H \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|}}{\mu^2} \right).$$

■

Lemma 18 *Under the same assumptions as lemma 17*

$$\|\nabla f(\boldsymbol{\theta}_{n+1})\| \leq \left(\delta_n \beta_n + \frac{L_H}{2} \beta_n^2 \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\| \right) \|\mathbf{P}_n \nabla f(\boldsymbol{\theta}_n)\|. \tag{32}$$

In particular, for any $\boldsymbol{\theta}_n$ in a small enough neighborhood of $\boldsymbol{\theta}_*$, $\|\nabla f(\boldsymbol{\theta}_{n+1})\| \leq c \|\nabla f(\boldsymbol{\theta}_n)\|$ for some $0 < c < 1$.

Proof Recall the next iteration is given by

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{B}_n^{-1} \nabla f(\boldsymbol{\theta}_n.)$$

Then by Taylor's theorem we can compute gradient at $\boldsymbol{\theta}_{n+1}$

$$\begin{aligned}
\nabla f(\boldsymbol{\theta}_{n+1}) &= \nabla f(\boldsymbol{\theta}_n - \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)) \\
&= \nabla f(\boldsymbol{\theta}_n) - \int_0^1 \nabla^2 f(\boldsymbol{\theta}_n - t\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n))dt\ \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) \\
&= \mathbf{B}_n\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) - \int_0^1 \nabla^2 f(\boldsymbol{\theta}_n - t\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n))dt\ \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) \\
&= \nabla^2 f(\boldsymbol{\theta}_n)\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) + \delta_n\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) \\
&\quad - \int_0^1 \nabla^2 f(\boldsymbol{\theta}_n - t\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n))dt\ \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) \\
&= \delta_n\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n) - \int_0^1 \left[\nabla^2 f(\boldsymbol{\theta}_n) - \nabla^2 f(\boldsymbol{\theta}_n - t\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n))\right] dt\ \mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n).
\end{aligned}$$

Apply norm to both sides above we have

$$\|\nabla f(\boldsymbol{\theta}_{n+1})\| \leq \delta_n\|\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\| + \frac{L_H}{2}\|\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\|^2. \tag{33}$$

Plug $\|\mathbf{B}_n^{-1}\nabla f(\boldsymbol{\theta}_n)\| = \beta_n\|\mathbf{P}\nabla f(\boldsymbol{\theta}_n)\|$ into (33) we get

$$\|\nabla f(\boldsymbol{\theta}_{n+1})\| \leq \left(\delta_n\beta_n + \frac{L_H\beta_n^2}{2}\|\mathbf{P}\nabla f(\boldsymbol{\theta}_n)\|\right)\|\mathbf{P}\nabla f(\boldsymbol{\theta}_n)\|.$$

And the last assertion follows from the fact that β_n is bounded from above by lemma 17 and that as $\boldsymbol{\theta}_n$ approaches $\boldsymbol{\theta}_*$, $\delta_n\beta_n + \frac{L_H\beta_n^2}{2}\|\mathbf{P}\nabla f(\boldsymbol{\theta}_n)\| = \sqrt{L_H\|\nabla f(\boldsymbol{\theta}_n)\|}\beta_n + \frac{L_H\beta_n^2}{2}\|\mathbf{P}\nabla f(\boldsymbol{\theta}_n)\|$ goes to zero.

■

Having good one step update, to get a local convergence result, it suffices to know that once one of the iterate is close enough to $\boldsymbol{\theta}_*$, then we can apply the per-iteration improvement (32) above for all future iterates. And this is shown to be true using Lyapunov stability from [27].

Definition 19 (Vanishing Steps) *An algorithm has Vanishing Steps property on set $\boldsymbol{\Theta}$ is there is an open neighborhood $U \supseteq \boldsymbol{\Theta}$ and a continuous function $\eta : \mathbb{R}^N \to [0, \infty)$ so that $\eta|_{\boldsymbol{\Theta}} = 0$ and if $\boldsymbol{\theta}_n \in U$, then the next iterate $\boldsymbol{\theta}_{n+1}$ satisfies*

$$\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| \leq \eta(\boldsymbol{\theta}_n)$$

Definition 20 (Bounded Path Length) *An algorithm has Bounded Path Length property on set $\boldsymbol{\Theta}$ is there is an open neighborhood $U \supseteq \boldsymbol{\Theta}$ and a continuous function $\gamma : \mathbb{R}^N \to [0, \infty)$ so that $\gamma|_{\boldsymbol{\Theta}} = 0$ and if consecutive iterates $\boldsymbol{\theta}_n, ..., \boldsymbol{\theta}_{n+m} \in U$, then*

$$\sum_{i=n}^{n+m-1} \|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\| \leq \gamma(\boldsymbol{\theta}_n)$$

Lemma 21 (Lyapunov stability, Prop. 3.5 in [27]) *Suppose the algorithm satisfies definition 19 and 20 at $\boldsymbol{\theta}_* \in \mathbb{R}^N$. Given a neighborhood U of $\boldsymbol{\theta}_*$, there is a neighborhood V of $\boldsymbol{\theta}_*$ so that once $\boldsymbol{\theta}_n \in V$, then all $\boldsymbol{\theta}_{n+i} \in U$ for all $i \geq 1$.*

Proof [**Proof of Theorem 4**] First let us show that if one iterate $\boldsymbol{\theta}_n$ is in a sufficiently small neighborhood of $\boldsymbol{\theta}_*$, then all the future iterates will also stay in that neighborhood. By lemma 18, there is $0 < c < 1$ so that

$$\|\nabla f(\boldsymbol{\theta}_{n+1})\| \leq c\|\nabla f(\boldsymbol{\theta}_n)\|$$

provided $\boldsymbol{\theta}_n$ is in a sufficiently small neighborhood of $\boldsymbol{\theta}_*$. Suppose for some $\boldsymbol{\theta}_n, ..., \boldsymbol{\theta}_{n+m}$ in this neighborhood, then

$$\begin{aligned}\sum_{i=n}^{n+m-1} \|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\| &= \sum_{i=n}^{n+m-1} \|\mathbf{B}_i^{-1}\nabla f(\boldsymbol{\theta}_i)\| \\ &\leq \sum_{i=n}^{n+m-1} \beta_i c^{i-n}\|\nabla f(\boldsymbol{\theta}_n)\| \\ &\leq (\max_{i\geq n}\beta_i)\frac{1}{1-c}\|\nabla f(\boldsymbol{\theta}_n)\|.\end{aligned}$$

Therefore our algorithm satisfies the definition 20 at local min $\boldsymbol{\theta}_*$. And by lemma 21, there is neighborhood U, V of $\boldsymbol{\theta}_*$ so that $\boldsymbol{\theta}_n \in V$ implies $\boldsymbol{\theta}_{n+i} \in U$ for all $i \geq 1$. And in neighborhood U we have per-iteration improvement (32), from which we have

$$\frac{\|\nabla f(\boldsymbol{\theta}_{n+1})\|}{\|\nabla f(\boldsymbol{\theta}_n)\|} \leq \left(\delta_n\beta_n + \frac{L_H\beta_n^2}{2}\|\nabla f(\boldsymbol{\theta}_n)\|\right),$$

which goes to zero as $n \to \infty$. So $\|\nabla f(\boldsymbol{\theta}_n)\|$ converges to 0 superlinearly. More explicitly, we have

$$\|\nabla f(\boldsymbol{\theta}_{n+1})\| \leq \left(\delta_n\beta_n\|\nabla f(\boldsymbol{\theta}_n)\| + \frac{L_H\beta_n^2}{2}\|\nabla f(\boldsymbol{\theta}_n)\|^2\right).$$

Recall that by lemma 17, $\beta = O(1/\mu)$, and that $\delta_n = \sqrt{L_H\|\nabla f(\boldsymbol{\theta}_n)\|}$. Thus

$$\|\nabla f(\boldsymbol{\theta}_{n+1})\| \leq C\|\nabla f(\boldsymbol{\theta}_n)\|^{3/2},$$

for some $C = O(1/\mu)$, where μ is the PL constant of defined in A3.

■

Remark 22 *We note that unlike other recent results giving superlinear convergence rate for classical quasi-Newton methods [14, 28, 29], our result requires $\boldsymbol{\theta}_0$ to be sufficiently close to a local minimizer $\boldsymbol{\theta}_*$ satisfying A3. This is likely due to the fact that we do not require the objective function to be strongly convex and that the Hessian of our objective function has many flat directions near $\boldsymbol{\theta}_*$.*