OJMO

Open Journal of Mathematical Optimization

Terunari Fuji, Pierre-Louis Poirion & Akiko Takeda

Theoretical analysis of the randomized subspace regularized Newton method for non-convex optimization

Volume 6 (2025), article no. 8 (35 pages)

https://doi.org/10.5802/ojmo.46

Article submitted on June 11, 2024, revised on April 23, 2025,
accepted on September 11, 2025.

CENTRE MERSENNE

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

Theoretical analysis of the randomized subspace regularized Newton method for non-convex optimization

Terunari Fuji
Graduate School of Information Science and Technology, University of Tokyo
teru0818258@gmail.com

Pierre-Louis Poirion
Center for Advanced Intelligence Project, RIKEN
pierre-louis.poirion@riken.jp

Akiko Takeda
Graduate School of Information Science and Technology, University of Tokyo; Center for Advanced Intelligence Project, RIKEN
takeda@mist.i.u-tokyo.ac.jp

Abstract

While there already exist randomized subspace Newton methods that restrict the search direction to a random subspace for a convex function, we propose a randomized subspace regularized Newton method for a non-convex function and more generally we investigate thoroughly, for the first time, the local convergence rate of the randomized subspace Newton method. In our proposed algorithm, we use a modified Hessian of the function restricted to some random subspace so that, with high probability, the function value decreases at each iteration, even when the objective function is non-convex. We show that our method has global convergence under appropriate assumptions and its convergence rate is the same as that of the full regularized Newton method. Furthermore, we obtain a local linear convergence rate under some additional assumptions, and prove that this rate is the best we can hope, in general, when using a random subspace. We furthermore prove that if the Hessian, at the local optimum, is rank deficient then super-linear convergence holds.

Digital Object Identifier 10.5802/ojmo.46

Keywords Random projection, Newton method, non-convex optimization, local convergence rate.

Acknowledgments This work was supported by Grant-in-Aid for Scientific Research (B) (No. 23H03351) and JST ERATO (JPMJER1903).

1 Introduction

While first-order optimization methods such as stochastic gradient descent methods are well studied for large-scale machine learning optimization, second-order optimization methods have not received much attention due to the high cost of computing second-order information until recently. However, in order to overcome relatively slow convergence of first-order methods, there has been recent interest in second-order methods that aim to achieve faster convergence speed by utilizing subsampled Hessian information and stochastic Hessian estimate (see e.g., [4, 44, 46] and references therein).

In this paper, we develop a Newton-type iterative method with random projections for the following unconstrained optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a possibly non-convex twice differentiable function. In our method, at each iteration, we restrict the function f to a random subspace and compute the next iterate by choosing a descent direction on this random subspace.

There are some existing studies on developing second-order methods with random subspace techniques for convex optimization problems (1). Let us now review randomized subspace Newton (RSN) existing work [18],

while gradient-based randomized subspace algorithms are reviewed in Section 2.1. RSN computes the descent direction d_k and the next iterate as

$$d_k^{\mathrm{RSN}} = -P_k^{\mathsf{T}}(P_k\nabla^2 f(x_k)P_k^{\mathsf{T}})^{-1}P_k\nabla f(x_k),$$
$$x_{k+1} = x_k + \frac{1}{\widehat{L}}d_k^{\mathrm{RSN}},$$

where $P_k \in \mathbb{R}^{s\times n}$ is a random matrix with $s < n$ and $\widehat{L}$ is some fixed constant. RSN is expected to be highly computationally efficient with respect to the original Newton method, since it does not require computation of the full Hessian inverse. RSN is also shown to achieve a global linear convergence for strongly convex f. We first note that the second-order Taylor approximation around x_k restricted in the affine subspace $\{x_k\} + \mathrm{Range}(P_k^{\mathsf{T}})$ is expressed as

$$f(x_k + P_k^{\mathsf{T}}u) \simeq f(x_k) + \nabla f(x_k)^{\mathsf{T}}P_k^{\mathsf{T}}u + \frac{1}{2}u^{\mathsf{T}}P_k\nabla^2 f(x_k)P_k^{\mathsf{T}}u,$$

and the direction d_k^{RSN} is obtained as $d_k^{\mathrm{RSN}} = P_k^{\mathsf{T}}u_k^*$ where u_k^* is the minimizer of the above subspace Taylor approximation, i.e.,

$$u_k^* = \operatorname*{arg\,min}_{u\in\mathbb{R}^s}\left(f(x_k) + \nabla f(x_k)^{\mathsf{T}}P_k^{\mathsf{T}}u + \frac{1}{2}u^{\mathsf{T}}P_k\nabla^2 f(x_k)P_k^{\mathsf{T}}u\right).$$

Hence, we can see that the next iterate of RSN is computed by using the Newton direction for the function

$$f_{x_k} \ : \ u \longmapsto f(x_k + P_k^{\top}u). \tag{2}$$

Other second-order subspace descent methods, such as cubically-regularized subspace Newton methods, [22], have been studied in the literature. More precisely, the method in [22] can be seen as a stochastic extension of the cubically-regularized Newton method [32] and also as a second-order enhancement of stochastic subspace descent [28]. In [27], a random subspace version of the BFGS method is proposed. The authors prove local linear convergence, if the function is assumed to be self-concordant. Apart in recent Shao's Ph.D thesis [37] and the associated papers [11, 12] which have been done parallelly to this paper, to the best of our knowledge, existing second-order subspace methods have iteration complexity analysis only for convex optimization problems. The thesis [37] and the paper [12] propose a random subspace adaptive regularized cubic method for unconstrained non-convex optimization and show a global convergence property with sub-linear rate to a stationary point[1]. In this paper we propose a new subspace method based on the regularized Newton method and discuss the local convergence rate together with global iteration complexity.[2] Notice indeed that, to the best of our knowledge, the local convergence of such methods never seems to have been thoroughly studied[3]; one would expect super-linear convergence for second order methods and no papers discuss whether super-linear convergence holds or not for second order methods. Indeed any iterative algorithm can be easily adapted to a random subspace method as it suffices to apply it to the function restricted to the subspace: $u \mapsto f(x_k + P_k^{\top}u)$. We therefore believe that it is important to investigate thoroughly whether the properties of such full-space algorithms are preserved or not when adapted to the random subspace setting. If the objective function f is not convex, the Hessian is not always positive semidefinite and d_k^{RSN} is not guaranteed to be a descent direction so that we need to use a modified Hessian. Based on the regularized Newton method (RNM) for the unconstrained non-convex optimization [39, 40], we propose the randomized subspace regularized Newton method (RS-RNM):

$$d_k = -P_k^{\mathsf{T}}(P_k\nabla^2 f(x_k)P_k^{\mathsf{T}} + \eta_k I_s)^{-1}P_k\nabla f(x_k),$$
$$x_{k+1} = x_k + t_k d_k,$$

where η_k is defined to ensure that search direction d_k is a descent direction and the step size t_k is chosen so that it satisfies Armijo's rule. As with RSN, this algorithm is expected to be computationally efficient since we use projections onto lower-dimensional spaces. In this paper, we first show that RS-RNM has global convergence

[1] The author also proves that if the Hessian matrix has low rank and scaled Gaussian sketching matrices are used, then the Hessian at the stationary point is approximately positive semidefinite with high probability.

[2] Just as the ordinary cubic method is superior to the Newton method in terms of iteration complexity, similar observation seems to hold between the subspace cubic method [37] and ours.

[3] Some papers, as we will see later, investigate when local linear convergence holds.

under appropriate assumptions, more precisely, we have $\|\nabla f(x_k)\| \leq \varepsilon$ after at most $O(\varepsilon^{-2})$ iterations with some probability, which is the same as the global convergence rate shown in [39] for the full regularized Newton method. We then prove that under additional assumptions, we can obtain a linear convergence rate locally. In particular, one contribution of the paper is to propose, to the best of our knowledge, the weakest conditions until now for local linear convergence. To do so we will extensively use the fact that the subspace is chosen at random. From these conditions, we can derive a random-projection version of the Polyak–Lojasiewicz (PL) inequality (3),

$$\forall\, x \in \mathbb{R}^n, \quad \|\nabla f(x)\|^2 \geq c_0(f(x) - f(x^*)), \tag{3}$$

which will be satisfied when the function is restricted to a random subspace. One other contribution of this paper is to prove that, in general, linear convergence is the best rate we can hope for this method. Furthermore, we also prove that if the Hessian at the local optima is rank deficient, then one can achieve super-linear convergence using a subspace dimension s large enough.

Our randomized subspace method for nonconvex optimization problems is based on the regularized Newton method in [39, 40]. While various other regularized Newton methods have been proposed in recent years, most of them are for convex problems or non-smooth optimization problems. For example, [31] presents a globally convergent proximal Newton-type method for non-smooth convex optimization and [8] develops coderivative-based Newton methods combined with Wolfe line-search for non-smooth problems. Recently [45] proposes a generalized regularization method that includes quadratic, cubic, and elastic net regularizations. Also [14] proposes, in the convex case, a variant of the Newton method with quadratic regularization and proves better global rate. Recent papers, [19, 47, 48], propose regularization methods for the non-convex case. However, although these methods obtained better iterations complexity, the subroutines involved to compute are quite complex and not as simple as in [39, 40]. By applying similar random subspace techniques to these methods, we may be able to develop random subspace variants with state-of-the-art theoretical guarantees, but that is a topic for future work.

The rest of this paper is organized as follows. After reviewing gradient-based randomized subspace algorithms and introducing properties of random projections in Section 2, we introduce our random subspace Newton method for non-convex functions in Section 3. In Section 4, we prove global convergence properties for our method. In Section 5, we investigate local linear convergence as well as local super-linear convergence. Finally, in Section 6, we show some numerical examples to illustrate the theoretical properties derived in the paper. In Section 7 we conclude the paper.

2 Preliminaries

▶ Notation. In this paper we call a matrix $P \in \mathbb{R}^{s\times n}$ a random projection matrix or a random matrix when its entries P_{ij} are independently sampled from the normal distribution $\mathrm{N}(0, 1/s)$. Let I_n be the identity matrix of size n. We denote by g_k the gradient of the k-th iterate of the obtained sequence and by H_k it's Hessian.

2.1 Related optimization algorithms using random subspace

As introduced in Section 1, random subspace techniques are used for second-order optimization methods with the aim of reducing the size of Hessian matrix. Here we refer to other types of subspace methods focusing on their convergence properties.

Cartis et al. [6] proposed a general random embedding framework for global optimization of a function f. The framework projects the original problem onto a random subspace and solves the reduced subproblem in each iteration:

$$\min_u f(x_k + P_k^\top u) \text{ subject to } x_k + P_k^\top u \in \mathcal{C}.$$

These subproblems need to be solved to some required accuracy by using a deterministic global optimization algorithm. This study is further expanded in [7] and [5], when f has low effective dimension.

There are also various subspace first-order methods based on coordinate descent methods (see e.g. [43]). In [9] a randomized coordinate descent algorithm is introduced assuming some subspace decomposition which is suited to the A-norm, where A is a given preconditioner. In [30], minimizing $f(\widetilde{A}x) + \frac{\lambda}{2}\|x\|^2$, where f is a strongly convex smooth function and $\widetilde{A}$ is a high-dimensional matrix, is considered and a new randomized optimization method that can be seen as a generalization of coordinate descent to random subspaces is proposed. The paper [20] deals with a convex optimization problem $\min_x f(x) + g(x)$, where f is convex and differentiable

and g is assumed to be convex, non-smooth and sparse inducing such as $\|x\|_1$. To solve the problem, they propose a randomized proximal algorithm leveraging structure identification: the variable space is sampled according to the structure of g. The approach in [38] is to optimize a smooth convex function by choosing, at each iteration a random direction on the sphere. Recently, in some contexts such as iteration complexity analysis, the assumption of strong convexity has been replaced by a weaker one, the PL inequality (3). Indeed, [29] has introduced a new first-order random subspace and proved that if the non-convex function is differentiable with a Lipschitz continuous first derivative and satisfies the PL inequality (3) then linear convergence rate is obtained in expectation. Notice that in all these papers a local linear convergence rate is only obtained when assuming that the objective function is, at least locally, strongly convex or satisfies the PL inequality.

From above, without (locally) strong convexity nor the PL inequality, it seems difficult to construct first-order algorithms having (local) linear convergence rates. Indeed, the probabilistic direct-search method [34] in reduced random spaces is applicable to both convex and non-convex problems but it obtains sub-linear convergence.

In this paper, we will prove that our algorithm achieves local linear convergence rates without locally strong convexity nor the PL inequality assumption on the full space. This is due to randomized Hessian information used in our algorithm. More precisely, our assumptions will allow us to prove that the function, restricted to a random subspace, satisfies a condition similar to the PL inequality.

2.2 Properties of random projection

In this section, we recall basic properties of random projection matrices. One of the most important features of a random projection defined by a random matrix is that it nearly preserves the norm of any given vector with arbitrary high probability. The following lemma is known as a variant of the Johnson–Lindenstrauss lemma [25].

▶ **Lemma 1** ([41, Lemma 5.3.2, Exercise 5.3.3]). *Let $P \in \mathbb{R}^{s\times n}$ be a random matrix whose entries P_{ij} are independently drawn from $\mathrm{N}(0, 1/s)$. Then for any $x \in \mathbb{R}^n$ and $\varepsilon \in (0, 1)$, we have*

$$\mathrm{Prob}[(1-\varepsilon)\|x\|^2 \le \|Px\|^2 \le (1+\varepsilon)\|x\|^2] \ge 1 - 2\exp(-\mathcal{C}_0\varepsilon^2 s),$$

where $\mathcal{C}_0$ is an absolute constant.

The next lemma shows that when P is a Gaussian matrix, we can obtain a bound on the norm of $PP^\top$.

▶ **Lemma 2.** *For a $s \times n$ random matrix P whose entries are sampled from $\mathrm{N}(0, 1/s)$, there exists a constant $\bar{\mathcal{C}} > 0$ such that*

$$\left\|PP^{\mathsf{T}}\right\| \left(= \left\|P^{\mathsf{T}}P\right\| = \|P\|^2\right) \le \bar{\mathcal{C}}\frac{n}{s},$$

with probability at least $1 - 2e^{-s}$.

Proof. By [41, Theorem 4.6.1], there exists a constant $\overline{C}$ such that

$$\left\|\frac{s}{n}PP^\top - I_s\right\| \le 2\overline{C}\sqrt{\frac{s}{n}}$$

holds with probability at least $1 - 2e^{-s}$. Therefore, we have

$$\left\|PP^{\mathsf{T}}\right\| \le \left\|PP^{\mathsf{T}} - \frac{n}{s}I_s\right\| + \left\|\frac{n}{s}I_s\right\| \le 2\overline{C}\sqrt{\frac{n}{s}} + \frac{n}{s} \le 2\overline{C}\frac{n}{s} + \frac{n}{s} = (2\overline{C}+1)\frac{n}{s}.$$

Setting $\bar{\mathcal{C}} = 2\overline{C} + 1$ ends the proof. ◀

All the results of this paper are stated in a probabilistic way. In the proofs we will constantly use the following fact:

$$\text{For any two events } E_1 \text{ and } E_2 : \ \mathrm{Prob}(E_1 \cap E_2) \ge 1 - ((1 - \mathrm{Prob}(E_1)) + (1 - \mathrm{Prob}(E_2))). \tag{4}$$

3 Randomized subspace regularized Newton method

In this section, we describe a randomized subspace regularized Newton method (RS-RNM) for the following unconstrained minimization problem,

$$\min_{x\in\mathbb{R}^n} f(x), \tag{5}$$

Algorithm 1 Randomized subspace regularized Newton method (RS-RNM)

input: $x_0 \in \mathbb{R}^n$, $0 \leq \gamma < 1, c_1 > 1, c_2 > 0, \alpha, \beta \in (0,1)$

1: $k \leftarrow 0$
2: **repeat**
3: sample a random matrix: $P_k \sim \mathcal{D}$
4: compute the regularized sketched Hessian:

$$M_k = P_k H_k P_k^\mathsf{T} + c_1 \Lambda_k I_s + c_2 \|g_k\|^\gamma I_s,$$

where $\Lambda_k = \max(0, -\lambda_{\min}(P_k H_k P_k^\mathsf{T}))$
5: compute the search direction: $d_k = -P_k^\mathsf{T} M_k^{-1} P_k g_k$
6: apply the backtracking line search with Armijo's rule by finding the smallest integer $l_k \geq 0$ such that (8) holds. Set $t_k = \beta^{l_k}$, $x_{k+1} = x_k + t_k d_k$ and $k \leftarrow k+1$
7: **until** some stopping criteria is satisfied **return** the last iterate x_k

where f is a twice continuously differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$. In what follows, we denote the gradient $\nabla f(x_k)$ and the Hessian $\nabla^2 f(x_k)$ as g_k and H_k, respectively.

The paper [39] develops a regularized Newton methods (RNM) that constructs a sequence of iterates with the following update rule:

$$x_{k+1} = x_k - t_k (H_k + c_1' \Lambda_k' I_n + c_2' \|g_k\|^{\gamma'} I_n)^{-1} g_k,$$

where $\Lambda_k' = \max(0, -\lambda_{\min}(H_k))$, c_1', c_2', γ' are some positive parameter values and t_k is the step-size chosen by Armijo's step size rule, and show that this algorithm achieves $\|g_k\| \leq \varepsilon$ after at most $O(\varepsilon^{-2})$ iterations and it has a super-linear rate of convergence in a neighborhood of a local optimal solution under appropriate conditions.

To increase the computational efficiency of this algorithm using random projections, based on the randomized subspace Newton method [18], we propose the randomized subspace regularized Newton method (RS-RNM) with Armijo's rule, which is described in Algorithm 1 and outlined below. Since RS-RNM is a subspace version of RNM, all discussions of global convergence guarantees made in Section 4 are based on the one in [39].

Let $\mathcal{D}$ denote the set of Gaussian matrices of size $s \times n$ whose entries are independently sampled from $\mathrm{N}(0, 1/s)$. With a Gaussian random matrix P_k from $\mathcal{D}$, the regularized sketched Hessian is defined by:

$$M_k := P_k H_k P_k^\mathsf{T} + \eta_k I_s \in \mathbb{R}^{s \times s}, \tag{6}$$

where $\eta_k := c_1 \Lambda_k + c_2 \|g_k\|^\gamma$ and $\Lambda_k := \max(0, -\lambda_{\min}(P_k H_k P_k^\mathsf{T}))$. We then compute the search direction:

$$d_k := -P_k^\mathsf{T} M_k^{-1} P_k g_k. \tag{7}$$

The costly part of Newton-based methods, the inverse computation of a (approximate) Hessian matrix, is done in the subspace of size s. We note that d_k defined by (7) is a descent direction for f at x_k, i.e., $g_k^\top d_k < 0$ if $g_k \neq 0$, since it turns out that M_k is positive definite from the definition of Λ_k, and therefore $x^\top P_k^\mathsf{T} M_k^{-1} P_k x > 0$ holds for $\forall\, x$ due to $P_k x \neq 0$ with high probability.

The backtracking line search with Armijo's rule finds the smallest integer $l_k \geq 0$ such that

$$f(x_k) - f(x_k + \beta^{l_k} d_k) \geq -\alpha \beta^{l_k} g_k^\mathsf{T} d_k. \tag{8}$$

Starting with $l_k = 0$, l_k is increased by $l_k \leftarrow l_k + 1$ until the condition (8) holds. The sufficient iteration number to find such a step-size is discussed in convergence analysis later.

4 Global convergence properties

In Section 4.1, we discuss the global convergence of the RS-RNM under Assumption 3. We further prove the global iteration complexity of the algorithm in Section 4.2 by considering further assumptions.

▶ **Assumption 3.** *The level set of f at the initial point x_0 is bounded, i.e., $\Omega := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is bounded.*

By (8), we have that for any $k \in \mathbb{N}$, $f(x_{k+1}) \leq f(x_k)$, implying all $x_k \in \Omega$. Since Ω is a bounded set and f is continuously differentiable, there exists $U_g > 0$ such that

$$\|g_k\| \leq U_g, \quad \forall\, k \geq 0. \tag{9}$$

Similarly, there exists $L > 0$ such that for all $x \in \Omega$,

$$\|\nabla^2 f(x)\| \leq L. \tag{10}$$

In particular, for all $k > 0$,

$$\|H_k\| \leq L. \tag{11}$$

Notice that by (10), ∇f is L-Lipschitz continuous. We also define $f^* = \inf_{x\in\Omega} f(x)$.

4.1 Global convergence

We first show that the norm of d_k can be bounded from above.

▶ **Lemma 4.** *Suppose that $\|d_k\| \neq 0$. Then, d_k defined by (7) satisfies*

$$\|d_k\| \leq \bar{\mathcal{C}}\frac{n}{s}\frac{\|g_k\|^{1-\gamma}}{c_2},$$

with probability at least $1 - 2e^{-s}$.

Proof. By Lemma 2 we have $\left\|P_k^\mathsf{T} P_k\right\| \leq \bar{\mathcal{C}}\frac{n}{s}$, holds with probability at least $1 - 2e^{-s}$. Then, it follows from (7) that

$$\begin{aligned}
\|d_k\| &= \left\|P_k^\mathsf{T} M_k^{-1} P_k g_k\right\| \\
&= \left\|P_k^\mathsf{T}(P_k H_k P_k^\mathsf{T} + \eta_k I_s)^{-1} P_k g_k\right\| \\
&\leq \left\|P_k^\mathsf{T}(P_k H_k P_k^\mathsf{T} + \eta_k I_s)^{-1} P_k\right\| \|g_k\| \\
&\leq \left\|P_k^\mathsf{T}\right\| \|P_k\| \left\|(P_k H_k P_k^\mathsf{T} + \eta_k I_s)^{-1}\right\| \|g_k\| \\
&= \frac{\left\|P_k^\mathsf{T} P_k\right\| \|g_k\|}{\lambda_{\min}(P_k H_k P_k^\mathsf{T} + c_1\Lambda_k I_s + c_2 \|g_k\|^\gamma I_s)} \quad (\text{as } \left\|P_k^\mathsf{T}\right\| \|P_k\| = \left\|P_k^\mathsf{T} P_k\right\|) \\
&\leq \bar{\mathcal{C}}\frac{n}{s}\frac{\|g_k\|^{1-\gamma}}{c_2}.
\end{aligned}$$

◀

We next show that, when $\|g_k\|$ is at least ε away from 0, $\|d_k\|$ is bounded above by some constant.

▶ **Lemma 5.** *Suppose that Assumption 3 holds. Suppose also that there exists $\varepsilon > 0$ such that $\|g_k\| \geq \varepsilon$. Then, with probability at least $1 - 2e^{-s}$, d_k defined by (7) satisfies*

$$\|d_k\| \leq r(\varepsilon), \tag{12}$$

where

$$r(\varepsilon) := \frac{\bar{\mathcal{C}}n}{c_2 s}\max\left(U_g^{1-\gamma}, \frac{1}{\varepsilon^{\gamma-1}}\right).$$

Proof. If $\gamma \leq 1$, it follows from Lemma 4 and (9) that

$$\|d_k\| \leq \frac{\bar{\mathcal{C}}n}{s}\frac{U_g^{1-\gamma}}{c_2}.$$

Meanwhile, if $\gamma > 1$, it follows from Lemma 4 and $\|g_k\| \geq \varepsilon$ that

$$\|d_k\| \leq \frac{\bar{\mathcal{C}}n}{s}\frac{1}{c_2\varepsilon^{\gamma-1}}.$$

This completes the proof. ◀

When $\|g_k\| \geq \varepsilon$, we have from Lemma 5 that

$$x_k + \tau d_k \in \Omega + B(0, r(\varepsilon)), \quad \forall\, \tau \in [0,1].$$

By boundedness of $\Omega + B(0, r(\varepsilon))$ and by using the fact that f is twice continuously differentiable, we deduce that there exists $U_H(\varepsilon) > 0$ such that

$$\left\|\nabla^2 f(x)\right\| \leq U_H(\varepsilon), \quad \forall\, x \in \Omega + B(0, r(\varepsilon)). \tag{13}$$

The following lemma indicates that a step size smaller than some constant satisfies Armijo's rule when $\|g_k\| \geq \varepsilon$.

▶ **Lemma 6.** *Suppose that Assumption 3 holds. Suppose also that there exists $\varepsilon > 0$ such that $\|g_k\| \geq \varepsilon$. Then, with probability at least $1 - 2e^{-s}$, a step size $t'_k > 0$ such that*

$$t'_k \leq \frac{2(1-\alpha)c_2^2 \varepsilon^{2\gamma} s}{((1+c_1)\frac{\bar{\mathcal{C}}n}{s} U_H(\varepsilon) + c_2 U_g^{\gamma}) U_H(\varepsilon) \bar{\mathcal{C}} n}$$

satisfies Armijo's rule, i.e.,

$$f(x_k) - f(x_k + t'_k d_k) \geq -\alpha t'_k g_k^{\mathsf{T}} d_k.$$

Proof. From Taylor's theorem, there exists $\tau'_k \in (0,1)$ such that

$$f(x_k + t'_k d_k) = f(x_k) + t'_k g_k^{\mathsf{T}} d_k + \frac{1}{2} {t'_k}^2 d_k^{\mathsf{T}} \nabla^2 f(x_k + \tau'_k t'_k d_k) d_k.$$

Then, we have

$$\begin{aligned}
&f(x_k) - f(x_k + t'_k d_k) + \alpha t'_k g_k^{\mathsf{T}} d_k \\
&= (\alpha - 1) t'_k g_k^{\mathsf{T}} d_k - \frac{1}{2} {t'_k}^2 d_k^{\mathsf{T}} \nabla^2 f(x_k + \tau'_k t'_k d_k) d_k \\
&= (1-\alpha) t'_k g_k^{\mathsf{T}} P_k^{\mathsf{T}} M_k^{-1} P_k g_k - \frac{1}{2} {t'_k}^2 g_k^{\mathsf{T}} P_k^{\mathsf{T}} M_k^{-1} P_k \nabla^2 f(x_k + \tau'_k t'_k d_k) P_k^{\mathsf{T}} M_k^{-1} P_k g_k \qquad (14) \\
&\qquad\qquad \text{(by (7))} \\
&\geq (1-\alpha) t'_k \lambda_{\min}(M_k^{-1}) \left\|P_k g_k\right\|^2 \\
&\qquad - \frac{1}{2} {t'_k}^2 \lambda_{\max}(\nabla^2 f(x_k + \tau'_k t'_k d_k)) \lambda_{\max}(M_k^{-1} P_k P_k^{\mathsf{T}} M_k^{-1}) \left\|P_k g_k\right\|^2 \\
&\geq (1-\alpha) t'_k \lambda_{\min}(M_k^{-1}) \left\|P_k g_k\right\|^2 - \frac{1}{2} {t'_k}^2 U_H(\varepsilon) \lambda_{\max}(M_k^{-1} P_k P_k^{\mathsf{T}} M_k^{-1}) \left\|P_k g_k\right\|^2, \\
&\qquad\qquad \text{(by (13))}
\end{aligned}$$

where the first inequality derives from the fact that

$$\begin{aligned}
g_k^{\mathsf{T}} P_k^{\mathsf{T}} M_k^{-1} P_k \nabla^2 f(x_k + \tau'_k t'_k d_k) P_k^{\mathsf{T}} M_k^{-1} P_k g_k &\leq \lambda_{\max}(M_k^{-1} P_k \nabla^2 f(x_k + \tau'_k t'_k d_k) P_k^{\mathsf{T}} M_k^{-1}) \|P_k g_k\|^2 \\
&\leq \lambda_{\max}(\nabla^2 f(x_k + \tau'_k t'_k d_k)) \lambda_{\max}(M_k^{-1} P_k P_k^{\mathsf{T}} M_k^{-1}) \left\|P_k g_k\right\|^2.
\end{aligned}$$

By Lemma 2, we have that, with probability at least $1 - 2e^{-s}$, $\left\|P_k P_k^{\mathsf{T}}\right\| \leq \frac{\bar{\mathcal{C}}n}{s}$. In addition, we have $\|H_k\| \leq U_H(\varepsilon)$ from (13), which gives us $\left\|P_k H_k P_k^{\mathsf{T}}\right\| \leq \frac{\bar{\mathcal{C}}n}{s} U_H(\varepsilon)$. For these reasons, we obtain evaluation of the values of $\lambda_{\min}(M_k^{-1})$ and $\lambda_{\max}(M_k^{-1} P_k P_k^{\mathsf{T}} M_k^{-1})$:

$$\begin{aligned}
\lambda_{\min}(M_k^{-1}) &= \frac{1}{\lambda_{\max}(M_k)} \\
&= \frac{1}{\lambda_{\max}(P_k H_k P_k^{\mathsf{T}} + c_1 \Lambda_k I_s + c_2 \left\|g_k\right\|^{\gamma} I_s)} \\
&\geq \frac{1}{\frac{\bar{\mathcal{C}}n}{s} U_H(\varepsilon) + c_1 \frac{\bar{\mathcal{C}}n}{s} U_H(\varepsilon) + c_2 \left\|g_k\right\|^{\gamma}}, \qquad (15)
\end{aligned}$$

$$\begin{aligned}
\lambda_{\max}(M_k^{-1} P_k P_k^{\mathsf{T}} M_k^{-1}) &\leq \left\|P_k P_k^{\mathsf{T}}\right\| \lambda_{\max}(M_k^{-1})^2 \\
&\leq \frac{\bar{\mathcal{C}}n}{s} \frac{1}{\lambda_{\min}(P_k H_k P_k^{\mathsf{T}} + c_1 \Lambda_k I_s + c_2 \left\|g_k\right\|^{\gamma} I_s)^2} \\
&\leq \frac{\bar{\mathcal{C}}n}{s} \frac{1}{c_2^2 \left\|g_k\right\|^{2\gamma}},
\end{aligned}$$

so that we have

$$
\begin{aligned}
&f(x_k) - f(x_k + t'_k d_k) + \alpha t'_k g_k^{\mathsf{T}} d_k \\
&\geq \frac{(1-\alpha)t'_k}{\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_1\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_2\left\|g_k\right\|^{\gamma}} \left\|P_k g_k\right\|^2 - \frac{1}{2}{t'_k}^2 \frac{\bar{\mathcal{C}}n}{s}\frac{U_H(\varepsilon)}{c_2^2\left\|g_k\right\|^{2\gamma}}\left\|P_k g_k\right\|^2 \\
&\geq \frac{(1-\alpha)t'_k}{\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_1\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_2 U_g^{\gamma}} \left\|P_k g_k\right\|^2 - \frac{1}{2}{t'_k}^2 \frac{\bar{\mathcal{C}}n}{s}\frac{U_H(\varepsilon)}{c_2^2\varepsilon^{2\gamma}}\left\|P_k g_k\right\|^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by (9) and } \|g_k\| \geq \varepsilon) \\
&= \frac{\bar{\mathcal{C}}U_H(\varepsilon)n}{2c_2^2\varepsilon^{2\gamma}s}t'_k\left(\frac{2(1-\alpha)c_2^2\varepsilon^{2\gamma}s}{((1+c_1)\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_2U_g^{\gamma})U_H(\varepsilon)\bar{\mathcal{C}}n} - t'_k\right)\left\|P_k g_k\right\|^2 \\
&\geq 0,
\end{aligned}
$$

which completes the proof. ◀

As a consequence of this lemma, it turns out that the step size t_k used in RS-RNM can be bounded from below by some constant.

▶ **Corollary 7.** *Suppose that Assumption 3 holds. Suppose also that there exists $\varepsilon > 0$ such that $\|g_k\| \geq \varepsilon$. Then, with probability at least $1 - 2e^{-s}$, the step size t_k chosen in Line 6 of RS-RNM satisfies*

$$
t_k \geq t_{\min}(\varepsilon), \tag{16}
$$

where

$$
t_{\min}(\varepsilon) = \min\left(1, \frac{2(1-\alpha)\beta c_2^2\varepsilon^{2\gamma}s}{((1+c_1)\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_2U_g^{\gamma})U_H(\varepsilon)\bar{\mathcal{C}}n}\right).
$$

Proof. If

$$
\frac{2(1-\alpha)c_2^2\varepsilon^{2\gamma}s}{((1+c_1)\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_2U_g^{\gamma})U_H(\varepsilon)\bar{\mathcal{C}}n} > 1,
$$

we know that $t_k = 1$ satisfies Armijo's rule (8) from Lemma 6. If not, there exists $l_k \in \{0, 1, 2, \dots\}$ such that

$$
\beta^{l_k+1} < \frac{2(1-\alpha)c_2^2\varepsilon^{2\gamma}s}{((1+c_1)\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_2U_g^{\gamma})U_H(\varepsilon)\bar{\mathcal{C}}n} \leq \beta^{l_k},
$$

and by Lemma 6, we have that the step size β^{l_k+1} satisfies Armijo's rule (8). Then, from the definition of β^{l_k} in Line 6 of RS-RNM, we have

$$
t_k = \beta^{l_k} \geq \beta^{l_k+1} = \beta\cdot\beta^{l_k} \geq \frac{2(1-\alpha)\beta c_2^2\varepsilon^{2\gamma}s}{((1+c_1)\frac{\bar{\mathcal{C}}n}{s}U_H(\varepsilon) + c_2U_g^{\gamma})U_H(\varepsilon)\bar{\mathcal{C}}n}.
$$

This completes the proof. ◀

Using Corollary 7, we can show the global convergence of RS-RNM under Assumption 3.

▶ **Theorem 8.** *Suppose that Assumption 3 holds. Let $\delta \in (0,1)$ and define $\delta_s := 2\left(\exp(-\frac{\mathcal{C}_0}{4}s) + \exp(-s)\right)$ and*

$$
m = \left\lfloor \frac{f(x_0) - f^*}{(1-\delta)(1-\delta_s)p(\varepsilon)\varepsilon^2} \right\rfloor + 1, \quad \textit{where} \quad p(\varepsilon) = \frac{\alpha t_{\min}(\varepsilon)}{2\bar{\mathcal{C}}(1+c_1)\frac{n}{s}U_H(\varepsilon) + 2c_2U_g^{\gamma}}.
$$

Then, with probability at least $1 - \exp\left(-\frac{\delta^2}{2}(1-\delta_s)m\right)$ there exists $k \in \{0, 1, \dots, m-1\}$ such that $\|g_k\| < \varepsilon$.

Proof. We first notice that, by Lemma 1, applied with $\varepsilon = 1/2$, and Lemma 2, we have, using (4), that $\left\|P_k g_k\right\|^2 \geq \frac{1}{2}\left\|g_k\right\|^2$ and $\|P_k P_k^{\top}\| \leq \bar{\mathcal{C}}\frac{n}{s}$ holds for all $k \in \{0, 1, \dots, m-1\}$ with the given probability.

Suppose, for the sake of contradiction, that $\|g_k\| \geq \varepsilon$ for all $k \in \{0, 1, \ldots, m-1\}$. From Armijo's rule (8), we can estimate how much the function value decreases in one iteration. We have that with probability $1 - 2\left(\exp(-\frac{\mathcal{C}_0}{4}s) + \exp(-s)\right)$:

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq -\alpha t_k g_k^\mathsf{T} d_k \\ &= \alpha t_k g_k^\mathsf{T} P_k^\mathsf{T} M_k^{-1} P_k g_k \\ &\geq \alpha t_k \lambda_{\min(M_k^{-1})} \|P_k g_k\|^2 \\ &\geq \frac{\alpha t_{\min}(\varepsilon)}{2(1+c_1)\frac{\bar{\mathcal{C}}n}{s} U_H(\varepsilon) + 2c_2 \|g_k\|^\gamma} \|g_k\|^2 \\ &\qquad\qquad \text{(by } \|P_k g_k\|^2 \geq \tfrac{1}{2}\|g_k\|^2\text{)} \\ &\geq p(\varepsilon)\varepsilon^2. \qquad \text{(by (9) and } \|g_k\| \geq \varepsilon\text{)} \end{aligned}$$

Let us denote by $\mathcal{A}_k$ the event, only depending of P_k, where the above inequality holds. Conditionally to the complement of $\mathcal{A}_k$ we have only that $f(x_k) - f(x_{k+1}) \geq 0$. Let us denote by $T_k \in \{0, 1\}$ the random variable equal to 1 if and only if $\mathcal{A}_k$ holds. Notice that the random variables $\{T_k\}$ are mutually independent because T_k depends only on P_k. By the above remark we have that for all k: $f(x_k) - f(x_{k+1}) \geq p(\varepsilon)\varepsilon^2 T_k$. Hence by adding up all these inequalities from $k = 0$ to $k = m-1$, we get

$$f(x_0) - f(x_m) \geq p(\varepsilon)\varepsilon^2 \sum_{k=0}^{m-1} T_k. \tag{17}$$

Since, for all k, $\mathbb{E}[T_k] \geq 1 - 2\left(\exp(-\frac{\mathcal{C}_0}{4}s) + \exp(-s)\right) := 1 - \delta_s$, we have by a Chernoff bound (see [41]), that for all $\delta \in (0, 1)$,

$$\mathbb{P}\left(\sum_{k=0}^{m-1} T_k \geq (1-\delta)(1-\delta_s)m\right) \geq 1 - \exp\left(-\frac{\delta^2}{2}(1-\delta_s)m\right). \tag{18}$$

Notice that by definition of m, we have that

$$m > \frac{f(x_0) - f^*}{(1-\delta)(1-\delta_s)p(\varepsilon)\varepsilon^2}.$$

Hence

$$(1-\delta)(1-\delta_s)p(\varepsilon)\varepsilon^2 m > f(x_0) - f^*. \tag{19}$$

Thus, with probability at least $1 - \exp\left(-\frac{\delta^2}{2}(1-\delta_s)m\right)$

$$\begin{aligned} f(x_0) - f^* &\geq f(x_0) - f(x_m) \\ &\geq (1-\delta)(1-\delta_s)mp(\varepsilon)\varepsilon^2 \\ &> f(x_0) - f^*, \end{aligned}$$

where the second inequality holds by (17) together with (18) and the strict inequality holds by (19). This is a contradiction, hence there exists $k \in \{0, 1, \ldots, m-1\}$ such that $\|g_k\| < \varepsilon$. ◀

Because of the dependency of $p(\varepsilon)$ on ε, the above discussion can not lead to the iteration complexity analysis, as we need to quantify the exact dependency of the iteration complexity bound with respect to ε. This will be done, under a few additional assumptions, in the next subsection.

4.2 Global iteration complexity

We now estimate the global iteration complexity of the RS-RNM under Assumption 3 and the following assumption.

▶ **Assumption 9.**

i. $\gamma \leq 1/2$,

ii. $\alpha \leq 1/2$,

iii. *There exists* $L_H > 0$ *such that*

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq L_H \left\|x - y\right\|, \quad \forall\, x, y \in \Omega + B(0, r_1),$$

where $r_1 := \dfrac{\bar{\mathcal{C}} U_g^{1-\gamma} n}{c_2 s}$.

From the definition of r_1 in iii, Lemma 4 and (9), we have

$$\|d_k\| \leq \frac{\bar{\mathcal{C}} n}{s} \frac{\|g_k\|^{1-\gamma}}{c_2} \leq \frac{\bar{\mathcal{C}} n}{s} \frac{U_g^{1-\gamma}}{c_2} = r_1.$$

Note that unlike (12), the bound has no dependency on ε. For this reason, we have

$$x_k + \tau d_k \in \Omega + B(0, r_1), \quad \forall\, \tau \in [0, 1].$$

Moreover, since $\Omega + B(0, r_1)$ is bounded and f is twice continuously differentiable, there exists $U_H > 0$ such that

$$\left\|\nabla^2 f(x)\right\| \leq U_H, \quad \forall\, x \in \Omega + B(0, r_1). \tag{20}$$

Similar to the result of Lemma 6, we can show that a step size smaller than some constant satisfies Armijo's rule and therefore, t_k can be bounded from below by some constant.

▶ **Lemma 10.** *Suppose that Assumption 3 and Assumption 9 hold. Then, with probability at least* $1 - 2e^{-s}$*, a step size* $t_k' > 0$ *such that*

$$t_k' \leq \min\left(1, \frac{c_2^2 s^2}{\bar{\mathcal{C}}^2 L_H U_g^{1-2\gamma} n^2}\right),$$

satisfies Armijo's rule, i.e.,

$$f(x_k) - f(x_k + t_k' d_k) \geq -\alpha t_k' g_k^\mathsf{T} d_k.$$

Proof. As (14) is obtained in the proof of Lemma 6, there exists $\tau_k' \in (0, 1)$ such that

$$\begin{aligned}
&f(x_k) - f(x_k + t_k' d_k) + \alpha t_k' g_k^\mathsf{T} d_k \\
&= (1 - \alpha) t_k' g_k^\mathsf{T} P_k^\mathsf{T} M_k^{-1} P_k g_k - \frac{1}{2} {t_k'}^2 g_k^\mathsf{T} P_k^\mathsf{T} M_k^{-1} P_k \nabla^2 f(x_k + \tau_k' t_k' d_k) P_k^\mathsf{T} M_k^{-1} P_k g_k.
\end{aligned}$$

Since we have $1 - \alpha \geq 1/2 \geq t_k'/2$ from Assumption 9.ii, we obtain

$$\begin{aligned}
&f(x_k) - f(x_k + t_k' d_k) + \alpha t_k' g_k^\mathsf{T} d_k \\
&\geq \frac{1}{2} {t_k'}^2 g_k^\mathsf{T} P_k^\mathsf{T} M_k^{-1} P_k g_k - \frac{1}{2} {t_k'}^2 g_k^\mathsf{T} P_k^\mathsf{T} M_k^{-1} P_k \nabla^2 f(x_k + \tau_k' t_k' d_k) P_k^\mathsf{T} M_k^{-1} P_k g_k \\
&= \frac{1}{2} {t_k'}^2 g_k^\mathsf{T} P_k^\mathsf{T} (M_k^{-1} - M_k^{-1} P_k H_k P_k^\mathsf{T} M_k^{-1}) P_k g_k \\
&\quad - \frac{1}{2} {t_k'}^2 g_k^\mathsf{T} P_k^\mathsf{T} M_k^{-1} P_k (\nabla^2 f(x_k + \tau_k' t_k' d_k) - H_k) P_k^\mathsf{T} M_k^{-1} P_k g_k.
\end{aligned} \tag{21}$$

We next evaluate the first and second terms respectively. Since we have

$$\begin{aligned}
M_k^{-1} - M_k^{-1} P_k H_k P_k^\mathsf{T} M_k^{-1} &= M_k^{-1} - M_k^{-1} (M_k - \eta_k I_s) M_k^{-1} \\
&= \eta_k (M_k^{-1})^2,
\end{aligned} \tag{22}$$

the first term can be bounded as follows:

$$\begin{aligned}
\frac{1}{2} {t_k'}^2 g_k^\mathsf{T} P_k^\mathsf{T} (M_k^{-1} - M_k^{-1} P_k H_k P_k^\mathsf{T} M_k^{-1}) P_k g_k &= \frac{1}{2} {t_k'}^2 \eta_k \left\|M_k^{-1} P_k g_k\right\|^2 \\
&\geq \frac{1}{2} {t_k'}^2 c_2 \left\|g_k\right\|^\gamma \left\|M_k^{-1} P_k g_k\right\|^2.
\end{aligned}$$

Using Lemma 2 and Assumption 9.iii, we also obtain, with probability at least $1-2e^{-s}$, the bound of the second term:

$$\begin{aligned}
&\frac{1}{2}{t'_k}^2 g_k^\mathsf{T} P_k^\mathsf{T} M_k^{-1} P_k (\nabla^2 f(x_k + \tau'_k t'_k d_k) - H_k) P_k^\mathsf{T} M_k^{-1} P_k g_k \\
&\leq \frac{1}{2}{t'_k}^2 \left\|\nabla^2 f(x_k + \tau'_k t'_k d_k) - H_k\right\| \left\|P_k P_k^\mathsf{T}\right\| \left\|M_k^{-1} P_k g_k\right\|^2 \\
&\leq \frac{\bar{\mathcal{C}} n}{2s} L_H {t'_k}^3 \left\|d_k\right\| \left\|M_k^{-1} P_k g_k\right\|^2 .
\end{aligned}$$

Thus, we have

$$\begin{aligned}
f(x_k) - f(x_k + t'_k d_k) + \alpha t'_k g_k^\mathsf{T} d_k &\geq \frac{1}{2}{t'_k}^2 \left(c_2 \left\|g_k\right\|^\gamma - \frac{\bar{\mathcal{C}} n}{s} L_H t'_k \left\|d_k\right\|\right) \left\|M_k^{-1} P_k g_k\right\|^2 \\
&= \frac{\bar{\mathcal{C}} n}{2s} L_H {t'_k}^2 \left\|d_k\right\| \left(\frac{c_2 s \left\|g_k\right\|^\gamma}{\bar{\mathcal{C}} L_H n \left\|d_k\right\|} - t'_k\right) \left\|M_k^{-1} P_k g_k\right\|^2 . \qquad (23)
\end{aligned}$$

Moreover, from (9), Lemma 4 and Assumption 9.i, we have

$$\frac{\left\|g_k\right\|^\gamma}{\left\|d_k\right\|} \geq \frac{c_2 s}{\bar{\mathcal{C}} n \left\|g_k\right\|^{1-2\gamma}} \geq \frac{c_2 s}{\bar{\mathcal{C}} U_g^{1-2\gamma} n},$$

so that we finally obtain

$$\begin{aligned}
f(x_k) - f(x_k + t'_k d_k) + \alpha t'_k g_k^\mathsf{T} d_k &\geq \frac{\bar{\mathcal{C}} n}{2s} L_H {t'_k}^2 \left\|d_k\right\| \left(\frac{c_2^2 s^2}{\bar{\mathcal{C}}^2 L_H U_g^{1-2\gamma} n^2} - t'_k\right) \left\|M_k^{-1} P_k g_k\right\|^2 \\
&\geq 0.
\end{aligned}$$

This completes the proof. ◀

▶ **Corollary 11.** *Suppose that Assumption 3 and Assumption 9 hold. Then, with probability at least $1-2e^{-s}$, the step size t_k chosen in Line 6 of RS-RNM satisfies*

$$t_k \geq t_{\min}, \qquad (24)$$

where

$$t_{\min} = \min\left(1, \frac{\beta c_2^2 s^2}{\bar{\mathcal{C}}^2 L_H U_g^{1-2\gamma} n^2}\right).$$

Proof. We get the conclusion in the same way as in the proof of Corollary 7 using Lemma 10. ◀

▶ Remark 12. Since (24) is equivalent to $\beta^{l_k} \geq t_{\min}$, and moreover

$$l_k \leq \log t_{\min} / \log \beta,$$

Corollary 11 tells us that the number of the backtracking steps is bounded above by some constant independent of k.

Now, we can obtain the global iteration complexity of RS-RNM.

▶ **Theorem 13.** *Suppose that Assumption 3 and Assumption 9 hold. Consider any $\delta \in (0,1)$. Let*

$$m = \left\lfloor \frac{f(x_0) - f^*}{(1-\delta)(1-\delta_s) p \varepsilon^2} \right\rfloor + 1, \quad \textit{where} \quad p = \frac{\alpha t_{\min}}{2\bar{\mathcal{C}}(1+c_1)\frac{n}{s} U_H + 2c_2 U_g^\gamma},$$

and where $\delta_s = 2\left(\exp(-\frac{\mathcal{C}_0}{4}s) - \exp(-s)\right)$. Then, we have that

$$\sqrt{\frac{f(x_0) - f^*}{mp}} \geq \min_{k=0,1,\ldots,m-1} \left\|g_k\right\|$$

holds with probability at least $1 - \exp\left(-\frac{\delta^2}{2}(1-\delta_s)m\right)$.

Proof. Replacing $U_H(\varepsilon)$ and $t_{\min(\varepsilon)}$ with U_H, in (20), and $t_{\min}$ respectively in the argument in the proof of Theorem 8, we have

$$f(x_k) - f(x_{k+1}) \geq p \|g_k\|^2 \quad (k = 0, 1, \ldots, m-1),$$

with the given probability. Therefore, by using the same notation as in the proof of Theorem 8, we obtain:

$$\begin{aligned} f(x_0) - f^* &\geq f(x_0) - f(x_m) \\ &= \sum_{k=0}^{m-1} (f(x_k) - f(x_{k+1})) \\ &\geq p \sum_{k=0}^{m-1} \|g_k\|^2 T_k \\ &\geq p \left(\min_{k=0,1,\ldots,m-1} \|g_k\|^2 \right) \sum_{k=0}^{m-1} T_k \\ &\geq (1-\delta)(1-\delta_s) mp \left(\min_{k=0,1,\ldots,m-1} \|g_k\|^2 \right), \end{aligned}$$

where the last inequality holds with probability $1 - \exp\left(-\frac{\delta^2}{2}(1-\delta_s)m\right)$ as shown in (18). This prove the theorem. ◀

If we ignore the probability, Theorem 13 shows that we get $\|g_k\| \leq \varepsilon$ after at most $O(\varepsilon^{-2})$ iterations. This global complexity $O(\varepsilon^{-2})$ is the same as that obtained in [39] for the regularized Newton method. Notice that, by a cubic regularization, the R-ARC algorithm in [37] achieves $O(\varepsilon^{-3/2})$ to obtain a first order stationary point.

5 Local convergence

In this section, we investigate local convergence properties of the sequence $\{x_k\}$ assuming that it converges to a strict local minimizer $\bar{x}$. First we will show that the sequence converges locally linearly to the strict local minimizer; then we will prove that, when f is strongly convex, we cannot aim at local super-linear convergence using random subspace. Finally, we will prove that when the Hessian at $\bar{x}$ is rank deficient then we can attain super-linear convergence for $s < n$ large enough.

▶ **Assumption 14.** *For all x, y*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_H \|x - y\|$$

holds in some neighborhood B_H of $\bar{x}$.

5.1 Local linear convergence

In this subsection we will show that the sequence $\{f(x_k) - f(\bar{x})\}$ converges locally linearly, i.e. there exists $\kappa \in (0,1)$ such that for k large enough,

$$f(x_{k+1}) - f(\bar{x}) \leq (1-\kappa)(f(x_k) - f(\bar{x})).$$

We will further prove that κ can be expressed as $\kappa = O(\frac{s}{n\widetilde{\kappa}(\nabla^2 f(\bar{x}))})$, where $\widetilde{\kappa}(\nabla^2 f(\bar{x}))$ is the ratio of the largest eigenvalue value over the smallest non-zero eigenvalue of $\nabla^2 f(\bar{x})$. Notice that, to the best of our knowledge, until now, local linear convergence is always proved for subspace algorithms assuming that the function is locally strongly convex or satisfies some PL-inequality (3). In this section we prove that under a Hölderian error bound condition, and an additional mild assumptions on the rank of the Hessian at the local minimizer, we can prove local linear convergence. More precisely let us denote by $r = \operatorname{rank}(\nabla^2 f(\bar{x}))$, which measures the number of positive eigenvalues of $\nabla^2 f(\bar{x})$. We will first prove, under some assumption on the rank of the Hessian at $\bar{x}$ and on s, that for any x in the a neighborhood of $\bar{x}$, the function

$$\widetilde{f}_x : u \longmapsto f(x + P^\top u), \quad \text{where } P \text{ is a random matrix sampled from } \mathcal{D} \tag{25}$$

is strongly convex with high probability in a neighborhood of 0. Let us fix $\sigma \in (0,1)$. We recall here that $P \in \mathbb{R}^{s \times n}$ is equal to $\frac{1}{\sqrt{s}}$ times a random Gaussian matrix. In this subsection, we make the following additional assumptions:

▶ **Assumption 15.**

i. *There exists* $\sigma \in (0,1)$ *such that* $r = \operatorname{rank}(\nabla^2 f(\bar{x})) \geq \sigma n$.

ii. *There exist* $\rho \in (0,3)$ *and* $\widetilde{C}$ *such that in a neighborhood of* $\bar{x}$, $f(x_k) - f(\bar{x}) \geq \widetilde{C}\|x_k - \bar{x}\|^\rho$ *holds.*

▶ **Assumption 16.** *We have that* $s \leq \min\left(\frac{\sigma}{4\mathcal{C}^2}, \frac{4(1-\sigma)}{\mathcal{C}^2}\right) n$.

From Assumption 15.i, $\nabla^2 f(\bar{x})$ has r positive eigenvalues, i.e, $\lambda_1(\bar{x}) \geq \dots \lambda_r(\bar{x}) > 0$. By continuity of the eigenvalues, there exists a neighborhood $\overline{B}$ of $\bar{x}$ such that for any $x \in \overline{B}$, $\lambda_r(x) \geq \frac{\lambda_r(\bar{x})}{2}$. Here, we assume, w.l.o.g. that $\overline{B} \subseteq B_H$, where B_H is defined in Assumption 14. Let us denote

$$\bar{\lambda} := \frac{\lambda_r(\bar{x})}{2}. \tag{26}$$

Assumption 15.ii is called a Hölderian growth condition or a Hölderian error bound condition [24]. The condition is weaker than local strong convexity in the sense that it holds with $\rho = 2$ if f is locally strongly convex.

▶ **Proposition 17.** *Assume that Assumption 15.i and Assumption 16 hold. Let us consider* $\widetilde{f}_x$ *defined by* (25). *There exists a neighborhood* $B^* \subseteq \overline{B}$ *such that for any* $x \in B^*$,

$$\nabla^2 \widetilde{f}_x(0) \succeq \frac{n}{8s}\sigma\bar{\lambda} I_s$$

holds with probability at least $1 - 6\exp(-s)$.

Proof. Let $x \in \overline{B}$ be fixed and let $P \in \mathbb{R}^{s\times n}$ be a Gaussian matrix. Because of $\nabla^2 \widetilde{f}_x(0) = P\nabla^2 f(x) P^\top$, we have $u^\top \nabla^2 \widetilde{f}_x(0) u = (P^\top u)^\top \nabla^2 f(x)(P^\top u)$ for any $u \in \mathbb{R}^s$. Let $\nabla^2 f(x) = U(x)D(x)U(x)^\top$ be the eigenvalue decomposition of $\nabla^2 f(x)$. Since $\nabla^2 \widetilde{f}_x(0) = (PU(x))D(x)(PU(x))^\top$ and $PU(x)$ has the same distribution as P, we can assume here w.l.o.g. that $PU(x) = P$. Here

$$D(x) = \begin{pmatrix} \lambda_1(x) & 0 & \dots & 0 \\ 0 & \lambda_2(x) & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n(x) \end{pmatrix},$$

where $\lambda_1(x) \geq \dots \geq \lambda_n(x)$ and $\lambda_r(x) \geq \bar{\lambda}$ (since $x \in \overline{B}$). Let us decompose $P^\top$ such that

$$P^\top = \begin{pmatrix} P^1 \\ P^2 \end{pmatrix}$$

where $P^1 \in \mathbb{R}^{n_1 \times s}$ and $P^2 \in \mathbb{R}^{n_2 \times s}$, where n_1 and n_2 are chosen such that $n_1 = r$ and $n_2 = n - r$. Furthermore let $D_1(x)$ and $D_2(x)$ be respectively the $n_1 \times n_1$ and $n_2 \times n_2$ diagonal matrix such that $D(x) = \begin{pmatrix} D_1(x) & 0 \\ 0 & D_2(x) \end{pmatrix}$. We have

$$(P^\top u)^\top D(x)(P^\top u) = (P^1 u)^\top D_1(x)(P^1 u) + (P^2 u)^\top D_2(x)(P^2 u). \tag{27}$$

By Assumption 15.i, and by definition of $\overline{B}$, we have that $D_1(x) \succeq \lambda_r(x) I_{n_1} \succeq \bar{\lambda} I_{n_1} \succ 0$, and $D_2(x) \succeq \lambda_n(x) I_{n_2}$. Hence from (27), we have

$$(P^\top u)^\top D(x)(P^\top u) \geq \bar{\lambda}\|P^1 u\|^2 + \lambda_n(x)\|P^2 u\|^2. \tag{28}$$

Let $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ denote respectively the largest and the smallest singular value of a matrix. Using [41, Theorem 4.6.1], there exists a constant $\mathcal{C}$ such that with probability at least $1 - 6\exp(-s)$:

$$\begin{aligned} &\sqrt{\frac{n}{s}} - \mathcal{C} \leq \sigma_{\min}(P^\top) \leq \sigma_{\max}(P^\top) \leq \sqrt{\frac{n}{s}} + \mathcal{C}, \\ &\sqrt{\frac{n_1}{s}} - \mathcal{C} \leq \sigma_{\min}(P^1) \leq \sigma_{\max}(P^1) \leq \sqrt{\frac{n_1}{s}} + \mathcal{C}, \\ &\sqrt{\frac{n_2}{s}} - \mathcal{C} \leq \sigma_{\min}(P^2) \leq \sigma_{\max}(P^2) \leq \sqrt{\frac{n_2}{s}} + \mathcal{C}. \end{aligned} \tag{29}$$

More precisely, since all the three matrices $P^\top, P^1$ and P^2 are Gaussian random matrices, we can apply [41, Theorem 4.6.1] and deduce that each of the three inequalities above holds with probability $1-2\exp(-s)$. The probability that all the three equations hold is derived using (4). Hence, with probability at least $1-6e^{-s}$, for any $u \in \mathbb{R}^s$,

$$\|P^1 u\| \geq \sqrt{n/s}\left(\frac{\sqrt{\frac{n_1}{s}}-\mathcal{C}}{\sqrt{n/s}}\right)\|u\|,$$
$$\|P^2 u\| \leq \sqrt{n/s}\left(\frac{\sqrt{\frac{n_2}{s}}+\mathcal{C}}{\sqrt{n/s}}\right)\|u\|.$$

We have that $n_1 \geq \sigma n$ and $n_2 \leq (1-\sigma)n$. Furthermore, we have by Assumption 16 that $s \leq \frac{\sigma}{4\mathcal{C}^2}n$ implies that $\sqrt{\frac{\sigma n}{s}} - \mathcal{C} \geq \frac{1}{2}\sqrt{\frac{\sigma n}{s}}$ and $s \leq \frac{4(1-\sigma)}{4\mathcal{C}^2}n$ implies that $\sqrt{\frac{(1-\sigma)n}{s}} + \mathcal{C} \leq 2\sqrt{\frac{(1-\sigma)n}{s}}$. Hence

$$\frac{\sqrt{\frac{n_1}{s}}-\mathcal{C}}{\sqrt{n/s}} \geq \frac{1}{2}\sqrt{\sigma} \quad \& \quad \frac{\sqrt{\frac{n_2}{s}}+\mathcal{C}}{\sqrt{n/s}} \leq 2\sqrt{(1-\sigma)}.$$

Therefore,

$$\|P^1 u\| \geq \frac{1}{2}\sqrt{\sigma(n/s)}\|u\|,$$
$$\|P^2 u\| \leq 2\sqrt{(1-\sigma)(n/s)}\|u\|.$$

Hence, from (28), we have that

$$(P^\top u)^\top D(x)(P^\top u) \geq n/s\left(\frac{1}{4}\sigma\bar{\lambda} + 4(1-\sigma)\min(\lambda_n(x),0)\right)\|u\|^2.$$

We conclude the proposition by noticing that $\min(\lambda_n(x),0)$ tends to 0, hence the claim holds by considering a neighborhood $B^* \subseteq \overline{B}$ of $\bar{x}$ small enough. ◀

We deduce the following PL inequality for $\widetilde{f}_x$ when $x \in B^*$.

▶ **Proposition 18.** *Assume that Assumption 14, Assumption 15.i and Assumption 16 hold, and let $P \in \mathbb{R}^{s\times n}$ be a Gaussian matrix. There exist neighborhoods $\widehat{B} \subset B^*$ and B_0 (a neighborhood of $0 \in \mathbb{R}^s$) such that for any $x \in \widehat{B}$,*

$$\nabla\widetilde{f}_x(0)^\top(P\nabla^2 f(x)P^\top)^{-1}\nabla\widetilde{f}_x(0) \geq f(x) - \min_{u\in B_0} f(x+P^\top u)$$

holds with probability at least $1-6\exp(-s)$.

Proof. Let $\widehat{B} \subset B^*$, and let $x \in \widehat{B}$. By the Taylor expansion of $\widetilde{f}_x$ at 0, there exists $\widetilde{x} \in [x, x+P^\top u]$ such that

$$f(x+P^\top u) = f(x) + (P\nabla f(x))^\top u + \frac{1}{2}u^\top P\nabla^2 f(\widetilde{x})P^\top u.$$

Since, by Proposition 17, we have that $P\nabla^2 f(\widetilde{x})P^\top \succ 0$ for any $x + P^\top u \in B^*$, we deduce by Assumption 14 that for u small enough:

$$f(x+P^\top u) \geq f(x) + (P\nabla f(x))^\top u + \frac{1}{4}u^\top P\nabla^2 f(x)P^\top u. \tag{30}$$

Let B_0 be a neighborhood of $0 \in \mathbb{R}^s$ such that, (30) holds, and $x + P^\top u \in B^*$ for any $x \in \widehat{B}$. Let $g(u) = (P\nabla f(x))^\top u + \frac{1}{4}u^\top P\nabla^2 f(x)P^\top u$. By the above inequality we have that

$$\min_{u\in B_0} f(x+P^\top u) \geq f(x) + \min_{u\in B_0} g(u). \tag{31}$$

By Proposition 17 we know that for any $u \in \mathbb{R}^s$ such that $x + P^\top u \in B^*$, g is convex. Thus, the minimum is attained at the point u^* satisfying

$$\nabla g(u^*) = P\nabla f(x) + \frac{1}{2}P\nabla^2 f(x)P^\top u^* = 0.$$

Hence, since $\|\nabla f(x)\|$ tends to 0 as x tends to $\bar{x}$, we can ensure, by taking $\widehat{B}$ small enough, that $u^* \in B_0$. Hence

$$\begin{aligned}\min_{u\in B_0} g(u) &= -2(P\nabla f(x))^\top (P\nabla^2 f(x)P^\top)^{-1}P\nabla f(x) + \frac{1}{4}4(P\nabla f(x))^\top (P\nabla^2 f(x)P^\top)^{-1}P\nabla f(x)\\ &= -(P\nabla f(x))^\top (P\nabla^2 f(x)P^\top)^{-1}P\nabla f(x)\end{aligned}$$

holds and (31) yields the desired inequality. ◀

Before proving local linear convergence, we prove the following technical proposition.

▶ **Proposition 19.** *Assume that Assumption 14, Assumption 16, and Assumption 15 hold. There exists $k_0 \in \mathbb{N}$ such that if $k \geq k_0$, we have with probability $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4}s))$:*

$$f(x_k) - \min_{u\in B_0} \widetilde{f}_{x_k}(u) \geq \frac{\lambda_0}{4\lambda_{\max}(\overline{H})\left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2}(f(x_k) - f(\bar{x})),$$

where λ_0 is the minimal non-zero eigenvalue of $\overline{H} := \nabla^2 f(\bar{x})$.

Proof. Using a Taylor expansion around $\bar{x}$, we have that for all $y \in \widehat{B}$,

$$|f(y) - f(\bar{x}) - \frac{1}{2}(y-\bar{x})^\top \overline{H}(y-\bar{x})| \leq L_H\|y-\bar{x}\|^3, \tag{32}$$

where we define

$$\overline{H} := \nabla^2 f(\bar{x}). \tag{33}$$

Also, for $u \in \mathbb{R}^d$ small enough, we have by setting $y = x_k + P_k^\top u$ in (32), that for k large enough such that $x_k + P_k^\top u \in \widehat{B}$,

$$\begin{aligned}&|f(x_k + P_k^\top u) - f(\bar{x}) - \frac{1}{2}(x_k-\bar{x})^\top \overline{H}(x_k-\bar{x}) - \frac{1}{2}u^\top P_k\overline{H}P_k^\top u - (P_k\overline{H}(x_k-\bar{x}))^\top u| \\ &\leq L_H\|x_k - \bar{x} + P_k^\top u\|^3\end{aligned} \tag{34}$$

holds. Let $g(u) = \frac{1}{2}u^\top P_k\overline{H}P_k^\top u + (P_k\overline{H}(x_k-\bar{x}))^\top u$. By a reasoning similar to that of Proposition 17, g is strongly convex with probability $1 - 6e^{-s}$ and hence is minimized at

$$u^* = -(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}(x_k-\bar{x}). \tag{35}$$

Notice that as k tends to infinity $\|u^*\|$ tends to 0, hence for k large enough we have $x_k + P_k^\top u^* \in \widehat{B}$ and $u^* \in B_0$. Plugging (35) in (34) yields

$$\begin{aligned}&f(x_k + P_k^\top u^*)\\ &\leq f(\bar{x}) + \frac{1}{2}(x_k-\bar{x})^\top \overline{H}(x_k-\bar{x}) - \frac{1}{2}(x_k-\bar{x})^\top \overline{H}P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}(x_k-\bar{x}) + L_H\|x_k-\bar{x}+P_k^\top u^*\|^3,\end{aligned}$$

from which we deduce

$$\begin{aligned}&f(x_k) - f(x_k + P_k^\top u^*)\\ &\geq f(x_k) - f(\bar{x}) - \frac{1}{2}(x_k-\bar{x})^\top \overline{H}(x_k-\bar{x}) + \frac{1}{2}(x_k-\bar{x})^\top \Pi(x_k-\bar{x}) - L_H\|x_k-\bar{x}+P_k^\top u^*\|^3,\end{aligned}$$

where $\Pi = \overline{H}P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}$. Using (32), we further obtain

$$f(x_k) - f(x_k + P_k^\top u^*) \geq \frac{1}{2}(x_k-\bar{x})^\top \Pi(x_k-\bar{x}) - L_H(\|x_k-\bar{x}+P_k^\top u^*\|^3 + \|x_k-\bar{x}\|^3). \tag{36}$$

We have $(x_k-\bar{x})^\top\Pi(x_k-\bar{x}) = (\overline{H}^{1/2}(x_k-\bar{x}))^\top\overline{\Pi}(\overline{H}^{1/2}(x_k-\bar{x}))$, where $\overline{\Pi} := \overline{H}^{1/2}P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}^{1/2}$ is an orthogonal projection matrix into $\mathrm{Range}(\overline{H}^{1/2}P_k^\top)$ parallel to $\ker P_k\overline{H}^{1/2}$. Hence

$$(x_k-\bar{x})^\top\Pi(x_k-\bar{x}) = \|\overline{\Pi}\overline{H}^{1/2}(x_k-\bar{x})\|^2.$$

Since $\|P_k\overline{H}^{1/2}\|^2\|\overline{\Pi}\overline{H}^{1/2}(x_k-\bar{x})\|^2 \geq \|P_k\overline{H}^{1/2}\overline{\Pi}\overline{H}^{1/2}(x_k-\bar{x})\|^2$, we have

$$\begin{aligned}(x_k-\bar{x})^\top\Pi(x_k-\bar{x}) &\geq \frac{1}{\|P_k\overline{H}^{1/2}\|^2}\|P_k\overline{H}^{1/2}\overline{\Pi}\overline{H}^{1/2}(x_k-\bar{x})\|^2\\ &= \frac{1}{\|P_k\overline{H}^{1/2}\|^2}\|P_k\overline{H}(x_k-\bar{x})\|^2\\ &\geq \frac{1}{2\|P_k\overline{H}^{1/2}\|^2}\|\overline{H}(x_k-\bar{x})\|^2\\ &\geq \frac{\lambda_0}{2\|P_k\overline{H}^{1/2}\|^2}\|\overline{H}^{1/2}(x_k-\bar{x})\|^2\\ &= \frac{\lambda_0}{2\lambda_{\max}(P_k\overline{H}P_k)}\|\overline{H}^{1/2}(x_k-\bar{x})\|^2\\ &= \frac{\lambda_0}{2\lambda_{\max}(P_k\overline{H}P_k)}(x_k-\bar{x})^\top\overline{H}(x_k-\bar{x}). \end{aligned} \tag{37}$$

where the second inequality holds with probability at least $1-2\exp(-\frac{c_0}{4}s)$ (by Lemma 1 with $\varepsilon=\frac{1}{2}$), and the third holds as λ_0 is the smallest non-zero eigenvalue of $\overline{H}$. The second equality holds as $\sigma_{\max}(P_k\overline{H}^{1/2})^2 = \lambda_{\max}(P_k\overline{H}_kP_k)$. We have therefore proved that

$$(\overline{H}^{1/2}(x_k-\bar{x}))^\top\overline{\Pi}(\overline{H}^{1/2}(x_k-\bar{x})) \geq \frac{\lambda_0}{2\lambda_{\max}(P_k\overline{H}P_k)}(x_k-\bar{x})^\top\overline{H}(x_k-\bar{x}). \tag{38}$$

Hence, by (36), we have

$$\begin{aligned}&f(x_k)-f(x_k+P_k^\top u^*)\\ &\geq \frac{\lambda_0}{4\lambda_{\max}(P_k\overline{H}P_k)}(x_k-\bar{x})^\top\overline{H}(x_k-\bar{x}) - L_H(\|x_k-\bar{x}+P_k^\top u^*\|^3+\|x_k-\bar{x}\|^3).\end{aligned} \tag{39}$$

From (35), we have that $\|x_k-\bar{x}+P_k^\top u^*\| = \|(I_n-P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H})(x_k-\bar{x})\|$. Hence

$$\|x_k-\bar{x}+P_k^\top u^*\| \leq \|I_n-P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}\|\|x_k-\bar{x}\|. \tag{40}$$

Since $P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}$ is projection matrix (along $\mathrm{Im}(P_k^\top)$ parallel to $\mathrm{Ker}(P_kH)$), we have by [1] that

$$\|I_n-P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}\| = \|P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}\|. \tag{41}$$

Furthermore, by Proposition 17, we have that with probability at least $1-6\exp(-s)$,

$$P_k\overline{H}P_k^\top \succeq \frac{n}{8s}\sigma\bar{\lambda}I_s.$$

Hence, we deduce from (41) that

$$\|I_n-P_k^\top(P_k\overline{H}P_k^\top)^{-1}P_k\overline{H}\| \leq \frac{\|P_k^\top\|^2\|\overline{H}\|}{\frac{n}{8s}\sigma\bar{\lambda}}. \tag{42}$$

Therefore, we deduce by (39), (40) and (42) for $\beta_1>0$ suitably chosen, we have

$$f(x_k)-f(x_k+P_k^\top u^*) \geq \frac{\lambda_0}{4\lambda_{\max}(P_k\overline{H}P_k)}(x_k-\bar{x})^\top\overline{H}(x_k-\bar{x}) - \beta_1\|x_k-\bar{x}\|^3. \tag{43}$$

By taking $y=x_k$ in (32), we have that

$$\frac{1}{2}(x_k-\bar{x})^\top\overline{H}(x_k-\bar{x}) \geq f(x_k)-f(\bar{x})-L_H\|x_k-\bar{x}\|^3.$$

Hence, by (43)

$$f(x_k)-f(x_k+P_k^\top u^*) \geq \frac{\lambda_0}{2\lambda_{\max}(P_k\overline{H}P_k)}(f(x_k)-f(\bar{x})) - \left(\frac{\lambda_0}{2\lambda_{\max}(P_k\overline{H}P_k)}L_H+\beta_1\right)\|x_k-\bar{x}\|^3.$$

By Assumption 15.ii,

$$f(x_k) - f(x_k + P_k^\top u^*) \geq \left(\frac{\lambda_0}{2\lambda_{\max}(P_k \overline{H} P_k)} - \left(\frac{\lambda_0}{2\lambda_{\max}(P_k \overline{H} P_k)} L_H + \beta_1 \right) \frac{1}{\widetilde{C}} \|x_k - \bar{x}\|^{3-\rho} \right) (f(x_k) - f(\bar{x})).$$

Since $\|x_k - \bar{x}\|$ tends to 0 as k tends to infinity and $\rho < 3$, we have that for k large enough

$$f(x_k) - \min_{u \in B_0} f(x_k + P_k^\top u) \geq f(x_k) - f(x_k + P_k^\top u^*) \geq \frac{\lambda_0}{4\lambda_{\max}(P_k \overline{H} P_k)} (f(x_k) - f(\bar{x})),$$

where the first inequality holds as, by (35), $u^* \in B_0$ for k large enough. The probability bound in the statement of the theorem is obtained by using (4): in the whole proof we only use Lemma 1 with $\varepsilon = \frac{1}{2}$, which holds with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4} s)$, and the inequalities (29) which hold with probability at least $1 - 6\exp(-s)$. We also factorize the expression, using that $1 - 2\exp(-\frac{\mathcal{C}_0}{4} s) > 1 - 6\exp(-\frac{\mathcal{C}_0}{4} s)$. We end the proof by noticing that $\lambda_{\max}(P_k \overline{H} P_k) \leq \lambda_{\max}(\overline{H}) \sigma_{\max}(P_k)^2$, hence by the first equation of (29)

$$\lambda_{\max}(P_k \overline{H} P_k) \leq \lambda_{\max}(\overline{H}) \left(\sqrt{\frac{n}{s}} + \mathcal{C} \right)^2. \tag{44}$$

◀

We are now ready to prove the main theorem of this section.

▶ **Theorem 20.** *Assume that Assumption 14, Assumption 15 and Assumption 16 hold. There exist* $0 < \kappa < 1$, $k_0 \in \mathbb{N}$, *such that if* $k \geq k_0$, *then*

$$f(x_{k+1}) - f(\bar{x}) \leq \left(1 - \frac{1}{2} \alpha (1 - \alpha) \frac{\lambda_0}{4\lambda_{\max}(\overline{H}) \left(\sqrt{\frac{n}{s}} + \mathcal{C} \right)^2} \right) (f(x_k) - f(\bar{x}))$$

holds with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4} s))$. *Here* $\alpha \in (0, 1)$ *is a parameter of Algorithm 1.*

Proof. We recall that we use a backtracking line search to find at each iteration k a step-size t_k such that

$$f(x_k + t_k d_k) \leq f(x_k) + \alpha t_k \nabla f(x_k)^\top d_k,$$

with $d_k = P_k^\top u_k$ and the update rule $t_k \leftarrow \beta t_k$ for $0 < \alpha < 1$ and $0 < \beta < 1$. We recall that

$$u_k = -(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k g_k, \tag{45}$$

where we recall that $\eta_k = c_1 \Lambda_k + c_2 \|g_k\|^\gamma$. By a Taylor expansion of f around x_k, there exists $x_k^* \in [x_k, x_{k+1}]$ such that

$$f(x_k + t_k P_k^\top u_k) = f(x_k) + t_k (P_k g_k)^\top u_k + \frac{t_k^2}{2} u_k^\top P_k \nabla^2 f(x_k^*) P_k^\top u_k. \tag{46}$$

Notice that $\nabla^2 f$ is Lipschitz continuous (by Assumption 14). Furthermore, by Proposition 17, for k large enough, $P_k H_k P_k^\top$ is positive definite with probability at least $1 - 6\exp(-s)$ as the sequence $\{x_k\}$ converges to $\bar{x}$. Hence, for k large enough

$$\begin{aligned} u_k^\top P_k \nabla^2 f(x_k^*) P_k^\top u_k &\leq u_k^\top P_k H_k P_k^\top u_k + \|P_k^\top u_k\|^2 \|H_k - \nabla^2 f(x_k^*)\| \\ &\leq u_k^\top P_k H_k P_k^\top u_k + L_H \|P_k^\top u_k\|^2 \|x_k - x_{k+1}\| \leq 2 u_k^\top P_k H_k P_k^\top u_k \end{aligned}$$

holds with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4} s))$. By (46), we deduce that for k large enough:

$$\begin{aligned} f(x_k + t_k P_k^\top u_k) &\leq f(x_k) + t_k (P_k g_k)^\top u_k + 2 \frac{t_k^2}{2} u_k^\top P_k H_k P_k^\top u_k \\ &\leq f(x_k) + t_k (P_k g_k)^\top u_k + t_k^2 u_k^\top (P_k H_k P_k^\top + \eta_k I_s) u_k, \end{aligned}$$

where the second inequality holds as $\eta_k \geq 0$. Let

$$\mu_k^2 := -g_k^\top d_k = (P_k g_k)^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1} (P_k g_k). \tag{47}$$

Since $(P_k g_k)^\top u_k = g_k^\top (P_k^\top u_k) = -\mu_k^2$, and by definition of u_k in (45), we can write

$$f(x_k + t_k P_k^\top u_k) \leq f(x_k) - t_k \mu_k^2 + t_k^2 u_k^\top (P_k H_k P_k^\top + \eta_k I_s) u_k = f(x_k) - t_k \mu_k^2 + t_k^2 \mu_k^2. \tag{48}$$

Hence, we have

$$f(x_{k+1}) \leq f(x_k) - t_k \left(1 - t_k\right) \mu_k^2.$$

Thus the step-size $t_k = 1 - \alpha$ satisfies the exit condition, $f(x_k) - f(x_k + t_k d_k) \geq -\alpha t_k g_k^\mathsf{T} d_k$, in the backtracking line search as we have

$$(1 - t_k) = \alpha$$

for such t_k. Therefore, the backtracking line search stops with some $t_k \geq 1 - \alpha$, and we have

$$f(x_{k+1}) \leq f(x_k) - \alpha(1 - \alpha)\mu_k^2. \tag{49}$$

Notice that since η_k tends to 0, we have that

$$\mu_k^2 = (P_k g_k)^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1} (P_k g_k) \geq \frac{1}{2} (P_k g_k)^\top (P_k \overline{H} P_k^\top)^{-1} (P_k g_k).$$

Hence, by Proposition 18, we have that when k is large enough,

$$f(x_{k+1}) - f(\bar{x}) \leq f(x_k) - f(\bar{x}) - \frac{1}{2}\alpha(1 - \alpha) \left(f(x_k) - \min_{u \in B_0} \widetilde{f}_{x_k}(u) \right) \tag{50}$$

holds with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4} s))$. By Proposition 19, we have that

$$f(x_k) - \min_{u \in B_0} \widetilde{f}_{x_k}(u) \geq \frac{\lambda_0}{4\lambda_{\max}(\overline{H}) \left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2} (f(x_k) - f(\bar{x}))$$

holds with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4} s))$. Hence

$$f(x_{k+1}) - f(\bar{x}) \leq \left(1 - \frac{1}{2}\alpha(1 - \alpha) \frac{\lambda_0}{4\lambda_{\max}(\overline{H}) \left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2} \right) \left(f(x_k) - f(\bar{x}) \right), \tag{51}$$

which proves the theorem. ◀

▶ Remark 21. Notice that the rate we obtain corresponds to a high probability estimation of the local convergence rate derived, when f is assumed to be strongly convex, in the stochastic subspace cubic Newton method [22]. This can be seen in the proof of Proposition 19, where the rate $\frac{\lambda_0}{4\lambda_{\max}(\overline{H})\left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2}$ corresponds to a lower bound of $\lambda_{\min}(\overline{H}^{1/2} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \overline{H}^{1/2})$, as seen in (38) and (44). More specifically, this corresponds to a high probability lower bound of the parameter $\zeta = \lambda_{\min}[\mathbb{E}(\overline{\Pi})] = \lambda_{\min}[\mathbb{E}(\overline{H}^{1/2} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \overline{H}^{1/2})]$ that appears in the local convergence rate in Theorem 6.2 of [22].

Let us define

$$\kappa := \frac{1}{2}\alpha(1 - \alpha) \frac{\lambda_0}{4\lambda_{\max}(\overline{H}) \left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2} < 1.$$

We have the following direct corollary:

▶ **Corollary 22.** *Assume that Assumption 14, Assumption 15 and Assumption 16 hold. There exist $k_0 \in \mathbb{N}$ such that if $k \geq k_0$, then, for any $m \in \mathbb{N}$,*

$$f(x_{k+m}) - f(\bar{x}) \leq (1 - \kappa)^m (f(x_k) - f(\bar{x}))$$

holds with probability at least $1 - 6m(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4} s))$.

Proof. This is a direct consequence of Theorem 20 where the success probability is obtained by union bound, using (4). ◀

Notice that one can also derive an expectation version of Theorem 20 as follows.

▶ **Corollary 23.** *Assume that Assumption 14, Assumption 15 and Assumption 16 hold. There exist $k_0 \in \mathbb{N}$ such that if $k \geq k_0$, then,*

$$\mathbb{E}\left[f(x_{k+1}) - f(\bar{x})\right] \leq (1 - p^2\kappa)\mathbb{E}\left[f(x_k) - f(\bar{x})\right],$$

where $p := 1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4}s))$. Here the expectation is taken with respect to the random variables $P_0, P_1, P_2, \ldots, P_k$.

Proof. By (50) we have that

$$f(x_{k+1}) - f(\bar{x}) \leq f(x_k) - f(\bar{x}) - \frac{1}{2}\alpha(1-\alpha)\left(f(x_k) - \min_{u \in B_0} \widetilde{f}_{x_k}(u)\right)$$

holds with probability $p = 1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4}s))$. Let us denotes by $\mathcal{E}$ the event, with respect to P_k, on which the above equation holds. Since $f(x_{k+1}) - f(\bar{x}) \leq f(x_k) - f(\bar{x})$ holds with probability one, we can write that

$$f(x_{k+1}) - f(\bar{x}) \leq f(x_k) - f(\bar{x}) - \frac{1}{2}\alpha(1-\alpha)\left(f(x_k) - \min_{u \in B_0} \widetilde{f}_{x_k}(u)\right)\mathbf{1}_{\mathcal{E}},$$

where $\mathbf{1}_{\mathcal{E}}$ is the indicator function over $\mathcal{E}$. Let us consider the following conditional expectation: $\mathbb{E}\left[\cdot \mid P_0, \ldots, P_{k-1}\right]$. We have that

$$\begin{aligned}
&\mathbb{E}\left[f(x_{k+1}) - f(\bar{x}) \mid P_0, \ldots, P_{k-1}\right] \\
&\leq f(x_k) - f(\bar{x}) - \frac{1}{2}\alpha(1-\alpha)\mathbb{E}\left[\left(f(x_k) - \min_{u \in B_0} \widetilde{f}_{x_k}(u)\right)\mathbf{1}_{\mathcal{E}} \;\middle|\; P_0, \ldots, P_{k-1}\right] \qquad (52)
\end{aligned}$$

holds as $f(x_k) - f(\bar{x})$ is measurable with respect to the sigma algebra generated by $P_1, \ldots, P_{k-1}$. Let us define the event

$$\mathcal{E}' = \left\{ f(x_k) - \min_{u \in B_0} \widetilde{f}_{x_k}(u) \geq \frac{\lambda_0}{4\lambda_{\max}(\overline{H})\left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2}(f(x_k) - f(\bar{x})) \;\middle|\; x_k \right\},$$

on this sigma algebra, which holds by probability at least $p = 1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4}s))$, by Proposition 19. By conditioning the right-hand-side of (52) with respect to this event, we obtain that when k is large enough

$$\begin{aligned}
&\mathbb{E}\left[f(x_{k+1}) - f(\bar{x}) \mid P_0, \ldots, P_{k-1}\right] \\
&\leq f(x_k) - f(\bar{x}) - \frac{1}{2}\alpha(1-\alpha)\mathbb{E}\left[\frac{\lambda_0}{4\lambda_{\max}(\overline{H})\left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2}(f(x_k) - f(\bar{x}))\mathbf{1}_{\mathcal{E}}\right] p \\
&\leq (f(x_k) - f(\bar{x}))\left(1 - \frac{1}{2}\alpha(1-\alpha)\frac{\lambda_0}{4\lambda_{\max}(\overline{H})\left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2}p^2\right).
\end{aligned}$$

Where the first inequality holds as in any case we have that $f(x_k) - \min_{u \in B_0} \widetilde{f}_{x_k}(u) \geq 0$. By taking the expectation with respect to $P_0, \ldots, P_{k-1}$ we deduce the corollary. ◀

Let consider the following assumption.

▶ **Assumption 24.** *There exists $\rho > 0$ such that for k large enough*

$$\|\nabla f(x_k)\| \geq \rho\|x_k - \bar{x}\|. \qquad (53)$$

Notice that Assumption 24 is actually stronger than Assumption 15.ii.

▶ **Lemma 25.** *We have, under Assumption 14 and Assumption 24, that for k large enough:*

$$\frac{\rho}{2\sqrt{\lambda_{\max}(\overline{H})}}\|x_k - \bar{x}\| \leq \|\sqrt{\overline{H}}(x_k - \bar{x})\|.$$

Proof. Using a Taylor expansion of $t \mapsto \nabla f(\bar{x} + t(x_k - \bar{x}))$ around 0, we have that

$$\nabla f(x_k) = \nabla f(\bar{x}) + \int_0^1 \nabla^2 f(\bar{x} + t(x_k - \bar{x}))(x_k - \bar{x})\mathrm{d}t = \int_0^1 \nabla^2 f(\bar{x} + t(x_k - \bar{x}))(x_k - \bar{x})\mathrm{d}t. \tag{54}$$

By Assumption 14, for any $t \in [0,1]$ we have $\|\nabla^2 f(\bar{x} + t(x_k - \bar{x})) - \overline{H}\| \leq t L_H \|x_k - \bar{x}\|$. Hence we deduce that

$$\|\nabla f(x_k)\| \leq \|\overline{H}(x_k - \bar{x})\| + \|\nabla^2 f(\bar{x} + t(x_k - \bar{x})) - \overline{H}\|\|x_k - \bar{x}\| \leq \|\overline{H}(x_k - \bar{x})\| + L_H \|x_k - \bar{x}\|^2. \tag{55}$$

Therefore, by (53), we deduce that

$$\rho\|x_k - \bar{x}\| - L_H\|x_k - \bar{x}\|^2 \leq \|\nabla f(x_k)\| - L_H\|x_k - \bar{x}\|^2 \overset{(55)}{\leq} \|\overline{H}(x_k - \bar{x})\|. \tag{56}$$

Since $\|x_k - \bar{x}\|$ tends to 0, we deduce that for k large enough:

$$\frac{\rho}{2}\|x_k - \bar{x}\| \leq \|\overline{H}(x_k - \bar{x})\| \leq \sqrt{\lambda_{\max}(\overline{H})}\|\sqrt{\overline{H}}(x_k - \bar{x})\|. \qquad \blacktriangleleft$$

Let us now define the semi-norm:

$$\|x\|_{\overline{H}}^2 := x^\top \overline{H} x. \tag{57}$$

Notice that by Lemma 25, under Assumption 24, when k is large enough, $\|\cdot\|_{\overline{H}}$ is a norm for $x_k - \bar{x}$ as we have that $\|x_k - \bar{x}\|_{\overline{H}} = 0$ if and only if $\|x_k - \bar{x}\| = 0$.

▶ **Proposition 26.** *Assume that Assumption 14, Assumption 16, Assumption 15.i and Assumption 24 hold. Then for k large enough:*

$$\|x_{k+1} - \bar{x}\|_{\overline{H}} \leq \left(\sqrt{1 - \frac{\lambda_0}{4\lambda_{\max}(\overline{H})\left(\sqrt{\frac{n}{s}} + \mathcal{C}\right)^2}}\right)\|x_k - \bar{x}\|_{\overline{H}}$$

holds with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4}s))$.

Proof.

$$\begin{aligned}
&\sqrt{\overline{H}}(x_{k+1} - \bar{x}) = \sqrt{\overline{H}}(x_{k+1} - x_k) + \sqrt{\overline{H}}(x_k - \bar{x}) \\
&= -\sqrt{\overline{H}}P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k g_k + \sqrt{\overline{H}}(x_k - \bar{x}) \\
&= -\sqrt{\overline{H}}P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k H_k(x_k - \bar{x}) + \sqrt{\overline{H}}P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k(g_k - H_k(x_k - \bar{x})) \qquad (58) \\
&\quad + \sqrt{\overline{H}}(x_k - \bar{x}) \\
&= -A + B + \sqrt{\overline{H}}(x_k - \bar{x}), \qquad (59)
\end{aligned}$$

where $A := \sqrt{\overline{H}}P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k H_k(x_k - \bar{x})$ and $B := \sqrt{\overline{H}}P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k(g_k - H_k(x_k - \bar{x}))$. First let us bound B. In order to do so, we bound $\|P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k\|$. Notice that from $P_k H_k P_k^\top \succ 0$, $\eta_k \geq 0$ and Proposition 17, we have

$$\|P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k\| \leq \|P_k^\top(P_k H_k P_k^\top)^{-1} P_k\| \leq \frac{\|P_k^\top\|^2}{\frac{n}{8s}\sigma\overline{\lambda}} \tag{60}$$

with probability at least $1 - 6\exp(-s)$. Therefore, by Lemma 2, we have

$$\|P_k^\top(P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k\| \leq \frac{8\overline{\mathcal{C}}}{\sigma\overline{\lambda}}. \tag{61}$$

By Taylor expansion at $\bar{x}$ of ∇f, as in (54), and by subtracting $H_k(x_k - \bar{x})$ to both sides, we obtain by Assumption 14 that

$$\|g_k - H_k(x_k - \bar{x})\| \leq \int_0^1 \|\nabla^2 f(\bar{x} + t(x_k - \bar{x})) - \nabla^2 f(\bar{x})\|\|x_k - \bar{x}\|\mathrm{d}t = O(\|x_k - \bar{x}\|^2). \tag{62}$$

Hence, by (61) and (62), there exists a constant $\beta_1 > 0$ such that

$$B \leq \|\sqrt{\overline{H}}\|\|P_k^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k\|\|(g_k - H_k(x_k - \bar{x}))\| \leq \beta_1 \|x_k - \bar{x}\|^2. \tag{63}$$

Let us now bound $A = \sqrt{\overline{H}} P_k^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k H_k (x_k - \bar{x})$. Let us furthermore decompose $A = A_1 + A_2$ such that

$$\begin{aligned}
&\sqrt{\overline{H}} P_k^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k H_k (x_k - \bar{x}) \\
&= \sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top + \eta_k I_s)^{-1} P_k \overline{H} (x_k - \bar{x}) + \sqrt{\overline{H}} P_k^\top ((P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k (H_k - \overline{H})(x_k - \bar{x}).
\end{aligned} \tag{64}$$

Notice that by Assumption 14, we have that $\|(\overline{H} - H_k)\|$ tends to 0. Therefore, we deduce from (61) and (63) that

$$\|\sqrt{\overline{H}} P_k^\top ((P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k H_k - (P_k \overline{H} P_k^\top + \eta_k I_s)^{-1} P_k \overline{H})(x_k - \bar{x})\| = o(\|x_k - \bar{x}\|).$$

Therefore by (58), (63) and (64), we deduce that

$$\sqrt{\overline{H}}(x_{k+1} - \bar{x}) = -A + B + \sqrt{\overline{H}}(x_k - \bar{x}) = -A_1 + \sqrt{\overline{H}}(x_k - \bar{x}) + o(\|x_k - \bar{x}\|).$$

Hence, by evaluating the norm of A_2 as $o(\|x_k - \bar{x}\|)$, we deduce that with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4} s))$

$$\|\sqrt{\overline{H}}(x_{k+1} - \bar{x})\| \leq \left\| \left(I_n - \sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top + \eta_k I_s)^{-1} P_k \sqrt{\overline{H}} \right) \sqrt{\overline{H}}(x_k - \bar{x}) \right\| + o(\|x_k - \bar{x}\|).$$

We can write

$$\begin{aligned}
&\left(I_n - \sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top + \eta_k I_s)^{-1} P_k \sqrt{\overline{H}} \right) \sqrt{\overline{H}}(x_k - \bar{x}) \\
&= \left(I_n - \sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \sqrt{\overline{H}} \right) \sqrt{\overline{H}}(x_k - \bar{x}) \\
&\quad - \sqrt{\overline{H}} P_k^\top ((P_k \overline{H} P_k^\top + \eta_k I_s)^{-1} - (P_k \overline{H} P_k^\top)^{-1}) P_k \overline{H}(x_k - \bar{x}).
\end{aligned}$$

Hence, using the same reasoning as before, we obtain that

$$\|\sqrt{\overline{H}}(x_{k+1} - \bar{x})\| \leq \left\| \left(I_n - \sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \sqrt{\overline{H}} \right) \sqrt{\overline{H}}(x_k - \bar{x}) \right\| + o(\|x_k - \bar{x}\|). \tag{65}$$

Notice that $\sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \sqrt{\overline{H}}$ is an orthogonal projection, hence

$$\begin{aligned}
&\left\| \left(I_n - \sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \sqrt{\overline{H}} \right) \sqrt{\overline{H}}(x_k - \bar{x}) \right\|^2 \\
&= \|\sqrt{\overline{H}}(x_k - \bar{x})\|^2 - \|\sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \overline{H}(x_k - \bar{x})\|^2.
\end{aligned}$$

Then similarly to the proof of Proposition 19 and similarly to (37), we have that with probability at least $1 - 6(\exp(-s) + \exp(-\frac{\mathcal{C}_0}{4} s))$,

$$\|\sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \overline{H}(x_k - \bar{x})\|^2 = (x_k - \bar{x})^\top \Pi (x_k - \bar{x}),$$

and

$$\|\sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \overline{H}(x_k - \bar{x})\|^2 \geq \frac{\lambda_0}{2\lambda_{\max}(P_k \overline{H} P_k^\top)} \|\sqrt{\overline{H}}(x_k - \bar{x})\|^2,$$

where λ_0 is the first non-zero eigenvalue of $\overline{H}$. Therefore, we have that

$$\left\| \left(I_n - \sqrt{\overline{H}} P_k^\top (P_k \overline{H} P_k^\top)^{-1} P_k \sqrt{\overline{H}} \right) \sqrt{\overline{H}}(x_k - \bar{x}) \right\| \leq \sqrt{1 - \frac{\lambda_0}{2\lambda_{\max}(P_k \overline{H} P_k^\top)}} \|\sqrt{\overline{H}}(x_k - \bar{x})\|.$$

Therefore, by (65), we have that

$$\|\sqrt{\overline{H}}(x_{k+1} - \bar{x})\| \leq \sqrt{1 - \frac{\lambda_0}{2\lambda_{\max}(P_k \overline{H} P_k^\top)}} \|\sqrt{\overline{H}}(x_k - \bar{x})\| + o(\|x_k - \bar{x}\|).$$

By Lemma 25, we have $o(\|x_k - \bar{x}\|) = o(\|\sqrt{\overline{H}}(x_k - \bar{x})\|)$, hence we deduce that when k is large enough,

$$\|\sqrt{\overline{H}}(x_{k+1} - \bar{x})\| \leq \sqrt{1 - \frac{\lambda_0}{4\lambda_{\max}(P_k \overline{H} P_k^\top)}} \|\sqrt{\overline{H}}(x_k - \bar{x})\|.$$

We complete the proof using (44). ◀

5.2 Impossibility of local super-linear convergence in general

In this section we will prove that when f is strongly convex locally around the strict local minimizer $\bar{x}$, we cannot aim, with high probability, at local super-linear convergence using random subspace. More precisely, the goal of this section is to prove that there exists a constant $c > 0$ such that when k is large enough, we have that with probability $1 - 2\exp(-\frac{C_0}{4}) - 2\exp(-s)$,

$$\|x_{k+1} - \bar{x}\| \geq c\|x_k - \bar{x}\|.$$

From that, we will easily deduce that there exists a constant c' such that

$$f(x_{k+1}) - f(\bar{x}) \geq c'(f(x_k) - f(\bar{x}))$$

holds with high probability when k is large enough. This will prove that the results obtained in the previous section are optimal when f is locally strongly-convex. Indeed, by local strong-convexity of f and Hessian Lipschitz continuity (i.e. Assumption 14), there exists $l_2 \geq l_1 > 0$ such that for k large enough,

$$l_1\|x_k - \bar{x}\|^2 \leq f(x_k) - f(\bar{x}) \leq l_2\|x_k - \bar{x}\|^2.$$

This immediately proves the existence of the constant c' described above. In this subsection we make the following additional assumption.

▶ **Assumption 27.** *We assume that*

$$(\mathcal{C} + 2)^2 s < n,$$

where $\mathcal{C}$ is the constant that appears in (29).

We recall here that for all k:

$$x_{k+1} = x_k - t_k P_k^\top ((P_k \nabla^2 f(x_k) P_k^\top) + \eta_k I_s)^{-1} P_k \nabla f(x_k),$$

where t_k is the step-size and $\eta_k > 0$ is a parameter that tends to 0 when k tends to infinity.

Let us fix k. Using a Taylor expansion of $t \mapsto \nabla f(\bar{x} + t(x_{k+1} - \bar{x}))$ around 0, as in (54), we have that

$$\|\nabla f(x_{k+1})\| \leq \int_0^1 \|\nabla^2 f(\bar{x} + t(x_{k+1} - \bar{x}))\|\|x_{k+1} - \bar{x}\|\mathrm{d}t \leq \int_0^1 2\lambda_{\max}(\nabla^2 f(\bar{x}))\|x_{k+1} - \bar{x}\|\mathrm{d}t, \tag{66}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue, and the second inequality holds for k large enough under Assumption 14. Hence, for k large enough and under Assumption 14,

$$\|x_{k+1} - \bar{x}\| \geq \frac{1}{2\lambda_{\max}(\nabla^2 f(\bar{x}))}\|\nabla f(x_{k+1})\| \tag{67}$$

holds. Using a Taylor expansion of ∇f around x_k, we have that

$$\nabla f(x_{k+1}) = \nabla f(x_k) + \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))(x_{k+1} - x_k)\,\mathrm{d}t.$$

Hence,

$$\nabla f(x_{k+1}) = \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) + \int_0^1 (\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x_k))(x_{k+1} - x_k).$$

We deduce therefore that

$$\|\nabla f(x_{k+1})\| \geq \|\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)\| - \int_0^1 \|(\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x_k))(x_{k+1} - x_k)\|.$$

By Assumption 14, the Hessian is L_H-Lipschitz in B_H. Since x_k and $x_k + t(x_{k+1} - x_k) \in B_H$ for k large enough, we have that for $t \leq 1$,

$$\|(\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x_k))(x_{k+1} - x_k))\| \leq L_H\|x_{k+1} - x_k\|^2.$$

Hence (67) leads to

$$\|x_{k+1} - \bar{x}\| \geq \frac{1}{2\lambda_{\max}(\nabla^2 f(\bar{x}))}\left(\|g_k + H_k(x_{k+1} - x_k)\| - L_H\|x_{k+1} - x_k\|^2\right). \tag{68}$$

▶ **Proposition 28.** *Assume that Assumption 14 and Assumption 27 hold and that f is strongly convex locally around $\bar{x}$. There exists a constant $\beta > 0$ such that if k is large enough, then with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}s) - 2\exp(-s)$, we have*

$$\|g_k + H_k(x_{k+1} - x_k)\| \geq \beta \|x_{k+1} - x_k\|.$$

Proof. Recalling the updated rule $x_{k+1} = x_k - t_k P_k^\top M_k^{-1} P_k g_k$ in Algorithm 1, we have

$$\|g_k + H_k(x_{k+1} - x_k)\| = \|(I_n - t_k H_k P_k^\top M_k^{-1} P_k) g_k\|,$$

where M_k is defined in (6). If k is large enough, H_k is invertible by strong convexity of f. Notice that $\|(I_n - t_k H_k P_k^\top M_k^{-1} P_k) g_k\| = \|H_k(H_k^{-1} - t_k P_k^\top M_k^{-1} P_k) g_k\|$. Hence since for any invertible matrix A we have $\|Ax\| \geq \frac{\|x\|}{\|A^{-1}\|}$, we deduce that

$$\|(I_n - t_k H_k P_k^\top M_k^{-1} P_k) g_k\| \geq \frac{1}{\|H_k^{-1}\|} \|(H_k^{-1} - t_k P_k^\top M_k^{-1} P_k) g_k\|.$$

Furthermore, we have

$$\|(H_k^{-1} - t_k P_k^\top M_k^{-1} P_k) g_k\|^2 = \|H_k^{-1} g_k\|^2 + \|t_k P_k^\top M_k^{-1} P_k g_k\|^2 - 2\langle H_k^{-1} g_k, t_k P_k^\top M_k^{-1} P_k g_k \rangle. \tag{69}$$

Let $H_k^{-1} g_k = P_k^\top z_1 + z_2$ be the orthogonal decomposition of $H_k^{-1} g_k$ on $\mathrm{Im}(P_k^\top)$ parallel to $\mathrm{Ker}(P_k)$. Since $P_k z_2 = 0$, we have

$$\langle H_k^{-1} g_k, t_k P_k^\top M_k^{-1} P_k g_k \rangle = \langle P_k^\top z_1, t_k P_k^\top M_k^{-1} P_k g_k \rangle.$$

Hence, by (69), we deduce that

$$\|(H_k^{-1} - t_k P_k^\top M_k^{-1} P_k) g_k\|^2 \geq \|H_k^{-1} g_k\|^2 + \|t_k P_k^\top M_k^{-1} P_k g_k\|^2 - 2\|P_k^\top z_1\| \|t_k P_k^\top M_k^{-1} P_k g_k\|. \tag{70}$$

Since $H_k^{-1} g_k = P_k^\top z_1 + z_2$ with $P_k z_2 = 0$, we have that $P_k H_k^{-1} g_k = P_k P_k^\top z_1$. Which implies (since $P_k P_k^\top$ is invertible with probability 1) that $z_1 = (P_k P_k^\top)^{-1} P_k H_k^{-1} g_k$. Hence

$$\|P_k^\top z_1\| = \|P_k^\top (P_k P_k^\top)^{-1} P_k H_k^{-1} g_k\| \leq \|P_k^\top (P_k P_k^\top)^{-1}\| \|P_k H_k^{-1} g_k\|.$$

By Lemma 1, we have that with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}s)$ that $\|P_k H_k^{-1} g_k\| \leq 2\|H_k^{-1} g_k\|$. Furthermore, by writing the singular value decomposition, $U\Sigma V^\top$, of $P_k^\top$, we have that $\|P_k^\top (P_k P_k^\top)^{-1}\| = \|U\Sigma^{-1} V^\top\| = \frac{1}{\sigma_{\min}(P_k^\top)}$. Since $\sigma_{\min}(P_k^\top) \geq \sqrt{\frac{n}{s}} - \mathcal{C}$ holds with probability at least $1 - 2e^{-s}$ (we only consider the first equation of (29)), we deduce that

$$\|P_k^\top z_1\| \leq \frac{2}{\sqrt{\frac{n}{s}} - \mathcal{C}} \|H_k^{-1} g_k\|.$$

Hence, from (70) we have

$$\begin{aligned} &\|(H_k^{-1} - t_k P_k^\top M_k^{-1} P_k) g_k\|^2 \\ &\geq \|H_k^{-1} g_k\|^2 + \|t_k P_k^\top M_k^{-1} P_k g_k\|^2 - \frac{4}{\sqrt{\frac{n}{s}} - \mathcal{C}} \|H_k^{-1} g_k\| \|t_k P_k^\top M_k^{-1} P_k g_k\| \\ &\geq \left(1 - \frac{2}{\sqrt{\frac{n}{s}} - \mathcal{C}}\right) \|H_k^{-1} g_k\|^2 + \left(1 - \frac{2}{\sqrt{\frac{n}{s}} - \mathcal{C}}\right) \|t_k P_k^\top M_k^{-1} P_k g_k\|^2, \end{aligned} \tag{71}$$

where we used that $2ab \leq a^2 + b^2$ in the last inequality, and that $\left(1 - \frac{2}{\sqrt{\frac{n}{s}} - \mathcal{C}}\right) > 0$ holds by Assumption 27. Hence, from (71) we proved that

$$\begin{aligned} \|(I_n - t_k H_k P_k^\top M_k^{-1} P_k) g_k\|^2 &\geq \frac{1}{\|H_k^{-1}\|^2} \left(1 - \frac{2}{\sqrt{\frac{n}{s}} - \mathcal{C}}\right) \|t_k P_k^\top M_k^{-1} P_k g_k\|^2 \\ &= \frac{1}{\|H_k^{-1}\|^2} \left(1 - \frac{2}{\sqrt{\frac{n}{s}} - \mathcal{C}}\right) \|x_{k+1} - x_k\|^2. \end{aligned}$$

That is

$$\|g_k + H_k(x_{k+1} - x_k)\| \geq \frac{\sqrt{1 - \frac{2}{\sqrt{\frac{n}{s}} - \mathcal{C}}}}{\|H_k^{-1}\|}\|x_{k+1} - x_k\|. \tag{72}$$

Considering k large enough, as x_k tends to $\bar{x}$, we can bound, using Assumption 14, $\frac{1}{\|H_k^{-1}\|} \geq \frac{1}{2\|\overline{H}^{-1}\|}$, where we recall that $\overline{H} = \nabla^2 f(\bar{x})$, which ends the proof. ◀

▶ **Theorem 29.** *Assume that Assumption 14 and Assumption 27 hold and that f is locally strongly convex around $\bar{x}$. There exists a constant $c > 0$ such that for k large enough,*

$$\|x_{k+1} - \bar{x}\| \geq c\|x_k - \bar{x}\|$$

holds with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}s) - 2\exp(-s)$.

Proof. From (68) and Proposition 28 we deduce that with probability at least $1 - 2\exp(-\frac{\mathcal{C}_0}{4}s) - 2\exp(-s)$, when k is large enough

$$\|x_{k+1} - \bar{x}\| \geq \frac{1}{2\lambda_{\max}(\nabla^2 f(\bar{x}))}\left(\beta - L_H\|x_{k+1} - x_k\|\right)\|x_{k+1} - x_k\|.$$

Since $\beta > 0$, we have that for k large enough so as to yield $L_H\|x_{k+1} - x_k\| \leq \beta/2$,

$$\|x_{k+1} - \bar{x}\| \geq \frac{1}{2\lambda_{\max}(\nabla^2 f(\bar{x}))}\frac{\beta}{2}\|x_{k+1} - x_k\|.$$

Hence

$$\|x_{k+1} - \bar{x}\| \geq \frac{\beta}{4\lambda_{\max}(\overline{H})}\|x_{k+1} - x_k\|. \tag{73}$$

Since f is assumed to be strongly convex, for all $\alpha \in (0, 1)$, as $g_k^{\mathsf{T}} d_k \leq 0$. Hence we have that $t_k = 1$Now we notice that

$$\|x_{k+1} - x_k\| = t_k\|P_k^\top M_k^{-1} P_k g_k\| \geq t_k \sigma_{\min}(P_k^\top)\|M_k^{-1}\|\|P_k g_k\|. \tag{74}$$

Using Lemma 1 (with $\varepsilon = 1/2$) and the bound (29) on $\sigma_{\min}(P_k^\top)$, we have that

$$t_k\sigma_{\min}(P_k^\top)\|M_k^{-1}\|\|P_k g_k\| \geq t_k\left(\sqrt{\frac{n}{s}} - \mathcal{C}\right)\|M_k^{-1}\|\frac{1}{2}\|g_k\|. \tag{75}$$

Since x_k converges to $\bar{x}$ and the Hessian is Lipschitz continuous, we have that H_k converges to $\overline{H}$. Therefore, when k is large enough, we have $\|M_k^{-1}\| \geq \frac{1}{2}\|(P_k\overline{H}P_k^\top)^{-1}\| = \frac{1}{2}\|\overline{M}^{-1}\|$, where $\overline{M} := P_k\overline{H}P_k^\top$. Since

$$0 \prec \overline{M} \preceq \lambda_{\max}(\overline{H})P_k P_k^\top,$$

we deduce by Lemma 2

$$\|M_k^{-1}\| \geq \frac{1}{2\mathcal{C}\lambda_{\max}(\overline{H})\frac{n}{s}}. \tag{76}$$

Hence, by (73) to (76) we have that there exists a constant $\kappa_2 > 0$ such that

$$\|x_{k+1} - \bar{x}\| \geq \kappa_2\|g_k\|.$$

By (54) we have that

$$g_k = \overline{H}(x_k - \bar{x}) + \int_0^1 (\nabla^2 f(\bar{x} + t(x_k - \bar{x})) - \overline{H})(x_k - \bar{x})\,\mathrm{d}t.$$

Hence, since f is assumed to be locally strongly convex, by Assumption 14 we have that for k large enough:

$$\|g_k\| \geq \frac{\lambda_{\min}(\overline{H})}{2}\|x_k - \bar{x}\|.$$

Using (23), we have

$$f(x_k) - f(x_k + t'_k d_k) + \alpha t'_k g_k^\mathsf{T} d_k \geq \frac{\bar{\mathcal{C}} n}{2s} L_H {t'_k}^2 \|d_k\| \left(\frac{c_2 s \|g_k\|^\gamma}{\bar{\mathcal{C}} L_H n \|d_k\|} - t'_k \right) \left\| M_k^{-1} P_k g_k \right\|^2,$$

and since f is assume to be strongly convex, $\frac{\|g_k\|^\gamma}{\|d_k\|}$ is in the order of $\mathcal{O}(\frac{1}{\|g_k\|^{1-\gamma}})$, hence t_k is bounded below by some constant for k large enough. Hence we have for k large enough that

$$\|x_{k+1} - \bar{x}\| \geq \frac{1}{2} \kappa_2 \lambda_{\min}(\overline{H}) \|x_k - \bar{x}\|,$$

which concludes the proof. ◀

We have the following deterministic corollary:

▶ **Corollary 30.** *Assume that Assumption 14 and Assumption 27 hold and that f is locally strongly convex around $\bar{x}$. Then for k large enough,*

$$\mathbb{E}(\|x_{k+1} - \bar{x}\|) \geq \bar{c} \mathbb{E}(\|x_k - \bar{x}\|),$$

where $\bar{c} = (1 - 2\exp(-\frac{\mathcal{C}_0}{4} s) - 2\exp(-s))c$ (c is the same constant as in Theorem 29), and where the expectation is taken with respect to the random variables $P_0, \ldots, P_k$.

Proof. The proof is very similar to the proof of Corollary 23. Let us consider the random variable $\mathbb{E}[\|x_{k+1} - \bar{x}\| \mid P_0, \ldots, P_{k-1}]$. Let $\mathcal{E} = \{\|x_{k+1} - \bar{x}\| \geq \bar{c}\|x_k - \bar{x}\| \mid x_k\}$ be an event with respect to the random variable P_k. Using the fact that $\|x_{k+1} - \bar{x}\| \geq 0$, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\|x_{k+1} - \bar{x}\| \mid P_0, \ldots, P_{k-1}\right] \\
&= \mathbb{E}\left[\|x_{k+1} - \bar{x}\| \mid P_0, \ldots, P_{k-1}, \mathcal{E}\right] P(\mathcal{E}) + \mathbb{E}\left[\|x_{k+1} - \bar{x}\| \;\middle|\; P_0, \ldots, P_{k-1}, \bar{\mathcal{E}}\right] (1 - P(\mathcal{E})) \\
&\geq \bar{c}\|x_k - \bar{x}\|
\end{aligned}$$

Taking the expectation with respect to $P_0, \ldots, P_{k-1}$ leads to the result. ◀

5.3 The rank deficient case

Previously we proved that when f is locally strongly convex, super-linear convergence cannot hold for RS-RNM. Here we prove that when the Hessian $\overline{H}$ at the local optimum $\bar{x}$ is rank deficient, then RS-RNM can achieve super-linear convergence. In this whole subsection, we assume that Assumption 14 and Assumption 24 are satisfied. We also denote by r ($< n$) the rank of $\overline{H}$. Notice that, as a special case of $r < n$, one can consider "functions with low dimensionality"[4] [42]. For such functions, there exists a projection matrix $\Pi \in \mathbb{R}^{n \times n}$ with $\text{rank}(\Pi) < n$ such that

$$\forall\, x \in \mathbb{R}^n, \quad f(x) = f(\Pi x). \tag{77}$$

Such functions are frequently encountered in many applications. For example, the loss functions of neural networks often have low rank Hessians [21, 36, 33]. This phenomenon is also prevalent in other areas such as hyper-parameter optimization for neural networks [3], heuristic algorithms for combinatorial optimization problems [23], complex engineering and physical simulation problems as in climate modeling [26], and policy search [17].

We first prove the following lemma which is very similar to Lemma 25.

▶ **Lemma 31.** *We have, under Assumption 14 and Assumption 24, that for k large enough:*

$$\frac{\rho}{2} \|x_k - \bar{x}\| \leq \|\overline{H}(x_k - \bar{x})\|.$$

Furthermore,

$$\|g_k\| \leq 2\lambda_{\max}(\overline{H}) \|x_k - \bar{x}\|.$$

[4] They are also called objectives with "active subspaces" [10], or "multi-ridge" [16].

Proof. As in the proof of Lemma 25, we have (56), i.e.,

$$\rho\|x_k - \bar{x}\| - L_H\|x_k - \bar{x}\|^2 \leq \|\overline{H}(x_k - \bar{x})\|.$$

Since $\|x_k - \bar{x}\|$ tends to 0, we deduce that for k large enough:

$$\frac{\rho}{2}\|x_k - \bar{x}\| \leq \|\overline{H}(x_k - \bar{x})\|.$$

The other inequality is easy to deduce from (54), as in (66):

$$\|g_k\| \leq \|\overline{H}\|\|x_k - \bar{x}\| + L_H\|x_k - \bar{x}\|^2 \leq 2\lambda_{\max}(\overline{H})\|x_k - \bar{x}\|, \tag{78}$$

when k is large enough such that $L_H\|x_k - \bar{x}\| \leq \lambda_{\max}(\overline{H})$ holds. ◀

The next lemma is the key to prove super-linear convergence. Notice that since $s \geq r$, we have that with probability one $\sigma_{\min}(P_k^1) > 0$.

▶ **Lemma 32.** *Under Assumption 14 and Assumption 24. If $s \geq r$, we have that for k large enough, with probability at least $1 - 2\exp(-s)$:*

$$\|P_k g_{k+1}\| \geq \rho\frac{\sigma_{\min}(P_k^1)}{8\lambda_{\max}(\overline{H})}\|g_{k+1}\|,$$

where $P_k^1 \in \mathbb{R}^{s\times r}$ is an $s \times r$ i.i.d. Gaussian matrix having the same distribution with P_k.

Proof. By (54) applied at $k+1$, we have that

$$\nabla f(x_{k+1}) = \int_0^1 \nabla^2 f(\bar{x} + t(x_{k+1} - \bar{x}))(x_{k+1} - \bar{x})\,\mathrm{d}t.$$

Hence,

$$P_k g_{k+1} = P_k\overline{H}(x_{k+1} - \bar{x}) + \int_0^1 P_k(\nabla^2 f(\bar{x} + t(x_{k+1} - \bar{x})) - \overline{H})(x_{k+1} - \bar{x}),$$

which leads to

$$\|P_k g_{k+1}\| \geq \|P_k\overline{H}(x_{k+1} - \bar{x})\| - L_H\|P_k\|\|x_{k+1} - \bar{x}\|^2. \tag{79}$$

Let $UDU^\top = \overline{H}$ be the diagonal decomposition of $\overline{H}$. Since $\bar{x}$ is a strict local minimizer, by Assumption 24, for k large enough, U is an orthogonal matrix independent of P_k, and hence, $\widetilde{P}_k := P_k U$ is an i.i.d. random Gaussian matrix with the same distribution as P_k. Let $y_{k+1} = U^\top(x_{k+1} - \bar{x})$. We have that

$$\overline{H}(x_{k+1} - \bar{x}) = UDy_{k+1} \quad \text{and thus,} \quad P_k\overline{H}(x_{k+1} - \bar{x}) = \widetilde{P}_k D y_{k+1}. \tag{80}$$

Furthermore, since D has rank $r < n$, we can write $Dy_{k+1} = \binom{z_{k+1}}{0}$, where $z_{k+1} \in \mathbb{R}^r$. We have therefore that

$$\|P_k\overline{H}(x_{k+1} - \bar{x})\| = \|P_k^1 z_{k+1}\|, \tag{81}$$

where $P_k^1 \in \mathbb{R}^{s\times r}$ is a submatrix of $\widetilde{P}_k$, i.e., $\widetilde{P}_k = \begin{pmatrix} P_k^1 & P_k^2 \end{pmatrix}$. Notice that from the definition of y_{k+1} and z_{k+1}, we have, by orthogonality of U, that

$$\|z_{k+1}\| = \|Dy_{k+1}\| \overset{(80)}{=} \|\overline{H}(x_{k+1} - \bar{x})\| \geq \frac{\rho}{2}\|x_{k+1} - \bar{x}\|,$$

where the inequality follows from Lemma 31. Hence, from (79) and (81), we deduce that

$$\|P_k g_{k+1}\| \geq \rho\frac{\sigma_{\min}(P_k^1)}{2}\|x_{k+1} - \bar{x}\| - L_H\|P_k\|\|x_{k+1} - \bar{x}\|^2.$$

Using that $\|P_k\|$ is bounded, with probability at least $1 - 2\exp(-s)$, by Lemma 2, we deduce, as in the proof of Lemma 31, that for k large enough:

$$\|P_k g_{k+1}\| \geq \rho\frac{\sigma_{\min}(P_k^1)}{4}\|x_{k+1} - \bar{x}\| \overset{(78)}{\geq} \rho\frac{\sigma_{\min}(P_k^1)}{4}\frac{\|g_{k+1}\|}{2\lambda_{\max}(\overline{H})}.$$

That is:

$$\|P_k g_{k+1}\| \geq \rho\frac{\sigma_{\min}(P_k^1)}{8\lambda_{\max}(\overline{H})}\|g_{k+1}\|.$$ ◀

Similarly, we have the following lemma.

▶ **Lemma 33.** *Let $M \in \mathbb{R}^{n\times n}$ be any matrix. Under Assumption 14 and Assumption 24, if k is large enough and $s \geq r$, we have*

$$\frac{\sigma_{\min}(P_k^1)}{2}\|H_k M\| \leq \|P_k H_k M\|.$$

Proof. The proof is very similar to the proof of Lemma 32. We have

$$\|P_k H_k M\| \geq \|P_k \overline{H} M\| - \|P_k(H_k - \overline{H})M\|. \tag{82}$$

Let $UDU^\top = \overline{H}$ be the diagonal decomposition of $\overline{H}$. Similarly to the proof of Lemma 32, $\widetilde{P}_k := P_k U$ is an i.i.d. random Gaussian matrix with the same distribution as P_k. Using $N := U^\top M$, we have that $P_k \overline{H} M = \widetilde{P}_k D N$. Furthermore, since D has rank $r < n$, we can write $DN = \begin{pmatrix} \widetilde{N} \\ 0 \end{pmatrix}$, where $\widetilde{N} \in \mathbb{R}^{r\times n}$. We have therefore that

$$\|P_k \overline{H} M\| = \|P_k^1 \widetilde{N}\|, \tag{83}$$

where $P_k^1 \in \mathbb{R}^{s\times r}$ is a submatrix of $\widetilde{P}_k$, i.e., $\widetilde{P}_k = \begin{pmatrix} P_k^1 & P_k^2 \end{pmatrix}$. Therefore

$$\|P_k \overline{H} M\| \geq \sigma_{\min}(P_k^1)\|\widetilde{N}\| = \sigma_{\min}(P_k^1)\|DN\| = \sigma_{\min}(P_k^1)\|\overline{H} M\|, \tag{84}$$

where the last equality holds by orthogonality of U. We deduce therefore, from (82) and (84) that

$$\|P_k H_k M\| \geq \sigma_{\min}(P_k^1)\|H_k M\| - \sigma_{\min}(P_k^1)\|(\overline{H} - H_k)M\| - \|P_k(H_k - \overline{H})M\|.$$

Since H_k tends to $\overline{H}$, we have the desired result for k large enough. ◀

The next lemma, similar to Lemma 5.2 of [39], is needed to control $\eta_k = c_1\Lambda_k + c_2\|g_k\|^\gamma$, where $\Lambda_k = \max(0, -\lambda_{\min}(P_k H_k P_k^\top))$.

▶ **Lemma 34.** *Under Assumption 14, for k large enough, we have that with probability at least $1 - 2\exp(-s)$,*

$$\Lambda_k \leq \frac{\bar{\mathcal{C}} n}{s} L_H \|x_k - \bar{x}\|.$$

Proof. The result is obvious when $\Lambda_k = 0$. Let us consider the case $\Lambda_k > 0$. Let $\lambda_k = (\lambda_k^{(1)}, \dots, \lambda_k^{(s)})$ be a vector of eigenvalues of $P_k \overline{H} P_k^\top$ and we write the eigenvalue decomposition of $P_k \overline{H} P_k^\top$ as follows:

$$P_k \overline{H} P_k^\top = U_k^\top \operatorname{diag}(\lambda_k) U_k.$$

Notice that $\lambda_{\min}(P_k H_k P_k^\top) I_s - U_k P_k H_k P_k^\top U_k^\top$ is singular. Furthermore,

$$\lambda_{\min}(P_k H_k P_k^\top) I_s - \operatorname{diag}(\lambda_k)$$

is not singular as $\lambda_{\min}(P_k H_k P_k^\top) < 0$ by assumption and $\operatorname{diag}(\lambda_k)$ is positive. We define

$$A_k = (\lambda_{\min}(P_k H_k P_k^\top) I_s - \operatorname{diag}(\lambda_k))^{-1}(\lambda_{\min}(P_k H_k P_k^\top) I_s - U_k P_k H_k P_k^\top U_k^\top),$$

which is therefore singular. Notice furthermore that since $\lambda_{\min}(P_k H_k P_k^\top) < 0$,

$$\|(\lambda_{\min}(P_k H_k P_k^\top) I_s - \operatorname{diag}(\lambda_k))^{-1}\| \leq \frac{1}{-\lambda_{\min}(P_k H_k P_k^\top)} = \frac{1}{\Lambda_k}. \tag{85}$$

Hence we have

$$\begin{aligned}
1 &\leq \|I_s - A_k\| \\
&= \|I_s - (\lambda_{\min}(P_k H_k P_k^\top) I_s - \operatorname{diag}(\lambda_k))^{-1}(\lambda_{\min}(P_k H_k P_k^\top) I_s - U_k P_k H_k P_k^\top U_k^\top)\| \\
&= \|I_s - (\lambda_{\min}(P_k H_k P_k^\top) I_s - \operatorname{diag}(\lambda_k))^{-1}(\lambda_{\min}(P_k H_k P_k^\top) I_s - \operatorname{diag}(\lambda_k) - U_k P_k (H_k - \overline{H}) P_k^\top U_k^\top)\| \\
&= \|(\lambda_{\min}(P_k H_k P_k^\top) I_s - \operatorname{diag}(\lambda_k))^{-1} U_k P_k (H_k - \overline{H}) P_k^\top U_k^\top\| \\
&\overset{(85)}{\leq} \frac{1}{\Lambda_k}\|P_k P_k^\top\|\|H_k - \overline{H}\| \\
&\overset{\text{Lemma 2}}{\leq} \frac{1}{\Lambda_k}\frac{\bar{\mathcal{C}} n}{s} L_H \|x_k - \bar{x}\|,
\end{aligned}$$

where the first inequality is a well known inequality for a singular matrix and is proved in [39, Lemma 5.1]. ◀

Let us recall that

$$d_k = -P_k^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1} P_k g_k,$$

and

$$M_k = P_k H_k P_k^\top + \eta_k I_s.$$

▶ **Lemma 35.** *Under Assumption 14 and Assumption 24, if $s \geq r$, we have that for k large enough, we have that with probability at least $1 - 2\exp(-s)$,*

$$\|d_k\| \leq \frac{4}{\sigma_{\min}(P_1^k)} \left(2 + \frac{1}{c_1 - 1}\right) \sqrt{\frac{\bar{\mathcal{C}} n}{s}} \|x_k - \bar{x}\|,$$

where $P_k^1 \in \mathbb{R}^{s \times r}$ is an $s \times r$ i.i.d. Gaussian matrix having the same distribution with P_k.

Proof. Notice first that by Taylor expansion of $t \mapsto \nabla f(\bar{x} + t(x_k - \bar{x}))$ and by Assumption 14, we have that

$$\|g_k - \nabla f(\bar{x}) - H_k(x_k - \bar{x})\| \leq \frac{L_H}{2} \|x_k - \bar{x}\|^2. \tag{86}$$

The definition of d_k leads to

$$\begin{aligned}
\|d_k\| &= \|P_k^\top M_k^{-1} P_k g_k\| \\
&\overset{\nabla f(\bar{x})=0}{=} \|P_k^\top M_k^{-1} P_k (g_k - \nabla f(\bar{x}) - H_k(x_k - \bar{x}) + H_k(x_k - \bar{x}))\| \\
&\leq \|P_k\|^2 \|M_k^{-1}\| \|g_k - \nabla f(\bar{x}) - H_k(x_k - \bar{x})\| + \|P_k^\top M_k^{-1} P_k H_k\| \|x_k - \bar{x}\| \\
&\overset{(86)}{\leq} \frac{L_H}{2} \|P_k\|^2 \|M_k^{-1}\| \|x_k - \bar{x}\|^2 + \|P_k^\top M_k^{-1} P_k H_k\| \|x_k - \bar{x}\|.
\end{aligned} \tag{87}$$

Let us first bound the first term in the right-hand side of (87). When k is large enough, with probability at least $1 - 2\exp(-s)$, we have by Lemma 2

$$\begin{aligned}
\frac{L_H}{2} \|P_k\|^2 \|M_k^{-1}\| &\leq \frac{L_H}{2} \cdot \frac{\bar{\mathcal{C}} n}{s} \cdot \frac{1}{\lambda_{\min}(P_k H_k P_k^\top + c_1 \Lambda_k I_s + c_2 \|g_k\|^\gamma I_s)} \\
&\leq \frac{L_H \bar{\mathcal{C}} n}{2 c_2 s \|g_k\|^\gamma} \\
&\overset{(53)}{\leq} \frac{L_H \bar{\mathcal{C}} n}{2 c_2 s \rho^\gamma \|x_k - \bar{x}\|^\gamma}.
\end{aligned}$$

Hence

$$\frac{L_H}{2} \|P_k\|^2 \|M_k^{-1}\| \|x_k - \bar{x}\|^2 \leq \frac{L_H \bar{\mathcal{C}} n}{2 c_2 s \rho^\gamma} \|x_k - \bar{x}\|^{2-\gamma}. \tag{88}$$

Next, we consider the second term $\|P_k^\top M_k^{-1} P_k H_k\| \|x_k - \bar{x}\|$. Notice that

$$\|P_k^\top M_k^{-1} P_k H_k\| = \|H_k P_k^\top M_k^{-1} P_k\| \leq \frac{2}{\sigma_{\min}(P_1^k)} \|P_k H_k P_k^\top M_k^{-1}\| \|P_k\|,$$

where the inequality follows from Lemma 33. We have

$$\begin{aligned}
\|P_k H_k P_k^\top M_k^{-1}\| &= \|P_k H_k P_k^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1}\| \\
&\leq \|(P_k H_k P_k^\top + \eta_k I_s)^\top (P_k H_k P_k^\top + \eta_k I_s)^{-1}\| + \eta_k \|(P_k H_k P_k^\top + \eta_k I_s)^{-1}\| \\
&\leq 1 + \frac{\eta_k}{\lambda_{\min}(P_k H_k P_k^\top + \eta_k I_s)} \\
&\leq 1 + \frac{c_1 \Lambda_k + c_2 \|g_k\|^\gamma}{(c_1 - 1)\Lambda_k + c_2 \|g_k\|^\gamma} \\
&\leq 2 + \frac{1}{c_1 - 1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\|P_k^\top M_k^{-1} P_k H_k\|\|x_k-\bar{x}\| &\leq \frac{2}{\sigma_{\min}(P_1^k)}\left(2+\frac{1}{c_1-1}\right)\|P_k\|\|x_k-\bar{x}\| \\ &\leq \frac{2}{\sigma_{\min}(P_1^k)}\left(2+\frac{1}{c_1-1}\right)\sqrt{\frac{\bar{\mathcal{C}}n}{s}}\|x_k-\bar{x}\|,\end{aligned}$$

where the second inequality follows from Lemma 2. The results follows from (87) and (88) noticing that $\frac{\|x_k-\bar{x}\|^{2-\gamma}}{\|x_k-\bar{x}\|}$ tends to 0, as $\gamma<1$, hence for k large enough

$$\frac{L_H\bar{\mathcal{C}}n}{2c_2s\rho^\gamma}\|x_k-\bar{x}\|^{2-\gamma} \leq \frac{2}{\sigma_{\min}(P_1^k)}\left(2+\frac{1}{c_1-1}\right)\sqrt{\frac{\bar{\mathcal{C}}n}{s}}\|x_k-\bar{x}\|. \qquad \blacktriangleleft$$

▶ **Theorem 36.** *Under Assumption 14 and Assumption 24, for k large enough and for any $s \geq r$, we have that with probability at least $1-2\exp(-s)$*

$$\|x_{k+1}-\bar{x}\| \leq \frac{c_2\Gamma}{\sigma_{\min}^2(P_k^1)}\|x_k-\bar{x}\|^{1+\gamma},$$

where Γ is some constant depending on n and s, and where $P_k^1 \in \mathbb{R}^{s\times r}$ is an $s\times r$ i.i.d. Gaussian matrix having the same distribution with P_k.

Proof. We have

$$\begin{aligned}\|x_{k+1}-\bar{x}\| &\overset{(53)}{\leq} \frac{1}{\rho}\|g_{k+1}\| \\ &\leq \frac{8\lambda_{\max}(\overline{H})}{\rho^2\sigma_{\min}(P_k^1)}\|P_k g_{k+1}\| \\ &\leq \frac{8\lambda_{\max}(\overline{H})}{\rho^2\sigma_{\min}(P_k^1)}\left(\|P_k(g_{k+1}-g_k-H_k(x_{k+1}-x_k))\|+\|P_k g_k+P_kH_k(x_{k+1}-x_k)\|\right), \qquad (89)\end{aligned}$$

where the first inequality holds by (53), and the second holds by Lemma 32. By Lemma 35 and an equation similar to (86) (where x_k is replaced by x_{k+1} and $\bar{x}$ is replaced by x_k), we have that

$$\|P_k(g_{k+1}-g_k-H_k(x_{k+1}-x_k))\| \leq L_H\|P_k\|\left(\frac{4}{\sigma_{\min}(P_1^k)}\left(2+\frac{1}{c_1-1}\right)\sqrt{\frac{\bar{\mathcal{C}}n}{s}}\right)^2\|x_k-\bar{x}\|^2. \qquad (90)$$

From the updated rule $x_{k+1}=x_k-t_kP_k^\mathsf{T}M_k^{-1}P_kg_k$ in Algorithm 1, we see that $x_{k+1}-x_k=-t_kP_k^\mathsf{T}M_k^{-1}P_kg_k$. From now on, we will show that $t_k=1$ for k large enough. Indeed by (23), we have that

$$f(x_k)-f(x_k+t_k'd_k)+\alpha t_k'g_k^\mathsf{T}d_k \geq \frac{\bar{\mathcal{C}}n}{2s}L_H{t_k'}^2\|d_k\|\left(\frac{c_2s\|g_k\|^\gamma}{\bar{\mathcal{C}}L_Hn\|d_k\|}-t_k'\right)\left\|M_k^{-1}P_kg_k\right\|^2.$$

Hence, by Assumption 24 and Lemma 35, we deduce that there exists some constant $\mathcal{C}_1$ such that

$$f(x_k)-f(x_k+t_k'd_k)+\alpha t_k'g_k^\mathsf{T}d_k \geq \frac{\bar{\mathcal{C}}n}{2s}L_H{t_k'}^2\|d_k\|\left(\frac{\mathcal{C}_1}{\|x_k-\bar{x}\|^{1-\gamma}}-t_k'\right)\left\|M_k^{-1}P_kg_k\right\|^2,$$

proving that we can take $t_k'=1$ if $\|x_k-\bar{x}\|$ is small enough.
Now notice that for k large enough, $t_k=1$, hence

$$\begin{aligned}\|P_kg_k+P_kH_k(x_{k+1}-x_k)\| &= \|(I_s-P_kH_kP_k^\top(P_kH_kP_k^\top+\eta_kI_s)^{-1})P_kg_k\| \\ &\leq \|\eta_k(P_kH_kP_k^\top+\eta_kI_s)^{-1}P_kg_k\| \\ &\leq \frac{\eta_k}{\sigma_{\min}(P_k^\top)}\|P_k^\top(P_kH_kP_k^\top+\eta_kI_s)^{-1}P_kg_k\| \\ &= \frac{\eta_k}{\sigma_{\min}(P_k^\top)}\|d_k\|.\end{aligned}$$

Using that $\|\eta_k\| \leq c_1\|\Lambda_k\| + c_2\|g_k\|^\gamma$ and that $\|g_k\| = O(\|x_k - \bar{x}\|)$ by Lemma 31, we deduce, by Lemmas 34 and 35, that there exists some constants α, β, $\beta' > 0$ such that with probability at least $1 - 2\exp(-s)$,

$$\begin{aligned}\frac{\eta_k}{\sigma_{\min}(P_k^\top)}\|d_k\| &\leq \frac{1}{\sigma_{\min}^2(P_k^\top)}\left(c_1\alpha\|x_k - \bar{x}\|^2 + c_2\beta\|x_k - \bar{x}\|\|g_k\|^\gamma\right)\\ &\leq \frac{1}{\sigma_{\min}^2(P_k^\top)}\left(c_1\alpha\|x_k - \bar{x}\|^2 + c_2\beta'\|x_k - \bar{x}\|^{1+\gamma}\right),\end{aligned}$$

where we have used in the second inequality that $\|g_k\| \leq O(\|x_k - \bar{x}\|)$. Now by (89), (90) and the above, we obtain the desired result. ◀

Notice that by using [35], we can furthermore bound $\frac{1}{\sigma_{\min}(P_k^1)}$, with high probability, by $O(\frac{1}{\sqrt{s}-\sqrt{r-1}})$.

Let us consider a function with low dimensionality, i.e. satisfying (77). Let us write $\Pi = R^\top R$, where $R \in \mathbb{R}^{s\times n}$ and let us define $g : y \in \mathbb{R}^s \mapsto f(R^\top)$. Hence, we have that $g(Rx) = f(\Pi x) = f(x)$. By denoting $y_k := Rx_k \in \mathbb{R}^s$ and assuming that the function $g(y)$ is strongly convex, locally near $\bar{y} := R\bar{x}$, it is easy to see that Assumption 24 is satisfied for the sequence $\{y_k\}$, locally, i.e., there exists $\rho > 0$ such that for k large enough;

$$\|\nabla g(y_k)\| \geq \rho\|y_k - \bar{y}\|$$

holds. Hence, we can prove that there exists some constant $\mathcal{K} >$ such that the following inequality holds with high probability.

$$\|y_{k+1} - \bar{y}\| \leq \mathcal{K}\|y_k - \bar{y}\|^{1+\gamma}.$$

By strong convexity of $g(y)$, we know that there exists two constant $l_1 > l_2 > 0$ such that

$$l_2(g(y_{k+1} - g(\bar{y}))) \leq \|y_{k+1} - \bar{y}\| \leq l_1(g(y_{k+1} - g(\bar{y}))).$$

Hence by following the same proof as in Corollary 23, we can obtain the following super-linear rate in expectation:

▶ **Theorem 37.** *Assume that there exists a function $g : y \in \mathbb{R}^s \mapsto g(y)$ such that $g(Rx) = f(x)$, for some matrix $R \in \mathbb{R}^{s\times n}$ $(s < n)$. If the function $g(y)$ is strongly convex, locally near $R\bar{x}$, then there exists a constant $\mathcal{K}' > 0$, such that if k is large enough:*

$$\mathbb{E}\left[f(x_{k+1}) - f(\bar{x})\right] \leq \mathcal{K}'\mathbb{E}\left[f(x_k) - f(\bar{x})\right]^{1+\gamma}, \tag{91}$$

6 Numerical illustration

In this section, we illustrate numerically the randomized subspace regularized Newton method (RS-RNM). All results are obtained using Python scripts on a 12th Gen Intel(R) Core(TM) i9-12900HK 2.50 GHz with 64GB of RAM. As a benchmark, we compare it against the gradient descent method (GD) and the regularized Newton method (RNM) [39]. Here we do not aim to prove that our method is faster to the state-of-the-art methods but rather to illustrate the theoretical results that have been proved in the previous sections.

6.1 Support vector regression

The methods are tested on a support vector regression problem formulated as minimizing sum of a loss function and a regularizer:

$$f(w) = \frac{1}{m}\sum_{i=1}^{m}\ell(y_i - x_i^\mathsf{T} w) + \lambda\|w\|^2. \tag{92}$$

Here, $(x_i, y_i) \in \mathbb{R}^n \times \{0, 1\}$ $(i = 1, 2, \ldots, m)$ denote the training example and ℓ is the loss function. λ is a constant of the regularizer and is fixed to 0.01 in the numerical experiments below. We note that (92) is a type of (generalized) linear model used in the numerical experiments of [18] and [22]. As the loss function ℓ, we use the following two functions known as robust loss functions: the Geman–McClure loss function (ℓ_1) and the Cauchy loss function (ℓ_2) [2] defined as

$$\begin{aligned}\ell_1(t) &= \frac{2t^2}{t^2 + 4},\\ \ell_2(t) &= \log\left(\frac{1}{2}t^2 + 1\right).\end{aligned}$$

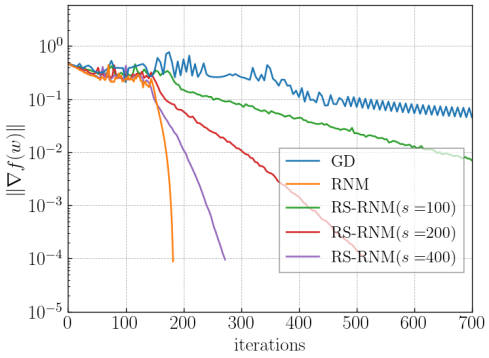

Figure 1 Iterations versus $\|\nabla f(w)\|$ ($\log_{10}$-scale) for Geman–McClure loss

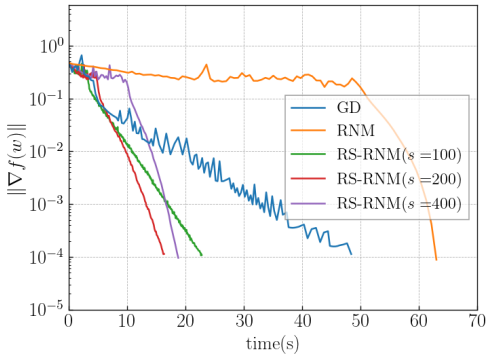

Figure 2 Computation time versus $\|\nabla f(w)\|$ ($\log_{10}$-scale) for Geman–McClure loss

Since both loss functions ℓ_1 and ℓ_2 are non-convex, the objective function (92) is non-convex.

The search directions at each iteration in GD and RNM are given by

$$d_k^{\mathrm{GD}} = -\nabla f(w_k),$$
$$d_k^{\mathrm{RNM}} = -(\nabla^2 f(w_k) + c_1'\Lambda_k' I_n + c_2' \|\nabla f(w_k)\|^{\gamma'} I_n)\nabla f(w_k),$$
$$(\Lambda_k' = \max(0, -\lambda_{\min}(\nabla^2 f(w_k)))$$

and the step sizes are all determined by Armijo backtracking line search (8) with the same parameters α and β for the sake of fairness. The parameters shown above and in Section 3 are fixed as follows:

$$c_1 = c_1' = 2, c_2 = c_2' = 1, \gamma = \gamma' = 0.5, \alpha = 0.3, \beta = 0.5, s \in \{100, 200, 400\}.$$

We test the methods on internet advertisements dataset from UCI repository [15] that is processed so that the number of instances is $600(= m)$ and the number of data attributes is $1500(= n)$, and the results, until the stop condition $\|\nabla f(w_k)\| < 10^{-4}$ is satisfied, are shown in Figures 1 to 4. Our first observation is that RS-RNM converges faster than GD. GD does not require the calculation of Hessian or its inverse, making the time per iteration small. However, it usually needs a large number of iterations, resulting in slow convergence. Next, we look at the comparison between RNM and RS-RNM. From Figures 1 and 3, we see that RNM has the same or a larger decrease in the function value in one iteration than RS-RNM, and it takes fewer iterations to converge. This is possibly due to the fact that RNM determines the search direction in full-dimensional space. In particular, it should be mentioned that RNM converges rapidly from a certain point on, as it is shown that RNM has a super-linear rate of convergence near a local optimal solution. However, as shown in Figures 2 and 4, since RNM takes a long time to get close to the local solution due to the heavy calculation of the full regularized Hessian, RS-RNM results in faster convergence than RNM. We also confirm on Figure 3 that for small dimensions $s = 100, 200$ a linear convergence rate seems to be achieved. However for $s = 400$ it seems that the method converges super-linearly.

6.2 Low rank Rosenbrock function

To properly illustrate the superlinear convergence proved in the low rank setting (cf. Section 5.3), we conducted numerical experiments on a low rank Rosenbrock function: $f(x) = R(U^\top U x)$, where

$$R(x) = \sum_{i=1}^{n-1} 100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2,$$

and $U \in \mathbb{R}^{r\times n}$ is a matrix whose columns are orthogonal. If we denote by $\Pi \in \mathbb{R}^{n\times n}$ the matrix $U^\top U$, we see that for all $x \in \mathbb{R}^n$, $f(x) = f(\Pi x)$, hence the Hessian of f is of rank r for all $x \in \mathbb{R}^n$. The parameters in Section 3 are fixed as follows:

$$c_1 = c_1' = 2,\ c_2 = c_2' = 1,\ \gamma = \gamma' = 0.5,\ \alpha = 0.3,\ \beta = 0.5,\ s \in \{100, 200, 600\}.$$

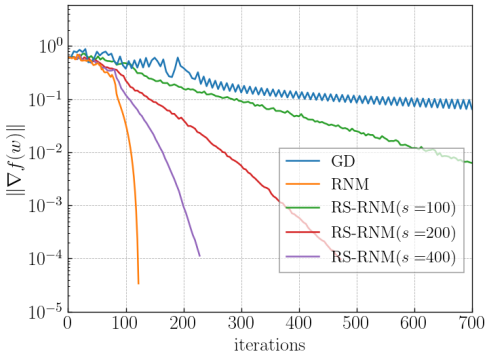

Figure 3 Iterations versus $\|\nabla f(w)\|$ ($\log_{10}$-scale) for Cauchy loss

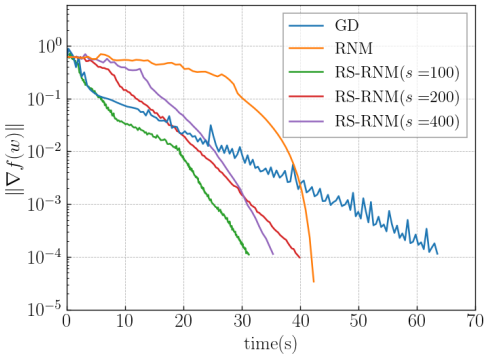

Figure 4 Computation time versus $\|\nabla f(w)\|$ ($\log_{10}$-scale) for Cauchy loss

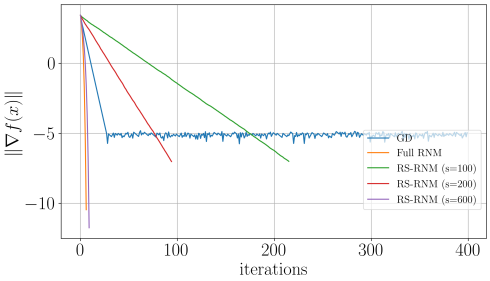

Figure 5 Iterations versus $\|\nabla f(x)\|$ ($\log_{10}$-scale) for low rank Rosenbrock function

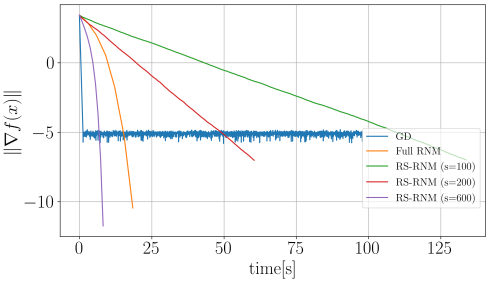

Figure 6 Computation time versus $\|\nabla f(x)\|$ ($\log_{10}$-scale) for low rank Rosenbrock function

Figures 5 and 6 show experiments for $n = 3000$ and $r = 500$. We selected three values for s, two ($s = 100, 200$) smaller than r and one ($s = 600$), larger than r. The results confirm the results of Section 5: when $s > r$ we have local superlinear convergence, otherwise the convergence is only linear locally.

6.3 Convolutional neural network

We tested our method on a micro Convolutional Neural Network (CNN) using the MNIST dataset in [13]. We used the cross-entropy loss function $m = 256$ images. Our CNN is made of the following factors:

- one convolutional layer (1 input channel, 1 output channel, kernel size 3),
- a ReLU activation,
- a max pooling layer (kernel size 2),
- a fully connected layer mapping the flattened feature vector to 10 classes.

This setup is intended to demonstrate the differences between the three methods in a controlled, small-scale scenario. This problem is formulated as

$$\min_{w\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\mathcal{M}(w, x_i), y_i),$$

where (x_i, y_i) denotes the MNIST dataset with $x_i \in \mathbb{R}^{784}$ and $y_i \in \{0, 1\}^{10}$ ($m = 256$), $\mathcal{L}$ denotes the Cross Entropy Loss function, and $\mathcal{M}$ denotes the CNN with $n = 1710$ parameters. The parameters in Section 3 are fixed as follows:

$$c_1 = c_1' = 2,\ c_2 = c_2' = 1,\ \gamma = \gamma' = 0.5,\ \alpha = 0.3,\ \beta = 0.5,\ s \in \{100, 200, 500\}.$$

The results are show in Figures 7 and 8. We notice that our method outperforms GD which is stuck at some stationary point and RNM which is to slow to converge.

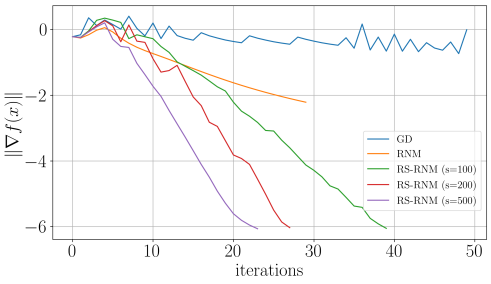

Figure 7 Iterations versus $\|\nabla f(w)\|$ ($\log_{10}$-scale) for CNN with the MNIST dataset

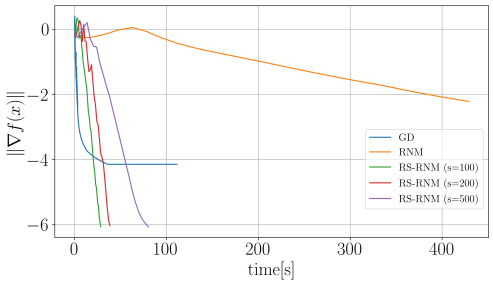

Figure 8 Computation time versus $\|\nabla f(w)\|$ ($\log_{10}$-scale) for CNN with the MNIST dataset

6.4 Choice of s

In the special case where the Hessian truly has low-rank structure, setting s to this value can substantially speed up convergence, provided the rank is not prohibitively large. However, in more general problems, especially where the Hessian does not exhibit pronounced low-rank properties or its effective rank is unknown, preselecting s is more challenging. One might try to start with some constant value of s and increasing it gradually since the best s ultimately depends on problem-specific characteristics and computational resources.

7 Conclusions

Random projections have been applied to solve optimization problems in suitable lower-dimensional spaces in various existing works. In this paper, we proposed the randomized subspace regularized Newton method (RS-RNM) for a non-convex twice differentiable function in the expectation that a framework for the full-space version [39, 40] could be used; indeed, we could prove the stochastic variant of the same order of iteration complexity, i.e., the global complexity bound of the algorithm: the worst-case iteration number m that achieves $\min_{k=0,\dots,m-1} \|\nabla f(x_k)\| \leq \varepsilon$ is $O(\varepsilon^{-2})$ when the objective function has Lipschitz Hessian. On the other hand, although RS-RNM uses second-order information similar to the regularized Newton method having a super-linear convergence, we proved that it is not possible, in general, to achieve local super-linear convergence and that local linear convergence is the best rate we can hope for in general. We were however able to prove super-linear convergence in the particular case where the Hessian is rank deficient at a local minimizer. In this paper we choose to thoroughly investigate local convergence rate for the Newton-based method. One could possibly, in a future work, extend these results to a state-of-the-art second order iterative method and compare the resulting subspace method with other state-of-the-art algorithms, as [19, 47, 48].

References

1 Roman Andreev. A note on the norm of oblique projections. *Appl. Math. E-Notes*, 14:43–44, 2014.
2 Jonathan T. Barron. A general and adaptive robust loss function. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4331–4339. IEEE, 2019.
3 James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(2):381–305, 2012.
4 Leon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.
5 Coralia Cartis, Estelle Massart, and Adilet Otemissov. Bound-constrained global optimization of functions with low effective dimensionality using multiple random embeddings. *Math. Program.*, 198(1(A)):997–1058, 2023.
6 Coralia Cartis, Estelle Massart, and Adilet Otemissov. Global optimization using random embeddings. *Math. Program.*, 200:781–829, 2023.
7 Coralia Cartis and Adilet Otemissov. A dimensionality reduction technique for unconstrained global optimization of functions with low effective dimensionality. *Inf. Inference*, 11(1):167–201, 2021.
8 Miantao Chao, Boris S. Mordukhovich, Zijian Shi, and Jin Zhang. Coderivative-Based Newton Methods with Wolfe Linesearch for Nonsmooth Optimization. `https://arxiv.org/abs/2407.02146`, 2024.
9 Long Chen, Xiaozhe Hu, and Huiwen Wu. Randomized fast subspace descent methods. `https://arxiv.org/abs/2006.06589`, 2020.

10 Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM J. Sci. Comput.*, 36(4):A1500–A1524, 2014.

11 Cartis Coralia, Fowkes Jaroslav, and Shao Zhen. A Randomised Subspace Gauss-Newton Method for Nonlinear Least-Squares. In *Thirty-seventh International Conference on Machine Learning, 2020. Workshop on Beyond First Order Methods in ML Systems*. 2020.

12 Cartis Coralia, Fowkes Jaroslav, and Shao Zhen. Randomised subspace methods for non-convex optimization, with applications to nonlinear least-squares. `https://arxiv.org/abs/2211.09873v1`, 2022.

13 Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.*, 29(6):141–142, 2012.

14 Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-Universal Regularized Newton Method. *SIAM J. Optim.*, 34(1):27–56, 2024.

15 Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. University of California, Irvine, School of Information and Computer Sciences, `http://archive.ics.uci.edu/ml`.

16 Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Found. Comput. Math.*, 12:229–262, 2012.

17 Lukas P. Fröhlich, Edgar D. Klenske, Christian G. Daniel, and Melanie N. Zeilinger. Bayesian optimization for policy search in high-dimensional systems via automatic domain selection. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 757–764. IEEE, 2019.

18 Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: Randomized Subspace Newton. *Adv. Neural Inf. Process. Syst.*, 32:616–625, 2019.

19 Serge Gratton, Sadok Jerad, and Philippe L. Toint. Yet another fast variant of Newton's method for nonconvex optimization. *IMA J. Numer. Anal.*, 45(2):971–1008, 2025.

20 Dmitry Grishchenko, Franck Iutzeler, and Jérôme Malick. Proximal gradient methods with adaptive subspace sampling. *Math. Oper. Res.*, 46(4):1303–1323, 2021.

21 Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. `https://arxiv.org/abs/1812.04754`, 2018.

22 Filip Hanzely, Nikita Doikov, Peter Richtárik, and Yurii Nesterov. Stochastic subspace cubic Newton method. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4027–4038. PMLR, 2020.

23 Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proceedings of the 31st International Conference on Machine Learning*, pages 754–762. PMLR, 2014.

24 Rujun Jiang and Xudong Li. Holderian Error Bounds and K.L. Inequality for the Trust Region Subproblem. *Math. Oper. Res.*, 47(4):3025–3050, 2022.

25 William Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In G. Hedlund, editor, *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

26 Christopher G. Knight, Sylvia H. E. Knight, Neil Massey, Tolu Aina, Carl Christensen, Dave J. Frame, Jamie A. Kettleborough, Andrew Martin, Stephen Pascoe, Ben Sanderson, David A. Stainforth, and Myles R. Allen. Association of parameter, software, and hardware variation with large-scale behavior across 57,000 climate models. *Proc. Natl. Acad. Sci. USA*, 104(30):12259–12264, 2007.

27 Dmitry Kovalev, Robert Gower, Peter Richtárik, and Alexander Rogozin. Fast linear convergence of randomized BFGS. `https://arxiv.org/abs/2002.11337`, 2020.

28 David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Comput. Optim. Appl.*, 79(2):339–368, 2021.

29 David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Comput. Optim. Appl.*, 79(2):339–368, 2021.

30 Jonathan Lacotte, Mert Pilanci, and Marco Pavone. *High-Dimensional Optimization in Adaptive Random Subspaces*, pages 10847–10857. Curran Associates, Inc., 2019.

31 Boris S. Mordukhovich, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A globally convergent proximal Newton-type method in nonsmooth convex optimization. *Math. Program.*, 198(1):899–936, 2023.

32 Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108:177–205, 2006.

33 Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. `https://arxiv.org/abs/1811.07062`, 2018.

34 Lindon Roberts and Clément W. Royer. Direct search based on probabilistic descent in reduced spaces. *SIAM J. Optim.*, 33(4):3057–3082, 2023.

35 Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Commun. Pure Appl. Math.*, 62(12):1707–1739, 2009.

36 Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of

over-parametrized neural networks. https://arxiv.org/abs/1706.04454, 2017.

37 Zhen Shao. On random embeddings and their application to optimisation. https://arxiv.org/abs/2206.03371, 2022.

38 Sebastian U. Stich, Christian L. Müller, and Bernd Gärtner. Optimization of Convex Functions with Random Pursuit. *SIAM J. Optim.*, 23(2):1284–1309, 2013.

39 Kenji Ueda and Nobuo Yamashita. Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization. *Appl. Math. Optim.*, 62(1):27–46, 2010.

40 Kenji Ueda and Nobuo Yamashita. A regularized Newton method without line search for unconstrained optimization. *Comput. Optim. Appl.*, 59(1-2):321–351, 2014.

41 Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018.

42 Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Feitas. Bayesian optimization in a billion dimensions via random embeddings. *J. Artif. Intell. Res.*, 55:361–387, 2016.

43 Stephen J. Wright. Coordinate descent algorithms. *Math. Program.*, 151(1):3–34, 2015.

44 Peng Xu, Fred Roosta, and Michael W. Mahoney. Second-order optimization for non-convex machine learning: an empirical study. In *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*, pages 199–207. 2020.

45 Yuya Yamakawa and Nobuo Yamashita. Convergence analysis of a regularized Newton method with generalized regularization terms for unconstrained convex optimization problems. *Appl. Math. Comput.*, 491: article no. 129219, 2025.

46 Zhewei Yao. *Efficient second-order methods for non-convex optimization and machine learning*. PhD thesis, UC Berkeley, 2021. https://escholarship.org/uc/item/0431q1ws.

47 Yuhao Zhou, Jintao Xu, Chenglong Bao, Chao Ding, and Jun Zhu. A Regularized Newton Method for Nonconvex Optimization with Global and Local Complexity Guarantees. https://arxiv.org/abs/2502.04799, 2025.

48 Hong Zhu and Yunhai Xiao. A hybrid inexact regularized Newton and negative curvature method. *Comput. Optim. Appl.*, 88(3):849–870, 2024.