Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Flexible modification of Gauss–Newton method and its stochastic extension

Nikita Yudin[1, 2], Alexander Gasnikov [1, 3]

submitted: February 17, 2021

[1] Moscow Institute of Physics and Technology
9 Institutskiy Pereulok
141701 Dolgoprudny, Moscow Region
Russian Federation
E-Mail: nikydin96@gmail.com

[2] Dorodnicyn Computing Center
Federal Research Center "Computer Science and Control"
of Russian Academy of Sciences
42 Vavilova Street
119991 Moscow
Russian Federation

[3] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: alexander.gasnikov@wias-berlin.de

No. 2813
Berlin 2021

2020 *Mathematics Subject Classification.* 90C30, 90C25, 90C26, 68Q25.

Key words and phrases. Systems of nonlinear equations, empirical risk minimization, Gauss–Newton method, trust region methods, non–convex optimization, inexact proximal mapping, inexact oracle, stochastic optimization, stochastic approximation, overparametrized model, weak growth condition, Polyak–Lojasiewicz condition, complexity estimate.

The research of A. Gasnikov was funded by Math+ AA4-2 Scholarship in Optimization.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Flexible modification of Gauss–Newton method and its stochastic extension

Nikita Yudin, Alexander Gasnikov

Abstract

This work presents a novel version of recently developed Gauss–Newton method for solving systems of nonlinear equations, based on upper bound of solution residual and quadratic regularization ideas. We obtained for such method global convergence bounds and under natural non–degeneracy assumptions we present local quadratic convergence results. We developed stochastic optimization algorithms for presented Gauss–Newton method and justified sub–linear and linear convergence rates for these algorithms using weak growth condition (WGC) and Polyak–Lojasiewicz (PL) inequality. We show that Gauss–Newton method in stochastic setting can effectively find solution under WGC and PL condition matching convergence rate of the deterministic optimization method. The suggested method unifies most practically used Gauss–Newton method modifications and can easily interpolate between them providing flexible and convenient method easily implementable using standard techniques of convex optimization.

1 Introduction

1.1 Motivation

We consider the problem of solving *systems of nonlinear equations*, which is one of the most fundamental in numerical methods. Corresponding problems are widespread among various works and monographs dedicated to numerical methods and optimization methods [26, 23, 22, 9]. The general form of system of nonlinear equations is defined via multidimensional mapping $F: \mathbb{R}^n \to \mathbb{R}^m$:

$$F(x) = \mathbf{0}_m,\ \mathbf{0}_m = (0,\ \ldots, 0)^T. \tag{1}$$

The next minimization problem of merit function is considered as a relaxation of the problem of solving systems of equations:

$$\min_{x \in \mathbb{R}^n} \left\{ f_1(x) \stackrel{\text{def}}{=} \|F(x)\| \right\}, \tag{2}$$

where $\|\ \|$ is the standard Euclidean norm (it can be straightforwardly generalized to other types of merit functions). This is quite typical way of dealing with problems like (1) [12, 15, 3, 5, 25]. The most standard way to solve (2) is to perform direct minimization of

$$f_2(x) \stackrel{\text{def}}{=} (f_1(x))^2,$$

which can cause some numerical instability and losses of performance, e.g. in case of linear F this transformation leads to squaring of the condition number of system (1). The direct optimization of f_2 is usually considered within *trust region methods* and *quasi-Newton methods* using variety of heuristics [30, 8, 28, 6]. However, it is possible to alleviate usage of iterative minimization schemes by

DOI 10.20347/WIAS.PREPRINT.2813 Berlin 2021

applying the Gauss–Newton method to the original problem (2), in which every iteration represents the following auxiliary optimization task

$$\begin{aligned} &\min_{h\in\mathbb{R}^n}\left\{\left\|F(x)+F^{'}(x)h\right\| :\ x+h\in \mathrm{D}(x)\right\},\\ &F^{'}(x)\stackrel{\text{def}}{=}\left(\frac{\partial F_i}{\partial x_j}(x)\right)_{i,j=1}^{m,n}\in\mathbb{R}^{m\times n}\text{ — Jacobian,}\end{aligned} \tag{3}$$

so we require smoothness for each function F_i, $i\in\{1,\ \ldots,m\}$, $\mathrm{D}(x)$ is an appropriate neighborhood of point $x\in\mathbb{R}^n$. This scheme allows us to optimize (2) with local quadratic speed under some natural non-degeneracy conditions [23].

Our work establish different approach to problem (2). First of all we define normalized merit function, and it stands for the standard Euclidean norm divided by square root of the number of coordinates in this norm. Then, we construct iterative scheme in which the problem (3) is replaced by minimization problem of parameterized *local upper model* of the introduced merit function. Our upper model represents development of the idea of quadratic regularization of functionals, so auxiliary problem in our scheme can be seen as some parametrized proximal mapping. Our local model unifies previously introduced models [20, 21] and has undoubtedly convenient ability to interpolate between them smoothly.

We also consider different variants of relation between task dimensions m and n: $m\leq n$ and $m>n$. The first case is related to the classical setting for problems of solving systems of nonlinear equations. The second case is usually described as the least–squares regression problem. For both cases we established global and local convergence properties, for each case we developed stochastic algorithms to solve (2). The whole analysis performed is applicable to various *empirical risk minimization* problems, and optimized functional f_2 possesses *weak growth condition* (WGC) [27, 29, 1]. The WGC states the majorization of squared norm of the gradient of f_2 by proportional to f_2 function value, in stochastic setting the gradient is replaced by its stochastic estimate and expectation is taken. Besides WGC we consider *Polyak–Lojasiewicz condition* (PL) [24], which forces domination of squared norm of the gradient of f_2 over f_2 value multiplied by some constant, so in stochastic setting PL condition is satisfied for expected squared norm of stochastic gradient and f_2 value. WGC and PL condition combined in case of $m\leq n$ allow us to establish the existence of solution for the problem (1), moreover we proved the existence of stochastic iterative scheme with arbitrary batch size, which converges linearly to the solution of (1). The existence of such schemes is deeply connected to properties of *overparameterized models* in statistical learning theory, these properties are usually called *interpolation conditions* [2, 19, 11, 10, 17, 18, 31, 16].

1.2 Main results

Our main contribution consists of designed algorithms and its analysis in both deterministic and stochastic settings of Gauss–Newton method, we consider different relations between the most important parameters of related tasks and offer a solution for each case. Our contribution is summarized as follows:

- We develop general Gauss–Newton method with inexact proximal map, our analysis has convergence guarantees for provided algorithm.
- We characterize difference between convergence types for developed methods. We elaborate conditions for sublinear, linear and superlinear convergence within developed Gauss–Newton framework.

- We propose stochastic algorithms for solving merit optimization problems and derive convergence conditions for each algorithm.
- We establish existence of stochastic algorithms with convergence rate of the deterministic optimization method under WGC and PL condition.

1.3 Contents

The next subsection recall the most used and crucial mathematical terms and tools for theoretical background. Section 2 describes Gauss–Newton framework with inexact oracle and describes its convergence. Section 3 states stochastic algorithms for Gauss–Newton framework. Section 4 is dedicated to analysis of proposed stochastic algorithms. Section 5 demonstrates experimental results of developed algorithms. All proofs and auxiliary discussion are placed into Appendix (Supplementary material).

1.4 Notation

Let us denote finite Euclidean space using letter E (we also equally use symbols subscripts to denote other Euclidean spaces) with standard Euclidean norm $\|\ \|$. Denote Euclidean spaces E_1 with $\dim(E_1) = n$ and E_2 with $\dim(E_2) = m$. The dual space for E is denoted as E^* and represents the space of linear functions over E, the value of function $u \in E^*$, evaluated at point $x \in E$, equals inner (scalar) product $\langle u, x\rangle$. Norms $\|x\|, x \in E$ and $\|u\|, u \in E^*$ are connected via following relation:

$$\begin{cases} \|x\| = \max\limits_{u \in E^*} \{\langle u, x\rangle : \ \|u\| \leq 1\}; \\ \|u\| = \max\limits_{x \in E} \{\langle u, x\rangle : \ \|x\| \leq 1\}. \end{cases}$$

Consequently, these relations state Cauchy–Schwarz inequality: $\langle u, x\rangle \leq \|u\|\|x\|$.

For a smooth function $f : E_1 \to E_2$ we denote first and second derivatives, evaluated at $x \in E_1$: $\nabla_x f(x)$ and $\nabla_x^2 f(x)$ respectively (in case of unambiguity we drop subscripted x). For $E_2 \equiv \mathbb{R}$ first and second derivatives are called gradient and hessian respectively. Note that $\nabla f(x) \in E_1^*$, $\nabla^2 f(x) : E_1 \to E_1^*$ is a self–adjoint operator.

Further, for linear operator $A : E_1 \to E_2$ we denote its adjoint $A^* : E_2^* \to E_1^*$:

$$\langle u, Ax\rangle = \langle A^* u, x\rangle, \ u \in E_2^*, x \in E_1.$$

Introduce for linear operator $A : E_1 \to E_2$ its *operator norm* as maximal singular value $\sigma_{\max}(A)$:

$$\begin{aligned} \|A\| = \sigma_{\max}(A) = \max_{x \in E_1} \{\|Ax\| : \ \|x\| \leq 1\} = \\ = \sqrt{\lambda_{\max}(AA^*)} = \sqrt{\lambda_{\max}(A^*A)}, \end{aligned}$$

where $\lambda_{\max}(\cdot)$ is maximal eigenvalue. In addition, for operator A with corresponding matrix $\left(a_{ij}\right)_{i,j=1}^{m,n}$ we denote Frobenius norm as $\|A\|_F$:

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{m,n} |a_{ij}|^2} = \sqrt{\mathrm{Tr}\,(AA^*)} = \sqrt{\mathrm{Tr}\,(A^*A)}.$$

Clearly, $\|A\| \leq \|A\|_F$ according to the property of trace $\mathrm{Tr}(\cdot)$ of self–adjoint operator. We also define minimal singular value for operator A:

$$\sigma_{\min}(A) = \min_{x \in E_1} \{\|Ax\| : \ \|x\| \leq 1\}.$$

For multidimensional map $F : E_1 \to E_2$ we introduce Jacobian $F^{'}(x)$, evaluated at a point $x \in E_1$ using linear operator from E_1 to E_2:

$$F^{'}(x)h = \lim_{t \to 0} \left(\frac{1}{t} (F(x+th) - F(x)) \right) \in E_2, \ h \in E_1.$$

For linear self–adjoint operators and its corresponding matrices we define partial order on positive semi–definite cone:

$$\begin{aligned} A \preceq A_1, \ A_1 \succeq A, \ A : E \to E^*, \ A_1 : E \to E^* \Leftrightarrow \\ \Leftrightarrow \langle (A_1 - A)x, \ x \rangle \geq 0, \ \forall x \in E; \end{aligned}$$

$$\begin{aligned} B \preceq B_1, \ B_1 \succeq B, \ B : E^* \to E, \ B_1 : E^* \to E \Leftrightarrow \\ \Leftrightarrow \langle u, \ (B_1 - B)u \rangle \geq 0, \ \forall u \in E^*. \end{aligned}$$

Notice that it is easy to establish for linear operator $A : E_1 \to E_2$ these relations:

$$\begin{cases} AA^* \succeq \sigma_{\min}(A^*)^2 I_{\dim(E_2)}; \\ A^*A \succeq \sigma_{\min}(A)^2 I_{\dim(E_1)}. \end{cases}$$

Denote set of integers form 1 to m inclusively as $\overline{1,m}$. Notation $f(x) = \mathrm{O}(h(x))$ means upper estimate of function f using function h up to positive constant and possible polylogarithmic factors. In similar manner $f(x) = \Omega(h(x))$ defines lower estimate of f using function h up to positive constant and possible polylogarithmic factors. Finally, we introduce

$$f^* = \min_{x \in E_1} f(x), \ g^*(y) = \min_{x \in E_1} g(x,y),$$

to define minimal values w.r.t. x for functions f and g respectively.

2 Modified Gauss–Newton method

2.1 Local upper model

Let us restate the problem of finding solution $x^* \in E_1$ of the smooth nonlinear system of equations:

$$F(x) = \mathbf{0}_m, \tag{4}$$

where $F : E_1 \to E_2$ is smooth multidimensional map with Jacobian $F^{'}(x), \ x \in E_1$. To estimate closeness to the solution of system of equations (4) we consider the following merit function depending on $\hat{F}(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{m}} F(x)$:

$$\hat{f}_1(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{m}} \|F(x)\| = \left\|\hat{F}(x)\right\|.$$

Clearly it is possible to solve (4) using $\hat{f}_1(x)$ in the following optimization problem:

$$\hat{f}_1^* = \min_{x \in E_1} \left\{ \hat{f}_1(x) = \frac{1}{\sqrt{m}} \|F(x)\| = \frac{1}{\sqrt{m}} \left\| (F_1(x), \ldots, F_m(x))^* \right\| \right\}. \tag{5}$$

Existence of solution of (4) is equivalent to $\hat{f}_1^* = \hat{f}_1(x^*) = 0$. We consider an iterative procedure to construct solution of (5), based on minimization of *local model*:

$$\phi(x,y) \overset{\text{def}}{=} \left\| \hat{F}(x) + \hat{F}'(x)(y-x) \right\|, \ (x,y) \in E_1^2,$$
$$\hat{F}'(x) = \frac{1}{\sqrt{m}} F'(x).$$

Classical Gauss–Newton method uses the following mapping at each iteration $k \in \mathbb{Z}_+$ to construct minimization scheme for (4) through sequence of convex problems:

$$x_{k+1} \in \operatorname*{Argmin}_{y \in E_1} \{\phi(x_k, y)\}.$$

However, simple additive regularization for ϕ allows us to establish global convergence properties in addition to local properties. In this section we consider a unified modification of local models considered in [20, 21]. First of all, we introduce some basic assumptions about the problem. Consider $\mathscr{F} \subseteq E_1$ — closed convex set with non-empty interior.

Assumption 1. *Multidimensional map $\hat{F}(x)$ is smooth on $\mathscr{F}$ with Lipschitz continuous Jacobian:*

$$\exists L_{\hat{F}} > 0 : \ \left\| \hat{F}'(y) - \hat{F}'(x) \right\|_F \leq L_{\hat{F}} \|y - x\|, \ \forall (x,y) \in \mathscr{F}^2.$$

Assumption 1 leads to the following Lipschitz property:

$$\left\| \hat{F}'(y) - \hat{F}'(x) \right\| \leq L_{\hat{F}} \|y - x\|, \ \forall (x,y) \in \mathscr{F}^2.$$

Denote level set $\mathscr{L}(v)$ of function $\hat{f}_1$:

$$\mathscr{L}(v) \overset{\text{def}}{=} \left\{ x : \hat{f}_1(x) \leq v \right\},$$

supposing that $\mathscr{F}$ is large enough:

$$\mathscr{L}(\hat{f}_1(x_0)) \subseteq \mathscr{F}, \ x_0 \in \mathscr{F} \text{ — initialization.}$$

Assumption 2. *Suppose the following PL condition is satisfied:*

$$\exists \mu > 0, \ \sigma_{\min}(\hat{F}'(x)^*) \geq \sqrt{\mu}, \ \forall x \in \mathscr{F}.$$

Assumption 2 is a PL–type condition because the inequality below is consequent from this assumption:

$$\left\| \nabla \hat{f}_2(x) \right\|^2 = 4 \left\| \hat{F}'(x)^* \hat{F}(x) \right\|^2 \geq 4\mu \hat{f}_2(x), \ \forall x \in \mathscr{F}.$$

Note that assumption 2 implicitly requires $\dim(E_1) \leq \dim(E_2)$. Based on these assumptions, we consider the following *general modification* of local model in the Gauss–Newton method recently introduced by Yurii Nesterov [21]:

$$\hat{f}_1(y) \leq \psi_{x,L,\tau}(y) \overset{\text{def}}{=} \frac{\tau}{2} + \frac{(\phi(x,y))^2}{2\tau} + \frac{L}{2} \|y - x\|^2, \ L \geq L_{\hat{F}},$$
$$\tau > 0, \ (x,y) \in \mathscr{F}^2.$$

Algorithm 1 General method of Normalized Squares with an inexact proximal map

Input: setting (6).
for $k = 0, 1,$ **to** $N-1$ **do**
  define $\tau_k = \mathscr{T}(x_k, L_k, \varepsilon_k)$, $\varepsilon_k = \mathscr{E}(k, x_k, x_{k-1})$.
  compute such $x_{k+1} \in E_1$,
  that $\psi_{x_k,L_k,\tau_k}(x_{k+1}) - \psi_{x_k,L_k,\tau_k}(T_{L_k,\tau_k}(x_k)) \leq \varepsilon_k$.
  if $\hat{f}_1(x_{k+1}) > \psi_{x_k,L_k,\tau_k}(x_{k+1})$ **then**
    define $L_k := \min\left\{2L_k, 2L_{\hat{F}}\right\}$ and return
    to step 3.
  end if
  $L_{k+1} = \max\left\{\frac{L_k}{2},\, L\right\}$.
end for
Output: x_N.

This local model provides a natural way of updating approximation of the solution:

$$T_{L,\tau}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in E_1} \{\psi_{x,L,\tau}(y)\};$$
$$T_{L,\tau}(x) = x - \left(\hat{F}'(x)^* \hat{F}'(x) + \tau L I_n\right)^{-1} \hat{F}'(x)^* \hat{F}(x).$$

2.2 Analysis of the scheme

The designed scheme of iterations in deterministic setting naturally unifies a variety of Gauss–Newton methods and possesses some convenient properties, such as strong convexity of the local model $\psi_{x,L,\tau}(y)$ w.r.t. y and strict convexity w.r.t. τ. It results into the uniqueness of optimal y at every iteration and even allows us to find approximation of the closest local model to our criterion w.r.t. τ. The developed optimization scheme is presented as algorithm 1, and because of f_1 structure we call the corresponding method as *Method of Normalized Squares*, the name we adopted from Yurii Nesterov's preprint [21]. This method requires objects outlined below:

$$\begin{cases} x_0 \in E_1,\ \mathscr{L}(\hat{f}_1(x_0)) \subseteq \mathscr{F} \text{ — initialization, } x_{-1} = x_0; \\ \mathscr{E}(\cdot) \text{ — error value function;} \\ N \in \mathbb{N} \text{ — number of outer iterations;} \\ L \text{ — local Lipschitz constant estimate, } L \in (0,\, L_{\hat{F}}],\ L_0 = L; \\ \mathscr{T}(\cdot) \text{ — function to specify } \tau. \end{cases} \tag{6}$$

Algorithm 1 is quite conceptual as it has some degrees of freedom in $(\tau_k,\ \varepsilon_k)$ selection, and this algorithm also exploits the idea of binary search of appropriate Lipschitz constant, which adaptively exploits geometric properties of a merit function surface. Note that the presented method uses an inexact oracle with some computational error ε_k of x_{k+1} and such error should be small enough to ensure $x_{k+1} \in \mathscr{F}$. We established global convergence properties for this method in listed below terms:

- norm of the proximal gradient mapping: $\|L_k\left(T_{L_k,\tau_k}(x_k) - x_k\right)\|$;
- local decrease:
  $\Delta_r(x_k) \stackrel{\text{def}}{=} \hat{f}_2(x_k) - \min\limits_{y \in E_1}\left\{(\phi(x_k, y))^2 :\ \|y - x_k\| \leq r\right\}$, $r > 0$; $\hat{f}_2(x) \stackrel{\text{def}}{=} \left(\hat{f}_1(x)\right)^2$, $x \in E_1$.

These values equivalently represent sets of stationary points in the following way:

- $\{x^* : x^* \in E_1,\ \|L(T_{L,\tau}(x^*) - x^*)\| = 0,\ \forall L > 0,\ \forall \tau > 0\}$;
- $\{x^* : x^* \in E_1,\ \Delta_r(x^*) = 0,\ \forall r > 0\}$.

Formally convergence to these stationary points is justified in theorem 1.

Theorem 1. *Suppose that assumption 1 is satisfied, $k \in \mathbb{N},\ r > 0$. Then Gauss–Newton method, implemented using scheme 1 with $\tau_k = \hat{f}_1(x_k)$, $\varepsilon_k = \varepsilon \geq 0$, has the following estimates:*

$$\begin{cases} \frac{8L_{\hat{F}}^2}{L}\left(\varepsilon + \frac{\left(\hat{f}_1(x_0) - \hat{f}_1(x_k)\right)}{k}\right) \geq \min\limits_{i \in \overline{0,k-1}} \left\{ \left\| 2L_{\hat{F}} \left(T_{2L_{\hat{F}}, \hat{f}_1(x_i)}(x_i) - x_i \right) \right\|^2 \right\}; \\ L_{\hat{F}}\left(\varepsilon + \frac{\left(\hat{f}_1(x_0) - \hat{f}_1(x_k)\right)}{k}\right) \geq \min\limits_{i \in \overline{0,k-1}} \left\{ 2\left(L_{\hat{F}} r\right)^2 \varkappa\left(\frac{\Delta_r(x_i)}{4\hat{f}_1(x_i) L_{\hat{F}} r^2}\right) \right\}; \end{cases}$$

where $\varkappa(t) = \frac{t^2}{2}\mathbb{1}_{\{t \in [0,1]\}} + \left(t - \frac{1}{2}\right)\mathbb{1}_{\{t > 1\}}$.

Theorem 1 states global sublinear rates of convergence, and does not imply existence of the solution of (5). The next theorem establish local superlinear convergence under natural non–degeneracy conditions for solvable systems (4).

Theorem 2. *Suppose that assumption 1 is satisfied, Jacobian is bounded: $\left\|\hat{F}'(x)\right\| \leq M_{\hat{F}}$ for all $x \in \mathscr{F}$, and the solution $x^* \in \mathscr{L}(\hat{f}_1(x_0)),\ \hat{F}(x^*) = \mathbf{0}_m$ with $\sigma_{\min}\left(\hat{F}'(x^*)\right) \geq \varsigma > 0$ exists. Then Gauss–Newton method 1 with $\tau_k = \hat{f}_1(x_k)$, $\varepsilon_k = 0$ in region*

$$\|x_k - x^*\| \leq \min\left\{ \frac{2\varsigma}{5L_{\hat{F}}},\ \frac{1}{12L_{\hat{F}}}\left(\left(3M_{\hat{F}} + 5\varsigma\right) - \sqrt{\left(3M_{\hat{F}} + 5\varsigma\right)^2 - 24\varsigma^2} \right) \right\},\ k \quad \in \mathbb{Z}_+$$

superlinearly converges

$$\|x_{k+1} - x^*\| \leq \frac{\frac{3L_{\hat{F}}\|x_k - x^*\|^2}{2} + \|x_k - x^*\|\sqrt{\hat{f}_1(x_k)L_k + \frac{L_{\hat{F}}^2\|x_k - x^*\|^2}{4}}}{\varsigma - L_{\hat{F}}\|x_k - x^*\|} \leq \|x_k - x^*\|,$$
$$x_{k+1} \in \mathscr{L}(\hat{f}_1(x_0)),\ \hat{f}_1(x_k) = \mathrm{O}\left(\|x_k - x^*\|\right).$$

Singular value bounds in theorem 2 require structural limitation $\dim(E_1) \leq \dim(E_2)$, so in case of $\dim(E_1) > \dim(E_2)$ there is no $\varsigma > 0$ exists, however, assumption 2 can be held for (4) according to the theorem below.

Theorem 3. *Assume that assumptions 1 and 2 are held for Gauss–Newton method 1 with $\tau_k = \hat{f}_1(x_k)$. Then any sequence $\{x_k\}_{k \in \mathbb{Z}_+}$ has the property:*

$$\hat{f}_1(x_{k+1}) \leq \varepsilon_k + \begin{cases} \frac{\hat{f}_1(x_k)}{2} + \frac{L_{\hat{F}}}{\mu}\hat{f}_2(x_k) \leq \frac{3}{4}\hat{f}_1(x_k), \text{ if } \hat{f}_1(x_k) \leq \frac{\mu}{4L_{\hat{F}}}; \\ \hat{f}_1(x_k) - \frac{\mu}{16L_{\hat{F}}}, \text{ otherwise.} \end{cases}$$

Theorem 3 reveals a quite important property of independence of linear convergence rate from μ for Gauss–Newton method.

3 Stochastic modification of Gauss–Newton method

3.1 Stochastic local model

We consider the only two sources of randomness: sampling of initialization $x_0 \in E_1$ and independent sampling of batches B_k of functions from (4) at each iteration of Gauss–Newton method. Batch B_k at each iteration k has size $|B_k| = b \in \{1, \ldots, m\}$ and is sampled without replacement, independently and from random uniform distribution q over subsets of size b:

$$B_k = \left\{ F_{i_j}(x) \middle| \; j \in \{1, \ldots, b\},\; i_j \in \{1, \ldots, m\} \right\}.$$

The whole set (finite population) of functions from F is denoted as

$$\mathscr{B} \stackrel{\text{def}}{=} \{F_i(x) |\; i \in \{1, \ldots, m\}\}.$$

Based on sampling strategy we define following stochastic estimates of $\hat{F}$ and $\hat{F}'$ w.r.t. batch B of size b:

$$\hat{G}(x,B) \stackrel{\text{def}}{=} \frac{1}{\sqrt{b}} \left(F_{i_1}(x), \ldots, F_{i_b}(x)\right)^*;$$
$$\hat{G}'(x,B) \stackrel{\text{def}}{=} \frac{1}{\sqrt{b}} \left(\nabla F_{i_1}(x), \ldots, \nabla F_{i_b}(x)\right)^*.$$

These multidimensional maps define corresponding stochastic optimization criteria:

$$\hat{g}_1(x,B) \stackrel{\text{def}}{=} \left\| \hat{G}(x,B) \right\|;$$
$$\hat{g}_2(x,B) \stackrel{\text{def}}{=} \left(\hat{g}_1(x,B)\right)^2.$$

And for such functions we are able to deduce the *stochastic local model*:

$$\hat{g}_1(y,B) \leq \hat{\psi}_{x,L,\tau}(y,B) \stackrel{\text{def}}{=} \frac{\tau}{2} + \frac{L}{2}\|y-x\|^2 + \frac{1}{2\tau}\left\| \hat{G}(x,B) + \hat{G}'(x,B)(y-x) \right\|^2,\; L \geq L_{\hat{F}},$$
$$(x,y) \in E_1^2,\; \tau > 0,\; B \subseteq \mathscr{B}.$$

This local model offers directly convenient proximal map for construction iterative optimization schemes:

$$x_{k+1} = \hat{T}_{L_k,\tau_k}(x_k, B_k) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in E_1} \left\{ \hat{\psi}_{x_k,L_k,\tau_k}(y, B_k) \right\},\; k \in \mathbb{Z}_+;$$
$$x_{k+1} = x_k - \left(\hat{G}'(x_k,B_k)^* \hat{G}'(x_k,B_k) + \tau_k L_k I_n \right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k).$$

3.2 Used assumptions

The stochastic version of Gauss–Newton method uses the next set of assumptions instead of the previous one to extend all main optimization criteria properties up to stochastic settings.

Assumption 3. *There are exist $L_{\hat{F}} > 0$, $l_{\hat{F}} > 0$, for which the following is satisfied*

$$\|\nabla F_i(x) - \nabla F_i(y)\| \leq L_{\hat{F}} \|x-y\|,$$
$$\left| (F_i(x))^2 - (F_i(y))^2 \right| \leq l_{\hat{F}} \|x-y\|,\; \forall (x,y) \in E_1^2,\; \forall i \in \overline{1,m}.$$

Unlike assumption 1, the lipschitzness is considered relatively F_i and not over the whole set of functions from (4).

Assumption 4. *Let $M_{\hat{G}} > 0$, for which $\left\|\hat{G}'(x,B)\right\| \leq M_{\hat{G}}$ for all $x \in E_1$ and $B \subseteq \mathscr{B}$, $|B| = b \in \overline{1,m}$. In case of $b = m$ exists $M_{\hat{F}} > 0$, for which $\left\|\hat{F}'(x)\right\| \leq M_{\hat{F}}$ at all $x \in E_1$.*

Assumption 5. *Let $P_{\hat{g}_1} > 0$, for which $\hat{g}_1(x,B) \leq P_{\hat{g}_1}$, for all $x \in E_1$ and $B \subseteq \mathscr{B}$, $|B| = b \in \overline{1,m}$. In case of $b = m$ exists $P_{\hat{f}_1} > 0$, for which $\left\|\hat{f}_1(x)\right\| \leq P_{\hat{f}_1}$ at all $x \in E_1$.*

Assumptions 4 and 5 mean lipschitzness of $(F_i(x))^2$ and $\hat{g}_2(x,B)$. By Lipschitz continuity the best (the least) value of the Lipschitz constant equals $\sup\limits_{x \in E_1} \{\|\nabla_x \hat{g}_2(x,B)\|\}$ [14] and this value is bounded:

$$\sup_{x \in E_1} \{\|\nabla_x \hat{g}_2(x,B)\|\} \leq \min\left\{l_{\hat{F}},\ 2M_{\hat{G}}P_{\hat{g}_1}\right\},\ \forall B \subseteq \mathscr{B},$$

because

$$|\hat{g}_2(z,B) - \hat{g}_2(y,B)| \leq \underbrace{\sup_{x \in E_1} \{\|\nabla_x \hat{g}_2(x,B)\|\}}_{\leq l_{\hat{F}} \text{ (lemma 7)}} \|z-y\|,\ \forall (y,z) \in E_1^2,\ \forall B \subseteq \mathscr{B}$$

and

$$\sup_{x \in E_1} \{\|\nabla_x \hat{g}_2(x,B)\|\} = \sup_{x \in E_1} \left\{\left\|2\hat{G}'(x,B)^* \hat{G}(x,B)\right\|\right\} \leq 2 \sup_{x \in E_1} \left\{\left\|\hat{G}'(x,B)\right\| \left\|\hat{G}(x,B)\right\|\right\} \leq$$
$$\leq 2M_{\hat{G}}P_{\hat{g}_1},\ \forall B \subseteq \mathscr{B}.$$

Assumption 6. *There is exists $\sigma > 0$, for which $\mathbb{E}_B\left[\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|^2\right] \leq \sigma^2$ at all $x \in E_1$ and $B \subseteq \mathscr{B}$, $|B| = 1$.*

Assumption 6 is automatically satisfied under assumption 5 and it is introduced due to convenience reason.

Assumption 7. *Let $\mu > 0$, for which $\hat{G}'(x,B)\hat{G}'(x,B)^* \succeq \mu I_b$ at all $x \in E_1$ and $B \subseteq \mathscr{B}$, $|B| = b \leq \min\{m,\ n\}$.*

Assumption 7 introduces lower bound for singular values of jacobian $\hat{G}'(x,B)^*$. Usually it is satisfied for cases with $m \leq n$, but theoretically it can be true for systems (4) with $m > n$.

3.3 Optimization scheme

According to stochastic local model $\hat{\psi}_{x_k,L_k,\tau_k}(y,B_k)$ the next update rule uses scaled descent direction to find another approximation of solution:

$$x_{k+1} = x_k - \eta_k \left(\hat{G}'(x_k,B_k)^* \hat{G}'(x_k,B_k) + \tau_k L_k I_n\right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k),\ \eta_k \geq 0. \tag{7}$$

Algorithm 2 General Three Stochastic Squares method with an inexact proximal map

Input: settings (9).
for $k = 0, 1,$ **to** $N-1$ **do**
  sample batch B_k of size b from $\mathscr{B}$.
  define $\tau_k = \mathscr{T}(x_k, L_k, B_k)$.
  compute $x_{k+1} \in E_1$ using (7).
  if $\hat{g}_1(x_{k+1}, B_k) > \hat{\psi}_{x_k, L_k, \tau_k}(x_{k+1}, B_k)$ **then**
    set $L_k := \min\left\{2L_k, \gamma L_{\hat{F}}\right\}$ and return to step 4.
  end if
  $L_{k+1} = \max\left\{\frac{L_k}{2},\ L\right\}$.
end for
Output: x_N.

The rule below is called *doubly stochastic* and is derived from stochastic local model, for which gradient and hessian are estimated using independently sampled batches while x_{k+1} is computed via scaled Newton method step:

$$x_{k+1} = x_k - \eta_k \left(\hat{G}'(x_k, \tilde{B}_k)^* \hat{G}'(x_k, \tilde{B}_k) + \tilde{\tau}_k L_k I_n\right)^{-1} \hat{G}'(x_k, B_k)^* \hat{G}(x_k, B_k),\ \eta_k \geq 0, \\ \tilde{B}_k \subseteq \mathscr{B} \text{ and } B_k \text{ are independent samples},\ \tilde{\tau}_k > 0. \tag{8}$$

The developed optimization scheme is a straightforward modification of scheme 1 and is present as algorithm 2. All presented stochastic Gauss–Newton methods are named with infix *three stochastic squares* because the local model is fully estimated on batches, unlike the *Method of Stochastic Squares* developed in [21], for which only the hessian of local model is batched. The whole process depends on the settings listed below:

$$\begin{cases} x_0 \in E_1 \text{ — initialization},\ x_{-1} = x_0; \\ N \in \mathbb{N} \text{ — number of outer iterations}; \\ \gamma \geq 1 \text{ — upper factor for } L_{\hat{F}} \text{ search}; \\ L \text{ — local Lipschitz constant estimate},\ L \in (0,\ \gamma L_{\hat{F}}], \\ \qquad L_0 = L; \\ \mathscr{T}(\cdot) \text{ — function to specify } \tau; \\ \mathscr{B} \text{ — population}; \\ b \in \overline{1, m} \text{ — size of batch } B_k \subseteq \mathscr{B},\ k \in \mathbb{Z}_+. \end{cases} \tag{9}$$

Scheme 2 uses more flexible upper bound of Lipschitz constant search, its typical value is no less than $2L_{\hat{F}}$ in case of unknown $L_{\hat{F}}$. Doubly stochastic step is used in another stochastic gradient–like strategy, based on the next settings:

$$\begin{cases} x_0 \in E_1 \text{ — initialization}; \\ N \in \mathbb{N} \text{ — number of iterations}; \\ \mathscr{T}(\cdot) \text{ — function to specify } \tilde{\tau} L; \\ \mathscr{B} \text{ — population}; \\ b, \tilde{b} \in \overline{1, m} \text{ — sizes of batches } B_k, \tilde{B}_k \subseteq \mathscr{B}, \\ \qquad\qquad \text{respectively},\ k \in \mathbb{Z}_+. \end{cases} \tag{10}$$

Algorithm 3 General Three Stochastic Squares method with a doubly stochastic step

Input: settings (10).
for $k = 0, 1,$ **to** $N-1$ **do**
  sample batches $B_k, \tilde{B}_k$ from $\mathscr{B}$ of corresponding sizes $b, \tilde{b}$.
  determine $\tilde{\tau}_k L_k = \mathscr{T}(x_k, \tilde{B}_k)$.
  compute $x_{k+1} \in E_1$ using (8).
end for
Output: x_N.

And corresponding scheme 3 does not contain adaptive Lipschitz constant search procedures, it only relies on step scale η_k.

4 Convergence analysis

4.1 Scaled step usage

Theorem 4 states general convergence result to approximate stationary point in mean.

Theorem 4. *Suppose assumptions 3, 4, 5, 6 are satisfied. Consider Stochastic Gauss–Newton method 2 with* $\tau_k = \hat{g}_1(x_k, B_k)$, $\eta_k \in [\eta, 1]$, $\eta \in (0, 1]$ *and some finite* $\tilde{\sigma} \geq \sigma$. *Then:*

$$\mathbb{E}\left[\min_{i \in \overline{0,k-1}} \left\|\nabla \hat{f}_2(x_i)\right\|^2\right] \leq \frac{8\left(M_{\hat{G}}^2 + \gamma P_{\hat{g}_1} L_{\hat{F}}\right)}{\eta(2-\eta)} \left(\frac{\mathbb{E}\left[\hat{f}_2(x_0)\right]}{k} + 2 l_{\hat{F}} \min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}}, \frac{M_{\hat{G}}}{L}\right\} \mathbb{1}_{\{b<m\}} + \right.$$
$$\left. + \tilde{\sigma}\sqrt{\frac{1}{b} - \frac{1}{m}}\right), \; k \in \mathbb{N}.$$

Expectation operator $\mathbb{E}[\cdot]$ *averages over all randomness in optimization procedure.*

The following proposition states linear convergence in mean in case of PL condition (assumption 7).

Theorem 5. *Suppose assumptions 3, 4, 5, 6, 7 are satisfied. Consider Stochastic Gauss–Newton method 2 with* $\tau_k = \hat{g}_1(x_k, B_k)$, $\eta_k \in [\eta, 1]$, $\eta \in (0, 1]$ *and some finite* $\tilde{\sigma} \geq \sigma$. *Then:*

$$\begin{cases} \mathbb{E}\left[\left\|\nabla \hat{f}_2(x_k)\right\|^2\right] \leq 4M_{\hat{G}}^2 \Delta_{k,b}; \\ \mathbb{E}\left[\hat{f}_2(x_k)\right] \leq \hat{f}_2^* + \Delta_{k,b}; \\ \Delta_{k,b} \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{f}_2(x_0)\right] \exp\left(-\dfrac{k\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} P_{\hat{g}_1} + \mu\right)}\right) + 4\left(l_{\hat{F}} \min\left\{\sqrt{\dfrac{2P_{\hat{g}_1}}{L}}, \dfrac{M_{\hat{G}}}{L}\right\} \mathbb{1}_{\{b<m\}} + \right. \\ \left. + \tilde{\sigma}\sqrt{\dfrac{1}{b} - \dfrac{1}{m}}\right) \left(\dfrac{\gamma L_{\hat{F}} P_{\hat{g}_1} + \mu}{\eta(2-\eta)\mu}\right), \; k \in \mathbb{Z}_+, \; b \in \overline{1, \min\{m,n\}}. \end{cases}$$

Expectation operator $\mathbb{E}[\cdot]$ *averages over all randomness in optimization procedure.*

4.2 Doubly stochastic step usage

The whole convergence of doubly stochastic step is justified using local model of $\hat{f}_2$, denoted as $\varphi_{x,l}$. Scheme 3 is used to optimize in mean the following $\varphi_{x,l}$ function:

$$\hat{f}_2(y) \leq \varphi_{x,l}(y) \stackrel{\text{def}}{=} \hat{f}_2(x) + \left\langle \nabla \hat{f}_2(x),\, y-x \right\rangle + \frac{l}{2}\|y-x\|^2,$$
$$l \geq l_{\hat{f}_2} \stackrel{\text{def}}{=} 2\left(M_{\hat{F}}^2 + L_{\hat{F}} P_{\hat{f}_1}\right),\ (x,y) \in E_1^2.$$

So, that's why the value $\tilde{\tau} L_k$ is less principal and it is enough for appropriate η_k just to ensure that $\tilde{\tau} L_k > 0$.

Now consider solving nonlinear equation regime in problem (5):

$$\dim(E_1) \geq \dim(E_2),\ n \geq m,$$

the structure of used functions allows us to use weak growth condition (WGC) alongside PL condition in current regime in context of assumptions 4 and 7:

$$\begin{cases} \|\nabla_x \hat{g}_2(x,B)\|^2 = \left\| 2\hat{G}'(x,B)^* \hat{G}(x,B) \right\|^2 \leq 4M_{\hat{G}}^2 \hat{g}_2(x,B); \\ \|\nabla_x \hat{g}_2(x,B)\|^2 = 4\left\| \hat{G}'(x,B)^* \hat{G}(x,B) \right\|^2 \geq 4\mu \hat{g}_2(x,B). \end{cases}$$

After averaging over batches $B \subseteq \mathscr{B}$ these inequalities lead to the following bounds:

$$4\mu \hat{f}_2(x) \leq \mathbb{E}_B\left[\|\nabla_x \hat{g}_2(x,B)\|^2\right] \leq 4M_{\hat{G}}^2 \hat{f}_2(x). \tag{11}$$

Together WGC and PL condition mean for function $\hat{f}_2$ satisfaction of so called *strong growth condition* (SGC):

$$\left\|\nabla \hat{f}_2(x)\right\|^2 = \left\| 2\hat{F}'(x)^* \hat{F}(x) \right\|^2 \geq \{\text{assumption 7}\} \geq$$
$$\geq 4\mu \hat{f}_2(x) \Rightarrow \{(11)\} \Rightarrow$$
$$\Rightarrow \mathbb{E}_B\left[\|\nabla_x \hat{g}_2(x,B)\|^2\right] \leq \frac{M_{\hat{G}}^2}{\mu}\left\|\nabla \hat{f}_2(x)\right\|^2.$$

These conditions forces all sampled gradients to be equal zero in stationary points, which are also global minimizers:

$$\nabla_{x^*} \hat{g}_2(x^*, B) = \mathbf{0}_n,\ B \subseteq \mathscr{B},\ x^* :\ F(x^*) = \mathbf{0}_m.$$

Thus, WGC and PL condition cause possibility to solve problem (5) in stochastic regime with arbitrary batch size and arbitrary accuracy, as the theorem below states.

Theorem 6. *Suppose that assumptions 3, 4, 5, 7 are satisfied. Consider Stochastic Gauss–Newton method 3 with $\tilde{\tau}_k \geq \tilde{\tau} > 0$, $L_k \geq L > 0$. Then, for sequence*

$$\eta_k = \frac{\mu\left(\tilde{\tau}_k L_k\right)^2}{\left(M_{\hat{G}}^2 + \tilde{\tau}_k L_k\right)\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2},\ k \in \mathbb{Z}_+$$

the next estimate holds

$$\mathbb{E}\left[\hat{f}_2(x_k)\right] \leq \mathbb{E}\left[\hat{f}_2(x_0)\right] \exp\left(-\frac{k}{\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2}\left(\frac{\mu\tilde{\tau}L}{M_{\hat{G}}^2 + \tilde{\tau}L}\right)^2\right),$$
$$k \in \mathbb{Z}_+.$$

In case of $\eta_k = 1,\ k \in \mathbb{Z}_+$ *convergence estimate is no better than*

$$\begin{cases} \mathbb{E}\left[\hat{f}_2(x_k)\right] & \leq \mathbb{E}\left[\hat{f}_2(x_0)\right] \exp\left(-\frac{k\mu^2}{M_{\hat{G}}^2}\left(\frac{2}{\mu+\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)c} - \frac{1}{\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)c^2}\right)\right); \\ & c \overset{\text{def}}{=} \frac{1}{3}\left(1+7\sqrt[3]{\frac{2}{47+3\sqrt{93}}}+\sqrt[3]{\frac{47+3\sqrt{93}}{2}}\right),\ k \in \mathbb{Z}_+. \end{cases}$$

Expectation operator $\mathbb{E}[\cdot]$ *averages over all randomness in optimization procedure.*

Theorem 6 states linear convergence for all considered cases. As corollary, the most optimal choice in the worst case scenario for $\tilde{\tau}$ and L is $\tilde{\tau}L \to +\infty$, which morphs Gauss–Newton step into gradient method step, if we use dynamic η_k. In case of $\eta_k = 1$ convergence speed is slower in the worst case scenario.

5 Experiments

We conduct numerical experiments to evaluate performance of algorithm 1, algorithm 2 and algorithm 3. The whole set of algorithms is implemented in Python 3.8 running on a Linux–based ASUS laptop with Intel Core i5–4200H CPU @ 2.80GHz $\times$ 4 processor and 16 Gb RAM. The estimated runtime for experiments is 6 hours, 5 minutes and 46 seconds. Details of our experiments are in supplementary material.

We consider three benchmark functions to test main features of presented methods. The main task is unconstrained minimization, which is achievable for smooth convex functions using equivalence between unconstrained minimization task and solving system of equations, which represents first order optimality conditions. More formally, it means $\nabla f(x) \equiv F(x)$, if we have to minimize function $f(x)$ using optimization of merit $\|F(x)\|$. So, we have to find a solution for this task:

$$\min_{x \in E_1} f(x).$$

But our benchmark functions are non–convex, so our solution of the system of equations $F(x) = \mathbf{0}_n$ can represent a local minimum point or even a saddle point.

We test the following three different doubly smooth functions $f(x), x = (x^1, \ldots, x^n)$:

- Nesterov–Skokov function [9]:
  $f_{NS}(x) = \frac{1}{4}\left(x^1 - 1\right)^2 + \sum\limits_{i=1}^{n-1}\left(x^{i+1} - 2\left(x^i\right)^2 + 1\right)^2$;
- Hat function: $f_H(x) = (\|x\|^2 - 1)^2$;
- PL function: $f_{PL}(x) = \|x\|^2 + 3\sum\limits_{i=1}^{n}\sin^2(x^i)$.

Clearly, in such conditions we always have $m = n$.

Function f_{NS} is one of the hardest to optimize because of its fluctuating landscape, achieved using superpositions with Chebyshev polynomials of first kind $P_2(x^i) = 2\left(x^i\right)^2 - 1$. Function f_H is non–convex and possesses quadratic growth property:

$$\exists \nu > 0 :\ f(x) - f^* \geq \frac{\nu}{2}\|x - \mathscr{P}(x)\|^2,\ \forall x \in E_1,$$
$$\mathscr{P} : E_1 \to E_1,$$

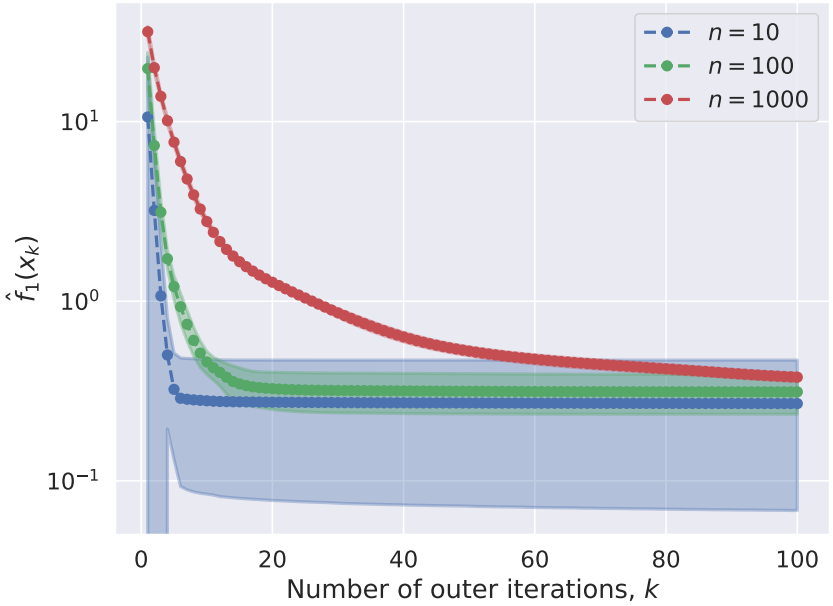

Figure 1: Performance of deterministic Gauss–Newton method on Nesterov–Skokov function

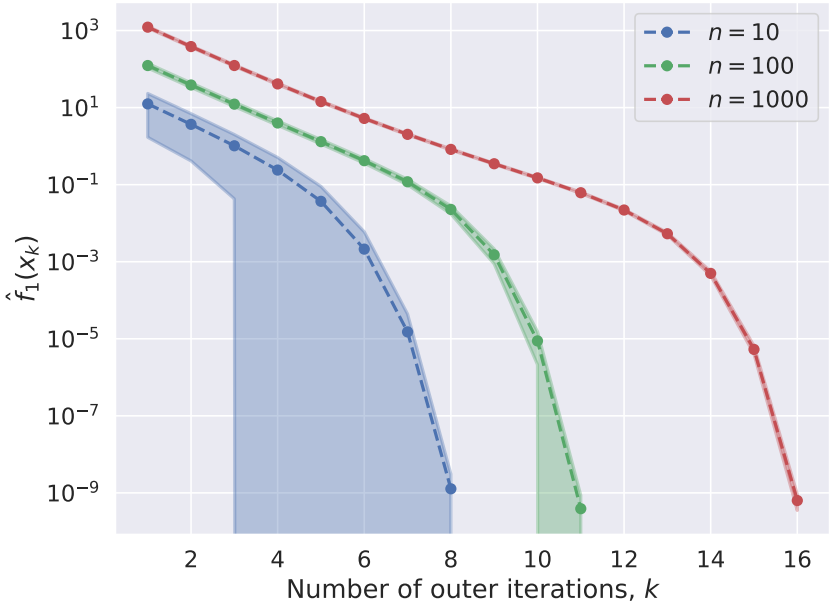

Figure 2: Performance of deterministic Gauss–Newton method on Hat function

where $\mathscr{P}$ is the projection of x onto set of global minimizers of f. Function f_{PL} is non–convex, bounded by two paraboloids and also satisfies quadratic growth property. All of three functions have global minimum equal 0.

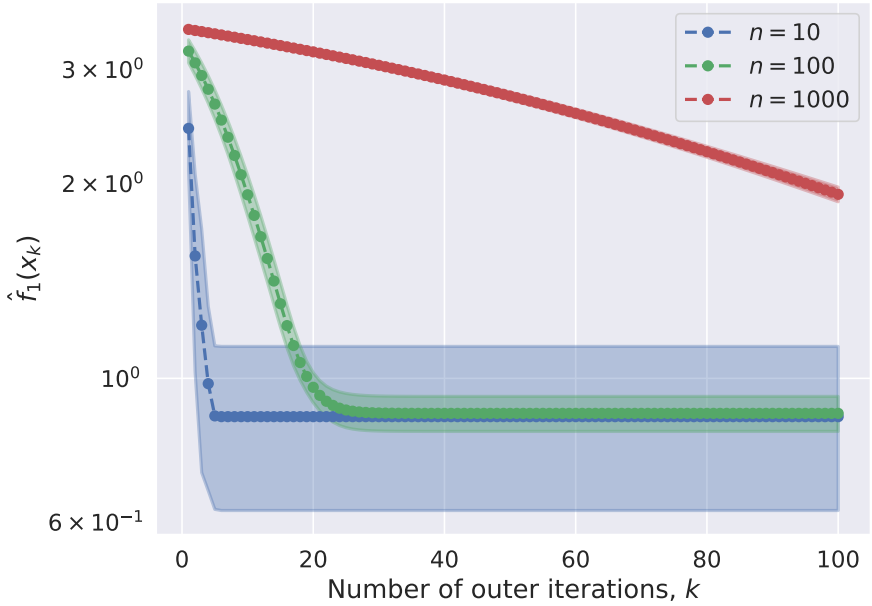

Figure 3: Performance of deterministic Gauss–Newton method on PL function

For deterministic Gauss–Newton method we use $\tau_k = \hat{f}_1(x_k)$ and test three different values of dimension n: 10, 100 and 1000. We use the exact proximal map: $\varepsilon_k \equiv 0$. All settings are averaged over 5 random initializations. Depicted uncertainty intervals have two standard deviations width.

All test runs show us the hardness of optimization f_{NS} (figure 1) dispite having a unique global minimum, the convergence speed is sublinear. f_{PL} (figure 3) demonstrates linear speed of convergence to a saddle point, however, trigonometric fluctuations slow down the whole process. And f_H (figure 2) shows the best properties to achieve even superlinear speed of convergence to the global minimum in later iterations demonstrating typical change of slope between linear and superlinear regions of convergence.

Results of stochastic algorithms performance are in supplementary material.

Acknowledgements

We express special thanks to Yurii Nesterov for the problem statement, which seeded the cause for our work to come up to the light. We also thank Dmitry Kamzolov and Pavel Dvurechensky for productive discussions done at various steps of writing this paper. Our work is dedicated to Yurii Nesterov 65 anniversary.

References

[1] Ahmad Ajalloeian and Sebastian U Stich. Analysis of sgd with biased gradient estimators. *arXiv preprint arXiv:2008.00051*, 2020.

[2] Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Neural Information Processing Systems (NIPS)*, 2011.

[3] Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical gauss-newton optimisation for deep learning. In *International Conference on Machine Learning*, pages 557–565. PMLR, 2017.

[4] Richard L. Burden and J. Douglas Faires. *Numerical Analysis*. The Prindle, Weber and Schmidt Series in Mathematics. PWS-Kent Publishing Company, Boston, fourth edition, 1989.

[5] Tianle Cai, Ruiqi Gao, Jikai Hou, Siyu Chen, Dong Wang, Di He, Zhihua Zhang, and Liwei Wang. Gram-gauss-newton method: Learning overparameterized neural networks for regression problems. *arXiv preprint arXiv:1905.11675*, 2019.

[6] Coralia Cartis and Lindon Roberts. A derivative-free gauss–newton method. *Mathematical Programming Computation*, 11(4):631–674, 2019.

[7] Jan JM Cuppen. A divide and conquer method for the symmetric tridiagonal eigenproblem. *Numerische Mathematik*, 36(2):177–195, 1980.

[8] Matilde Gargiani, Andrea Zanelli, Moritz Diehl, and Frank Hutter. On the promise of the stochastic generalized gauss-newton method for training dnns. *arXiv preprint arXiv:2006.02409*, 2020.

[9] Alexander Gasnikov. Universal gradient descent. *arXiv preprint arXiv:1711.00394*, 2017.

[10] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

[11] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

[12] Serge Gratton, Amos S Lawless, and Nancy K Nichols. Approximate gauss–newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106–132, 2007.

[13] Alston S Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM (JACM)*, 5(4):339–342, 1958.

[14] Matt Jordan and Alexandros G Dimakis. Exactly computing the local lipschitz constant of relu networks. *arXiv preprint arXiv:2003.01219*, 2020.

[15] Huu Le, Christopher Zach, Edward Rosten, and Oliver J Woodford. Progressive batching for efficient non-linear least squares. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[16] Chaoyue Liu and Mikhail Belkin. Mass: an accelerated stochastic method for over-parametrized learning. *arXiv preprint arXiv:1810.13395*, 2018.

[17] Nicolas Loizou, Sharan Vaswani, Issam Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. *arXiv preprint arXiv:2002.10542*, 2020.

[18] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[19] Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *arXiv preprint arXiv:1310.5715*, 2013.

[20] Yu Nesterov. Modified gauss–newton scheme with worst case guarantees for global performance. *Optimisation methods and software*, 22(3):469–483, 2007.

[21] Yu Nesterov. Flexible modification of gauss–newton method. *CORE Discussion Papers*, 2021.

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

[23] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[24] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[25] Yi Ren and Donald Goldfarb. Efficient subsampled gauss-newton and natural gradient methods for training neural networks. *arXiv preprint arXiv:1906.02353*, 2019.

[26] AA Samarskiı and AV Gulin. Numerical methods, 1989.

[27] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

[28] Marek J Śmietański. On a nonsmooth gauss–newton algorithms for solving nonlinear complementarity problems. *Algorithms*, 13(8):190, 2020.

[29] Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pages 2934–2992. PMLR, 2019.

[30] Christopher Thiele, Mauricio Araya-Polo, and Detlef Hohl. Deep neural network learning with second-order optimizers–a practical study with a stochastic quasi-gauss-newton method. *arXiv preprint arXiv:2004.03040*, 2020.

[31] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

SUPPLEMENTARY MATERIAL

A The proof of auxiliary results for Gauss–Newton method

The next lemma derives the local upper model for problem (5).

Lemma 1 ([21]). *Let $(x,y) \in \mathscr{F}^2, L \geq L_{\hat{F}}, \tau > 0$ and assumption 1 is satisfied. Then $\hat{f}_1(y) \leq \psi_{x,L,\tau}(y)$.*

Proof. We deduce an upper estimate for $\left\|\hat{F}(y) - \hat{F}(x) - \hat{F}'(x)(y-x)\right\|$:

$$\begin{aligned}
&\left\|\hat{F}(y) - \hat{F}(x) - \hat{F}'(x)(y-x)\right\| = \left\{\hat{F}(y) = \hat{F}(x) + \int\limits_0^1 \hat{F}'(x+t(y-x))(y-x)\,\mathrm{d}t\right\} = \\
&= \left\|\int\limits_0^1 \left(\hat{F}'(x+t(y-x)) - \hat{F}'(x)\right)(y-x)\,\mathrm{d}t\right\| \leq \{\|\cdot\| \text{ is convex, Jensen's inequality}\} \leq \\
&\leq \int\limits_0^1 \left\|\left(\hat{F}'(x+t(y-x)) - \hat{F}'(x)\right)(y-x)\right\|\mathrm{d}t \leq \int\limits_0^1 \left\|\hat{F}'(x+t(y-x)) - \hat{F}'(x)\right\|\|y-x\|\,\mathrm{d}t \leq \\
&\leq \{\text{assumption 1}\} \leq \int\limits_0^1 L_{\hat{F}}\|y-x\|^2 t\,\mathrm{d}t = \frac{L_{\hat{F}}}{2}\|y-x\|^2.
\end{aligned} \tag{12}$$

Consider an auxiliary inequality:

$$\begin{aligned}
\left(\sqrt{\frac{\tau}{2}} - \frac{1}{\sqrt{2\tau}}\left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\|\right)^2 &= \frac{\tau}{2} + \frac{1}{2\tau}\left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\|^2 - \\
&- \left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\| \geq 0 \Rightarrow \\
&\Rightarrow \frac{\tau}{2} + \frac{1}{2\tau}\left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\|^2 \geq \\
&\geq \left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\|.
\end{aligned} \tag{13}$$

Then for $\hat{f}_1$ we have

$$\begin{aligned}
\hat{f}_1(y) = \left\|\hat{F}(y)\right\| &= \left\|\hat{F}(y) - \hat{F}(x) - \hat{F}'(x)(y-x) + \hat{F}(x) + \hat{F}'(x)(y-x)\right\| \leq \\
&\leq \left\|\hat{F}(y) - \hat{F}(x) - \hat{F}'(x)(y-x)\right\| + \left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\| \leq \{\text{inequality from (12)}\} \leq \\
&\leq \frac{L_{\hat{F}}}{2}\|y-x\|^2 + \left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\| \leq \{\text{inequality from (13)}\} \leq \frac{\tau}{2} + \frac{L_{\hat{F}}}{2}\|y-x\|^2 + \\
&+ \frac{1}{2\tau}\left\|\hat{F}(x) + \hat{F}'(x)(y-x)\right\|^2 = \psi_{x,L_{\hat{F}},\tau}(y) \leq \psi_{x,L,\tau}(y).
\end{aligned}$$

□

Corollary 1.1. *For $\tau = \phi(x,y)$ the gap between $\hat{f}_1(y)$ and $\psi_{x,L,\tau}(y)$ is minimal, according to* (13).

The following lemma characterizes decrease relatively to the point of minimum of local model. The decrease is measured proportional to the squared norm of proximal gradient map.

Lemma 2. *Suppose that assumption 1 holds and* $x \in \mathscr{F}$, $T_{L,\tau}(x) \in \mathscr{F}$, $\tau > 0$, $L \geq L_{\hat{F}}$. *Then*

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \geq \frac{L}{2}\|T_{L,\tau}(x) - x\|^2.$$

Proof. Consider function

$$h(t) = \min_{y \in E_1} \left\{ \psi_{x,t^{-1},\tau}(y) \right\} = \min_{y \in E_1} \left\{ \frac{\tau}{2} + \frac{1}{2\tau} \left\| \hat{F}(x) + \hat{F}'(x)(y-x) \right\|^2 + \frac{1}{2t}\|y-x\|^2 \right\}.$$

The local model $\psi_{x,t^{-1},\tau}(y)$ is convex w.r.t. (τ, y, t) on convex set

$$\left\{ (y, \tau, t, \alpha) \in E_1 \times \mathbb{R}^3_+ : \|y-x\|^2 \leq \alpha t \right\}.$$

The function $h(t)$ has convex epigraph as the result of convex set projection, therefore, $h(t)$ is convex (Theorem 3.1.7, [22]). So we have for convex function

$$\begin{aligned}
&h(0) \geq h(t) + h'(t)(0-t) = h(t) - h'(t)t; \\
&h'(t) = \Bigg\langle \underbrace{\frac{1}{\tau}\hat{F}'(x)^* \left(\hat{F}(x) + \hat{F}'(x)\left(T_{t^{-1},\tau}(x) - x \right) \right) + \frac{1}{t}\left(T_{t^{-1},\tau}(x) - x \right)}_{= \nabla_y \psi_{x,t^{-1},\tau}(y) = \mathbf{0}_n \text{ as result of taking minimum w.r.t. } y}, \frac{\partial T_{t^{-1},\tau}(x)}{\partial t} \Bigg\rangle - \\
&\quad - \frac{1}{2t^2}\left\| T_{t^{-1},\tau}(x) - x \right\|^2 = -\frac{1}{2t^2}\left\| T_{t^{-1},\tau}(x) - x \right\|^2.
\end{aligned}$$

Using the main property of proximal map respectively t we have $\lim\limits_{t \to 0} \operatorname{argmin}\limits_{y \in E_1} \left\{ \psi_{x,t^{-1},\tau}(y) \right\} = x \Rightarrow$ $h(0) = \frac{\tau}{2} + \frac{\left\|\hat{F}(x)\right\|^2}{2\tau} = \frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau}$. Thus,

$$\begin{aligned}
\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} &\geq \psi_{x,t^{-1},\tau}(T_{t^{-1},\tau}(x)) + \frac{1}{2t}\left\| T_{t^{-1},\tau}(x) - x \right\|^2 \geq \{\text{lemma 1}\} \geq \hat{f}_1(T_{t^{-1},\tau}(x)) + \\
&+ \frac{1}{2t}\left\| T_{t^{-1},\tau}(x) - x \right\|^2 \Rightarrow \left\{ t^{-1} = L \right\} \Rightarrow \frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \geq \frac{L}{2}\|T_{L,\tau}(x) - x\|^2.
\end{aligned}$$

□

Corollary 2.1. *The results of lemma also mean the inequality below:*

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \psi_{x,L,\tau}(T_{L,\tau}(x)) \geq \frac{L}{2}\|T_{L,\tau}(x) - x\|^2,$$

which is true for $L > 0$ *and* $x \in E_1$.

Corollary 2.2. $T_{L,\tau}(x) = \operatorname{argmin}\limits_{y \in E_1} \{\psi_{x,L,\tau}(y)\}$ *has explicit form for* $L > 0$*:*

$$T_{L,\tau}(x) = x - \left(\hat{F}'(x)^* \hat{F}'(x) + \tau L I_n \right)^{-1} \hat{F}'(x)^* \hat{F}(x).$$

That's why $\lim\limits_{L\to+\infty} T_{L,\tau}(x) = x$ *and*

$$\begin{aligned}\frac{\tau}{2}+\frac{\hat{f}_2(x)}{2\tau}-\hat{f}_1(x) &\geq \frac{1}{2}\lim_{L\to+\infty}\left(L\|T_{L,\tau}(x)-x\|^2\right)= \\ &= \frac{1}{2}\lim_{L\to+\infty}\left(L\left\|\left(\hat{F}'(x)^*\hat{F}'(x)+\tau LI_n\right)^{-1}\hat{F}'(x)^*\hat{F}(x)\right\|^2\right)= \\ &= \frac{1}{2}\lim_{L\to+\infty}\left\|\left(\frac{1}{\sqrt{L}}\hat{F}'(x)^*\hat{F}'(x)+\tau\sqrt{L}I_n\right)^{-1}\hat{F}'(x)^*\hat{F}(x)\right\|^2=0.\end{aligned}$$

However, the value $\|L(T_{L,\tau}(x)-x)\|$ *converges to the norm of gradient of* $\psi_{x,L,\tau}(y)$ *w.r.t.* y *evaluated at* $y=x$ *when taking the limit in* $L\to+\infty$*:*

$$\begin{aligned}\lim_{L\to+\infty}\|L(T_{L,\tau}(x)-x)\| &= \lim_{L\to+\infty}\left\|L\left(\hat{F}'(x)^*\hat{F}'(x)+\tau LI_n\right)^{-1}\hat{F}'(x)^*\hat{F}(x)\right\|= \\ &= \lim_{L\to+\infty}\left\|\left(\frac{1}{L}\hat{F}'(x)^*\hat{F}'(x)+\tau I_n\right)^{-1}\hat{F}'(x)^*\hat{F}(x)\right\|=\left\|\frac{1}{\tau}\hat{F}'(x)^*\hat{F}(x)\right\|.\end{aligned}$$

The function $\|T_{L,\tau}(x)-x\|^2$ *is decreasing of* L *and* τ*.*

Corollary 2.3. *If we set* $\tau=\hat{f}_1(x)>0$*, then for* $x\in\mathscr{L}(\hat{f}_1(x))\subseteq\mathscr{F}$ *the proved inequality means* $T_{L,\hat{f}_1(x)}(x)\in\mathscr{L}(\hat{f}_1(x))$*:*

$$\begin{aligned}\frac{\tau}{2}+\frac{\hat{f}_2(x)}{2\tau}-\hat{f}_1(T_{L,\tau}(x)) &\geq \frac{L}{2}\|T_{L,\tau}(x)-x\|^2\Rightarrow\left\{\tau=\hat{f}_1(x)\right\}\Rightarrow\hat{f}_1(x)-\hat{f}_1(T_{L,\hat{f}_1(x)}(x))\geq \\ &\geq \frac{L}{2}\left\|T_{L,\hat{f}_1(x)}(x)-x\right\|^2\geq 0\Rightarrow\hat{f}_1(x)\geq\hat{f}_1(T_{L,\hat{f}_1(x)}(x))\Rightarrow \\ &\Rightarrow T_{L,\hat{f}_1(x)}(x)\in\mathscr{L}(\hat{f}_1(T_{L,\hat{f}_1(x)}(x)))\subseteq\mathscr{L}(\hat{f}_1(x)).\end{aligned}$$

The lemma below estimates local decrease of the optimized functional using $\Delta_r(x)$.

Lemma 3. *Let assumption 1 holds and* $x\in\mathscr{F}$*,* $T_{L,\tau}(x)\in\mathscr{F}$*,* $\tau>0$*,* $L\geq L_{\hat{F}}$*. Then, for any* $r>0$ *the next inequality holds*

$$\frac{\tau}{2}+\frac{\hat{f}_2(x)}{2\tau}-\hat{f}_1(T_{L,\tau}(x))\geq Lr^2\varkappa\left(\frac{\Delta_r(x)}{2\tau Lr^2}\right),$$

where

$$\begin{cases}\Delta_r(x)\overset{\text{def}}{=}\hat{f}_2(x)-\min\limits_{y\in E_1}\left\{(\phi(x,y))^2:\|y-x\|\leq r\right\};\\ \varkappa(t)\overset{\text{def}}{=}\begin{bmatrix}\frac{t^2}{2},\ t\in[0,\ 1];\\ t-\frac{1}{2},\ t>1.\end{bmatrix}\end{cases}$$

Proof. We introduce $h_r = \underset{h\in E_1}{\operatorname{argmin}} \left\{ (\phi(x,x+h))^2 : \|h\| \le r \right\}$. Express local model at $T_{L,\tau}(x)$:

$$\hat{f}_1(T_{L,\tau}(x)) \le \{\text{lemma 1}\} \le \psi_{x,L,\tau}(T_{L,\tau}(x)) \le$$
$$\le \min_{t\in[0,\,1]} \left\{ \frac{\tau}{2} + \frac{1}{2\tau} \left\| \hat{F}(x) + t\hat{F}'(x)h_r \right\|^2 + \frac{L}{2}(tr)^2 \right\} = \frac{\tau}{2} +$$
$$+ \min_{t\in[0,\,1]} \left\{ \frac{1}{2\tau} \left\| (1-t)\hat{F}(x) + t\left(\hat{F}(x) + \hat{F}'(x)h_r\right) \right\|^2 + \frac{L}{2}(tr)^2 \right\} \le \left\{ \|\cdot\|^2 \text{ convex} \right\} \le \frac{\tau}{2} +$$
$$+ \min_{t\in[0,\,1]} \left\{ \frac{(1-t)}{2\tau}\hat{f}_2(x) + \frac{t}{2\tau}(\phi(x,x+h_r))^2 + \frac{L}{2}(tr)^2 \right\} = \frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} +$$
$$+ \min_{t\in[0,\,1]} \left\{ \frac{-t}{2\tau}\Delta_r(x) + \frac{L}{2}(tr)^2 \right\} \Rightarrow \frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \ge Lr^2 \max_{t\in[0,\,1]} \left\{ \frac{\Delta_r(x)}{2\tau Lr^2}t - \frac{1}{2}t^2 \right\}.$$

The RHS of the inequality above hides under $\max$ operator a quadratic polynomial with negative coefficient at the highest degree term and with roots $t \in \left\{0,\ \frac{\Delta_r(x)}{\tau Lr^2}\right\}$, which means two cases for the computation of point of maximal value t^*: $\frac{\Delta_r(x)}{2r^2\tau L} \le 1$ and $\frac{\Delta_r(x)}{2r^2\tau L} > 1$. In the first case $t^* = \frac{\Delta_r(x)}{2\tau Lr^2}$, in the second one $t^* = 1$. The estimate obtained has the following expression:

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \ge Lr^2 \cdot \begin{bmatrix} \frac{1}{2}\left(\frac{\Delta_r(x)}{2\tau Lr^2}\right)^2, \text{ for } t^* = \frac{\Delta_r(x)}{2\tau Lr^2}; \\ \frac{\Delta_r(x)}{2\tau Lr^2} - \frac{1}{2}, \text{ for } t^* = 1. \end{bmatrix} \tag{14}$$

Define function $\varkappa(t) = \begin{bmatrix} \frac{t^2}{2},\ t \in [0,1]; \\ t - \frac{1}{2},\ t > 1. \end{bmatrix}$ Express the estimate (14) using this function:

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \ge Lr^2 \varkappa\left(\frac{\Delta_r(x)}{2\tau Lr^2}\right).$$

Note that $\hat{f}_2(x) \ge \Delta_\infty(x) \ge \Delta_r(x) \ge \Delta_0(x) = 0$ and $\varkappa(t) \ge 0$ by construction. □

Corollary 3.1. *The proved results hide the following inequality:*

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \psi_{x,L,\tau}(T_{L,\tau}(x)) \ge Lr^2 \varkappa\left(\frac{\Delta_r(x)}{2\tau Lr^2}\right),$$

which holds for $L > 0$ and $x \in E_1$. Moreover, for sufficiently small values of Lr^2, that $\frac{\Delta_r(x)}{2\tau Lr^2} \ge 1$, we have:

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \psi_{x,L,\tau}(T_{L,\tau}(x)) \ge \frac{\Delta_r(x)}{2\tau} - \frac{Lr^2}{2}.$$

For great values of Lr^2, for which $\frac{\Delta_r(x)}{2\tau Lr^2} \le 1$, we have different estimate:

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \psi_{x,L,\tau}(T_{L,\tau}(x)) \ge \frac{(\Delta_r(x))^2}{8\tau^2 Lr^2}.$$

Corollary 3.2. *For sufficiently great values of Lr^2, for which $\frac{\Delta_r(x)}{2\tau Lr^2} \le 1$, the obtained estimate simplifies:*

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \ge \frac{(\Delta_r(x))^2}{8\tau^2 Lr^2}.$$

For sufficiently small values of r, for which $\frac{\Delta_r(x)}{2\tau Lr^2} \geq 1$, the other estimate holds:

$$\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \geq \frac{\Delta_r(x)}{2\tau} - \frac{Lr^2}{2}.$$

In these inequalities the function $Lr^2 \varkappa\left(\frac{\Delta_r(x)}{2\tau Lr^2}\right)$ is decreasing of L and τ.

Corollary 3.3. *For $\tau = \hat{f}_1(x) > 0$ the obtained estimates for $x \in \mathscr{L}(\hat{f}_1(x)) \subseteq \mathscr{F}$ mean $T_{L,\hat{f}_1(x)}(x) \in \mathscr{L}(\hat{f}_1(x))$:*

$$\begin{aligned}\frac{\tau}{2} + \frac{\hat{f}_2(x)}{2\tau} - \hat{f}_1(T_{L,\tau}(x)) \geq Lr^2 \varkappa\left(\frac{\Delta_r(x)}{2\tau Lr^2}\right) \Rightarrow \{\tau = \hat{f}_1(x)\} \Rightarrow \hat{f}_1(x) - \hat{f}_1(T_{L,\hat{f}_1(x)}(x)) \geq \\ \geq Lr^2 \varkappa\left(\frac{\Delta_r(x)}{2\hat{f}_1(x) Lr^2}\right) \geq 0 \Rightarrow \hat{f}_1(x) \geq \hat{f}_1(T_{L,\hat{f}_1(x)}(x)) \Rightarrow \\ \Rightarrow T_{L,\hat{f}_1(x)}(x) \in \mathscr{L}(\hat{f}_1(T_{L,\hat{f}_1(x)}(x))) \subseteq \mathscr{L}(\hat{f}_1(x)).\end{aligned}$$

Lemma 4 bounds the local model for function $\hat{f}_1$. As corollary, this lemma also bounds the local model with distance to solution of the system of equations (4).

Lemma 4. *Let $x \in \mathscr{F}$, $T_{L,\tau}(x) \in \mathscr{F}$, $L > 0$, $\tau > 0$. Then*

$$\psi_{x,L,\tau}(T_{L,\tau}(x)) \leq \min_{y \in \mathscr{F}} \left\{ \frac{\tau}{2} + \frac{L\|y-x\|^2}{2} + \frac{\hat{f}_2(y)}{2\tau} + \frac{\hat{f}_1(y) L_{\hat{F}} \|y-x\|^2}{2\tau} + \frac{L_{\hat{F}}^2 \|y-x\|^4}{8\tau} \right\}.$$

Proof. By definition of $\psi_{x,L,\tau}(\cdot)$:

$$\begin{aligned}\psi_{x,L,\tau}(T_{L,\tau}(x)) = \min_{y \in \mathscr{F}} \left\{ \frac{\tau}{2} + \frac{1}{2\tau} \left\| \hat{F}(x) + \hat{F}'(x)(y-x) \right\|^2 + \frac{L}{2}\|y-x\|^2 \right\} = \frac{\tau}{2} + \\ + \min_{y \in \mathscr{F}} \left\{ \frac{1}{2\tau} \left(\left\| \hat{F}(y) - \left(\hat{F}(y) - \hat{F}(x) - \hat{F}'(x)(y-x) \right) \right\| \right)^2 + \frac{L}{2}\|y-x\|^2 \right\} \leq \frac{\tau}{2} + \\ + \min_{y \in \mathscr{F}} \left\{ \frac{1}{2\tau} \left(\hat{f}_1(y) + \left\| \hat{F}(y) - \hat{F}(x) - \hat{F}'(x)(y-x) \right\| \right)^2 + \frac{L}{2}\|y-x\|^2 \right\} \leq \\ \leq \{\text{inequality (12)}\} \leq \\ \leq \frac{\tau}{2} + \min_{y \in \mathscr{F}} \left\{ \frac{1}{2\tau} \left(\hat{f}_1(y) + \frac{L_{\hat{F}}}{2}\|y-x\|^2 \right)^2 + \frac{L}{2}\|y-x\|^2 \right\} \leq \frac{\tau}{2} + \\ + \min_{y \in \mathscr{F}} \left\{ \frac{L\|y-x\|^2}{2} + \frac{\hat{f}_2(y)}{2\tau} + \frac{\hat{f}_1(y) L_{\hat{F}} \|y-x\|^2}{2\tau} + \frac{L_{\hat{F}}^2 \|y-x\|^4}{8\tau} \right\}.\end{aligned}$$

□

Corollary 4.1. *Suppose $x^* \in \mathscr{F}$ is the solution of (4): $\hat{F}(x^*) = \mathbf{0}_m$, $\mathscr{L}(\hat{f}_1(x)) \subseteq \mathscr{F}$. Then*

$$\begin{aligned}\psi_{x,L,\tau}(T_{L,\tau}(x)) \leq \min_{y \in \mathscr{F}} \left\{ \frac{\tau}{2} + \frac{L\|y-x\|^2}{2} + \frac{\hat{f}_2(y)}{2\tau} + \frac{\hat{f}_1(y) L_{\hat{F}} \|y-x\|^2}{2\tau} + \frac{L_{\hat{F}}^2 \|y-x\|^4}{8\tau} \right\} \leq \frac{\tau}{2} + \\ + \frac{L\|y-x\|^2}{2} + \frac{\hat{f}_2(y)}{2\tau} + \frac{\hat{f}_1(y) L_{\hat{F}} \|y-x\|^2}{2\tau} + \frac{L_{\hat{F}}^2 \|y-x\|^4}{8\tau} = \{y = x^*\} = \\ = \frac{\tau}{2} + \frac{L\|x-x^*\|^2}{2} + \frac{L_{\hat{F}}^2 \|x-x^*\|^4}{8\tau}.\end{aligned}$$

B The proof of main results for Gauss–Newton method

The deterministic Gauss–Newton method is considered within settings (6). The general deterministic Gauss–Newton framework is conceptually described as algorithm 1. However, the scheme above deserves some criticism. Values ε_k can't be arbitrary big and we either should choose small enough values for ε_k or design procedures to force $x_{k+1} \in \mathscr{F}$ for scheme correctness reason. The main and maybe already conventional ways are listed below:

- add auxiliary rule for $x_{k+1} \in \mathscr{F}$ search: $\hat{f}_1(x_k) \geq \hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1})$ and force $x_{k+1} := x_k$ in case of inability to achieve strict inequality;
- choose small enough $\varepsilon_k \geq 0$ to guarantee $x_{k+1} \in \mathscr{F}$: $\mathscr{L}(\hat{f}_1(x_k) + \varepsilon_k) \subseteq \mathscr{F}$;
- introduce "correction procedure", e.g. projection onto $\mathscr{F}$ for every obtained x_{k+1}.

B.1 The proof of theorem 1

Theorem 1 states global sublinear convergence rate to approximate stationary point.

Theorem 1. *Suppose that assumption 1 is satisfied, $k \in \mathbb{N},\ r > 0$. Then Gauss–Newton method, implemented using scheme 1 with $\tau_k = \hat{f}_1(x_k)$, $\varepsilon_k = \varepsilon \geq 0$, has the following estimates:*

$$\begin{cases} \frac{8L_{\hat{F}}^2}{L}\left(\varepsilon + \frac{\left(\hat{f}_1(x_0) - \hat{f}_1(x_k)\right)}{k}\right) \geq \min\limits_{i\in\overline{0,k-1}}\left\{\left\|2L_{\hat{F}}\left(T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i) - x_i\right)\right\|^2\right\}; \\ L_{\hat{F}}\left(\varepsilon + \frac{\left(\hat{f}_1(x_0) - \hat{f}_1(x_k)\right)}{k}\right) \geq \min\limits_{i\in\overline{0,k-1}}\left\{2\left(L_{\hat{F}}r\right)^2 \varkappa\left(\frac{\Delta_r(x_i)}{4\hat{f}_1(x_i)L_{\hat{F}}r^2}\right)\right\}; \end{cases}$$

where $\varkappa(t) = \frac{t^2}{2}\mathbb{1}_{\{t\in[0,1]\}} + \left(t - \frac{1}{2}\right)\mathbb{1}_{\{t>1\}}$.

Proof. According to lemmas 2, 3 and corollaries 2.3, 3.3 for $\tau = \hat{f}_1(x_k),\ L = L_k,\ x = x_k$ we have

$$\begin{cases} \hat{f}_1(x_k) - \psi_{x_k,L_k,\hat{f}_1(x_k)}(T_{L_k,\hat{f}_1(x_k)}(x_k)) \geq \frac{L_k}{2}\left\|T_{L_k,\hat{f}_1(x_k)}(x_k) - x_k\right\|^2; \\ \hat{f}_1(x_k) - \psi_{x_k,L_k,\hat{f}_1(x_k)}(T_{L_k,\hat{f}_1(x_k)}(x_k)) \geq L_k r^2 \varkappa\left(\frac{\Delta_r(x_k)}{2\hat{f}_1(x_k)L_k r^2}\right). \end{cases}$$

Add and subtract $\psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1})$:

$$\begin{cases} \hat{f}_1(x_k) + \left(\psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1}) - \psi_{x_k,L_k,\hat{f}_1(x_k)}(T_{L_k,\hat{f}_1(x_k)}(x_k))\right) - \psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1}) \geq \\ \qquad \geq \frac{L_k}{2}\left\|T_{L_k,\hat{f}_1(x_k)}(x_k) - x_k\right\|^2; \\ \hat{f}_1(x_k) + \left(\psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1}) - \psi_{x_k,L_k,\hat{f}_1(x_k)}(T_{L_k,\hat{f}_1(x_k)}(x_k))\right) - \psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1}) \geq \\ \qquad \geq L_k r^2 \varkappa\left(\frac{\Delta_r(x_k)}{2\hat{f}_1(x_k)L_k r^2}\right). \end{cases}$$

We use conditions

$$\psi_{x_k,L_k,\tau_k}(x_{k+1}) - \psi_{x_k,L_k,\tau_k}(T_{L_k,\tau_k}(x_k)) \leq \varepsilon_k = \varepsilon$$

and $-\psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1}) \leq -\hat{f}_1(x_{k+1})$:

$$\begin{cases} \hat{f}_1(x_k)+\varepsilon-\hat{f}_1(x_{k+1}) \geq \dfrac{L_k}{2}\left\|T_{L_k,\hat{f}_1(x_k)}(x_k)-x_k\right\|^2; \\ \hat{f}_1(x_k)+\varepsilon-\hat{f}_1(x_{k+1}) \geq L_k r^2 \varkappa\left(\dfrac{\Delta_r(x_k)}{2\hat{f}_1(x_k)L_k r^2}\right). \end{cases}$$

We average both parts of inequalities over first k iterations:

$$\begin{cases} \varepsilon+\dfrac{\hat{f}_1(x_0)-\hat{f}_1(x_k)}{k} \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1}\dfrac{L_i}{2}\left\|T_{L_i,\hat{f}_1(x_i)}(x_i)-x_i\right\|^2; \\ \varepsilon+\dfrac{\hat{f}_1(x_0)-\hat{f}_1(x_k)}{k} \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1} L_i r^2 \varkappa\left(\dfrac{\Delta_r(x_i)}{2\hat{f}_1(x_i)L_i r^2}\right). \end{cases} \tag{15}$$

Using the following restriction $L_k \geq L$ in scheme 1 and monotonicity of $\left\|T_{L_i,\hat{f}_1(x_i)}(x_i)-x_i\right\|^2$ and $L_i r^2 \varkappa\left(\frac{\Delta_r(x_i)}{2\hat{f}_1(x_i)L_i r^2}\right)$ over L_i (corollaries 2.2 and 3.2):

$$\begin{cases} \varepsilon+\dfrac{\hat{f}_1(x_0)-\hat{f}_1(x_k)}{k} \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1}\dfrac{L_i}{2}\left\|T_{L_i,\hat{f}_1(x_i)}(x_i)-x_i\right\|^2 \geq \\ \quad \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1}\dfrac{L}{2}\left\|T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i)-x_i\right\|^2 \geq \min\limits_{i\in\overline{0,k-1}}\left\{\dfrac{L}{2}\left\|T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i)-x_i\right\|^2\right\}; \\ \varepsilon+\dfrac{\hat{f}_1(x_0)-\hat{f}_1(x_k)}{k} \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1} L_i r^2 \varkappa\left(\dfrac{\Delta_r(x_i)}{2\hat{f}_1(x_i)L_i r^2}\right) \geq \\ \quad \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1} 2L_{\hat{F}} r^2 \varkappa\left(\dfrac{\Delta_r(x_i)}{4\hat{f}_1(x_i)L_{\hat{F}} r^2}\right) \geq \min\limits_{i\in\overline{0,k-1}}\left\{2L_{\hat{F}} r^2 \varkappa\left(\dfrac{\Delta_r(x_i)}{4\hat{f}_1(x_i)L_{\hat{F}} r^2}\right)\right\}. \end{cases}$$

Finally, we multiply both sides of the inequalities above by constants to obtain bounds on *generalized proximal gradients*:

$$\begin{cases} \frac{8L_{\hat{F}}^2}{L}\left(\varepsilon+\frac{\left(\hat{f}_1(x_0)-\hat{f}_1(x_k)\right)}{k}\right) \geq \min\limits_{i\in\overline{0,k-1}}\left\{\left\|2L_{\hat{F}}\left(T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i)-x_i\right)\right\|^2\right\}; \\ L_{\hat{F}}\left(\varepsilon+\frac{\left(\hat{f}_1(x_0)-\hat{f}_1(x_k)\right)}{k}\right) \geq \min\limits_{i\in\overline{0,k-1}}\left\{2\left(L_{\hat{F}}r\right)^2 \varkappa\left(\frac{\Delta_r(x_i)}{4\hat{f}_1(x_i)L_{\hat{F}}r^2}\right)\right\}. \end{cases}$$

□

Corollary 1.1. *In case of adaptive accuracy for x_{k+1} computation, such as $\varepsilon_0 = \varepsilon\hat{f}_1(x_0)$, $\varepsilon_k = \varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1(x_k)\right)$, $k\in\mathbb{N}$, $\varepsilon\geq 0$, it is possible to achieve approximation of solution for* (4) *with arbitrary low approximation error. To prove that we consider* (15) *use the defined above ε_k computation strategy:*

$$\begin{cases} \dfrac{\varepsilon\left(2\hat{f}_1(x_0)-\hat{f}_1(x_{k-1})\right)}{k}+\dfrac{\hat{f}_1(x_0)-\hat{f}_1(x_k)}{k} \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1}\dfrac{L_i}{2}\left\|T_{L_i,\hat{f}_1(x_i)}(x_i)-x_i\right\|^2; \\ \dfrac{\varepsilon\left(2\hat{f}_1(x_0)-\hat{f}_1(x_{k-1})\right)}{k}+\dfrac{\hat{f}_1(x_0)-\hat{f}_1(x_k)}{k} \geq \dfrac{1}{k}\sum\limits_{i=0}^{k-1} L_i r^2 \varkappa\left(\dfrac{\Delta_r(x_i)}{2\hat{f}_1(x_i)L_i r^2}\right). \end{cases}$$

Using the proof scheme of theorem 1 we get:

$$\begin{cases} \frac{8L_{\hat{F}}^2}{kL}\left((1+2\varepsilon)\hat{f}_1(x_0)-\varepsilon\hat{f}_1(x_{k-1})-\hat{f}_1(x_k)\right) \geq \min\limits_{i\in\overline{0,k-1}}\left\{\left\|2L_{\hat{F}}\left(T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i)-x_i\right)\right\|^2\right\}; \\ \frac{L_{\hat{F}}}{k}\left((1+2\varepsilon)\hat{f}_1(x_0)-\varepsilon\hat{f}_1(x_{k-1})-\hat{f}_1(x_k)\right) \geq \min\limits_{i\in\overline{0,k-1}}\left\{2\left(L_{\hat{F}}r\right)^2\varkappa\left(\frac{\Delta_r(x_i)}{4\hat{f}_1(x_i)L_{\hat{F}}r^2}\right)\right\}. \end{cases}$$

Corollary 1.2. *If we substitute the initial iteration with the k–th one and substitute k–th iteration with $(k+N+1)\in\mathbb{N}$ iteration, we obtain estimate for the tail of optimization procedure $k\in\mathbb{Z}_+$:*

$$\begin{cases} \frac{8L_{\hat{F}}^2}{(N+1)L}\left(\varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1(x_{k+N})\right)+\hat{f}_1(x_k)-\hat{f}_1(x_{k+N+1})\right) \geq \\ \qquad \geq \min\limits_{i\in\overline{k,k+N}}\left\{\left\|2L_{\hat{F}}\left(T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i)-x_i\right)\right\|^2\right\}; \\ \frac{L_{\hat{F}}}{N+1}\left(\varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1(x_{k+N})\right)+\hat{f}_1(x_k)-\hat{f}_1(x_{k+N+1})\right) \geq \\ \qquad \geq \min\limits_{i\in\overline{k,k+N}}\left\{2\left(L_{\hat{F}}r\right)^2\varkappa\left(\frac{\Delta_r(x_i)}{4\hat{f}_1(x_i)L_{\hat{F}}r^2}\right)\right\}. \end{cases}$$

Unrolling theorem 1 proof for initial iteration $k>0$ and final iteration $k+N$ we have estimate for the sum of inequalities in (15)*:*

$$\begin{cases} \hat{f}_1(x_k)-\hat{f}_1(x_{k+N+1})+\varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1(x_{k+N})\right) \geq \frac{L}{2}\sum\limits_{i=k}^{k+N}\left\|T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i)-x_i\right\|^2 \geq \\ \qquad \geq \frac{L}{2}\left\|T_{2L_{\hat{F}},\hat{f}_1(x_k)}(x_k)-x_k\right\|^2; \\ \hat{f}_1(x_k)-\hat{f}_1(x_{k+N+1})+\varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1(x_{k+N})\right) \geq \sum\limits_{i=k}^{k+N}2L_{\hat{F}}r^2\varkappa\left(\frac{\Delta_r(x_i)}{4\hat{f}_1(x_i)L_{\hat{F}}r^2}\right) \geq \\ \qquad \geq 2L_{\hat{F}}r^2\varkappa\left(\frac{\Delta_r(x_k)}{4\hat{f}_1(x_k)L_{\hat{F}}r^2}\right). \end{cases}$$

In the limit $N\to+\infty$ we get

$$\begin{cases} \hat{f}_1(x_k)-\hat{f}_1^*+\varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1^*\right) \geq \frac{L}{2}\left\|T_{2L_{\hat{F}},\hat{f}_1(x_k)}(x_k)-x_k\right\|^2; \\ \hat{f}_1(x_k)-\hat{f}_1^*+\varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1^*\right) \geq 2L_{\hat{F}}r^2\varkappa\left(\frac{\Delta_r(x_k)}{4\hat{f}_1(x_k)L_{\hat{F}}r^2}\right). \end{cases} \tag{16}$$

Inequalities in (16) *conditioned on* $\lim\limits_{k\to+\infty}\varepsilon_k=\lim\limits_{k\to+\infty}\varepsilon\left(\hat{f}_1(x_{k-1})-\hat{f}_1(x_k)\right)=0$ *mean*

$$\lim_{k\to+\infty}x_{k+1}=\lim_{k\to+\infty}T_{2L_{\hat{F}},\hat{f}_1(x_k)}(x_k)=x^*$$

and

$$\begin{cases} \lim\limits_{k\to+\infty}\|x_{k+1}-x_k\|=0; \\ \lim\limits_{k\to+\infty}\frac{\Delta_r(x_k)}{\hat{f}_1(x_k)}=0. \end{cases} \tag{17}$$

Limits in (17) *are deduced as consequence after taking limits in* (16) *over* $k \to +\infty$*, these limits bound variation for sequence* $\{x_k\}_{k\in\mathbb{Z}_+}$ *and state connectivity for the set of limit points*

$$\{x^* : x^* \in E_1,\ \Delta_r(x^*) = 0\}$$

of sequence $\{x_k\}_{k\in\mathbb{Z}_+}$.

Corollary 1.3. *Formally, convergence condition up to the level* $\hat{\varepsilon} > 0$ *for the norm of proximal gradient map is presented below:*

$$\min_{i\in\overline{0,k-1}} \left\{ \left\| 2L_{\hat{F}} \left(T_{2L_{\hat{F}},\hat{f}_1(x_i)}(x_i) - x_i \right) \right\| \right\} \leq \hat{\varepsilon}.$$

And such condition puts limitations for k *and* ε*:*

$$\begin{cases} \frac{8L_{\hat{F}}^2\varepsilon}{L} \leq r\hat{\varepsilon}^2,\ r \in (0,1); \\ \frac{8L_{\hat{F}}^2\left(\hat{f}_1(x_0)-\hat{f}_1(x_k)\right)}{Lk} \leq (1-r)\hat{\varepsilon}^2. \end{cases}$$

The system of inequalities results into these asymptotics:

$$\varepsilon = \frac{r\hat{\varepsilon}^2 L}{8L_{\hat{F}}^2} = \mathrm{O}\left(\hat{\varepsilon}^2\right),\ k = \left\lceil \frac{8L_{\hat{F}}^2 \hat{f}_1(x_0)}{(1-r)\hat{\varepsilon}^2 L} \right\rceil = \mathrm{O}\left(\frac{1}{\hat{\varepsilon}^2}\right).$$

B.2 The proof of theorem 2

Theorem 2 states local superlinear convergence rate to solution of problem (5).

Theorem 2. *Suppose that assumption 1 is satisfied, Jacobian is bounded:* $\left\| \hat{F}'(x) \right\| \leq M_{\hat{F}}$ *for all* $x \in \mathscr{F}$*, and the solution* $x^* \in \mathscr{L}(\hat{f}_1(x_0)),\ \hat{F}(x^*) = \mathbf{0}_m$ *with* $\sigma_{\min}\left(\hat{F}'(x^*)\right) \geq \varsigma > 0$ *exists. Then Gauss–Newton method 1 with* $\tau_k = \hat{f}_1(x_k)$*,* $\varepsilon_k = 0$ *in region*

$$\|x_k - x^*\| \leq \min\left\{ \frac{2\varsigma}{5L_{\hat{F}}},\ \frac{1}{12L_{\hat{F}}} \left(\left(3M_{\hat{F}} + 5\varsigma\right) - \sqrt{\left(3M_{\hat{F}} + 5\varsigma\right)^2 - 24\varsigma^2} \right) \right\},\ k \in \mathbb{Z}_+$$

superlinearly converges

$$\|x_{k+1} - x^*\| \leq \frac{\frac{3L_{\hat{F}}\|x_k - x^*\|^2}{2} + \|x_k - x^*\| \sqrt{\hat{f}_1(x_k)L_k + \frac{L_{\hat{F}}^2\|x_k - x^*\|^2}{4}}}{\varsigma - L_{\hat{F}}\|x_k - x^*\|} \leq \|x_k - x^*\|,$$
$$x_{k+1} \in \mathscr{L}(\hat{f}_1(x_0)),\ \hat{f}_1(x_k) = \mathrm{O}\left(\|x_k - x^*\|\right).$$

Proof. According to lemma 4 (corollary 4.1) $\psi_{x_k,L_k,\tau_k}(T_{L_k,\tau_k}(x_k))$ has upper bound:

$$\begin{aligned}
\psi_{x_k,L_k,\tau_k}(T_{L_k,\tau_k}(x_k)) &\leq \frac{\tau_k}{2} + \frac{L_k\|x_k-x^*\|^2}{2} + \frac{L_{\hat{F}}^2\|x_k-x^*\|^4}{8\tau_k} \Rightarrow \\
&\Rightarrow \{\text{add } \psi_{x_k,L_k,\tau_k}(x_{k+1}) - \psi_{x_k,L_k,\tau_k}(T_{L_k,\tau_k}(x_k)) \leq \varepsilon_k\} \Rightarrow \\
&\Rightarrow \psi_{x_k,L_k,\tau_k}(x_{k+1}) \leq \frac{\tau_k}{2} + \frac{L_k\|x_k-x^*\|^2}{2} + \frac{L_{\hat{F}}^2\|x_k-x^*\|^4}{8\tau_k} + \varepsilon_k \Rightarrow \\
\Rightarrow \psi_{x_k,L_k,\tau_k}(x_{k+1}) &= \frac{\tau_k}{2} + \frac{(\phi(x_k,x_{k+1}))^2}{2\tau_k} + \frac{L_k}{2}\|x_{k+1}-x_k\|^2 \leq \\
&\leq \frac{\tau_k}{2} + \frac{L_k\|x_k-x^*\|^2}{2} + \frac{L_{\hat{F}}^2\|x_k-x^*\|^4}{8\tau_k} + \varepsilon_k \Rightarrow \\
\Rightarrow \frac{(\phi(x_k,x_{k+1}))^2}{2\tau_k} &\leq \frac{L_k\|x_k-x^*\|^2}{2} + \frac{L_{\hat{F}}^2\|x_k-x^*\|^4}{8\tau_k} + \varepsilon_k \Rightarrow \\
&\Rightarrow \sqrt{\tau_k L_k\|x_k-x^*\|^2 + \frac{L_{\hat{F}}^2\|x_k-x^*\|^4}{4} + 2\tau_k\varepsilon_k} \geq \\
&\geq \phi(x_k,x_{k+1}) \Rightarrow \sqrt{\|x_k-x^*\|^2\left(\tau_k L_k + \frac{L_{\hat{F}}^2\|x_k-x^*\|^2}{4}\right) + 2\tau_k\varepsilon_k} \geq \\
&\geq \left\|\hat{F}(x_k) + \hat{F}'(x_k)(x_{k+1}-x_k)\right\|.
\end{aligned}$$

Now we rewrite $\phi(x_k,x_{k+1})$ in a different way:

$$\left\|\hat{F}(x_k) + \hat{F}'(x_k)(x_{k+1}-x_k)\right\| = \left\| \underbrace{\hat{F}'(x^*)(x_{k+1}-x^*)}_{\stackrel{\text{def}}{=}A} + \underbrace{\left(\hat{F}(x_k) - \hat{F}(x^*) - \hat{F}'(x^*)(x_k-x^*)\right)}_{\stackrel{\text{def}}{=}B} + \right.$$
$$\left. + \underbrace{\left(\hat{F}'(x_k) - \hat{F}'(x^*)\right)(x_{k+1}-x_k)}_{\stackrel{\text{def}}{=}C} \right\|.$$

Using triangle inequality for norm $\|\cdot\|$:

$$\begin{aligned}
&\|A\| = \|A+B+C-B-C\| \leq \|A+B+C\| + \|-B\| + \|-C\| \Rightarrow \\
&\quad \Rightarrow \|A+B+C\| \geq \|A\| - \|B\| - \|C\|; \\
&\|A\| \geq \{\text{using minimal singular value definition}\} \geq \varsigma\|x_{k+1}-x^*\|; \\
&\|B\| \leq \{\text{inequality (12)}\} \leq \frac{L_{\hat{F}}}{2}\|x_k-x^*\|^2; \\
&\|C\| \leq \{\text{submultiplicativity of norm}\} \leq \left\|\hat{F}'(x_k) - \hat{F}'(x^*)\right\|\|x_{k+1}-x_k\| \leq \{\text{assumption 1}\} \leq \\
&\quad \leq L_{\hat{F}}\|x_k-x^*\|\|x_{k+1}-x^*+x^*-x_k\| \leq L_{\hat{F}}\|x_k-x^*\|^2 + L_{\hat{F}}\|x_k-x^*\|\|x_{k+1}-x^*\|.
\end{aligned}$$

Combining the inequalities above we get the lower bound for $\phi(x_k,x_{k+1})$:

$$\phi(x_k,x_{k+1}) \geq \left(\varsigma - L_{\hat{F}}\|x_k-x^*\|\right)\|x_{k+1}-x^*\| - \frac{3L_{\hat{F}}}{2}\|x_k-x^*\|^2.$$

Now we link lower and upper bounds for $\phi(x_k, x_{k+1})$ into the inequality below:

$$\sqrt{\|x_k - x^*\|^2 \left(\tau_k L_k + \frac{L_{\hat{F}}^2 \|x_k - x^*\|^2}{4}\right) + 2\tau_k \varepsilon_k} \geq$$
$$\geq \left(\varsigma - L_{\hat{F}} \|x_k - x^*\|\right) \|x_{k+1} - x^*\| - \frac{3L_{\hat{F}} \|x_k - x^*\|^2}{2} \Rightarrow$$
$$\Rightarrow \frac{\sqrt{\|x_k - x^*\|^2 \left(\tau_k L_k + \frac{L_{\hat{F}}^2 \|x_k - x^*\|^2}{4}\right) + 2\tau_k \varepsilon_k} + \frac{3L_{\hat{F}} \|x_k - x^*\|^2}{2}}{\varsigma - L_{\hat{F}} \|x_k - x^*\|} \geq \|x_{k+1} - x^*\|.$$

To prove the theorem for specific τ_k we deduce upper bound for $\hat{f}_1(x_k)$ using local model $\psi_{x^*, L_{\hat{F}}, \phi(x^*, x_k)}(x_k)$:

$$\hat{f}_1(x_k) \leq \{\text{lemma 1}\} \leq \left\| \underbrace{\hat{F}(x^*)}_{=\mathbf{0}_m} + \hat{F}'(x^*)(x_k - x^*) \right\| + \frac{L_{\hat{F}}}{2} \|x_k - x^*\|^2 \leq \underbrace{\left\| \hat{F}'(x^*)(x_k - x^*) \right\|}_{\leq M_{\hat{F}} \|x_k - x^*\|} +$$
$$+ \frac{L_{\hat{F}}}{2} \|x_k - x^*\|^2 \leq M_{\hat{F}} \|x_k - x^*\| + \frac{L_{\hat{F}}}{2} \|x_k - x^*\|^2 <$$
$$< \left\{ \begin{array}{l} \text{upper bound for} \\ \text{superlinear convergence region: } \|x_k - x^*\| < \dfrac{\varsigma}{L_{\hat{F}}} \end{array} \right\} < \left(M_{\hat{F}} + \frac{\varsigma}{2}\right) \|x_k - x^*\| \leq$$
$$\leq \left\{ \varsigma \leq \sigma_{\min}(\hat{F}'(x^*)) \leq \sigma_{\max}(\hat{F}'(x^*)) \leq M_{\hat{F}} \right\} \leq \frac{3M_{\hat{F}}}{2} \|x_k - x^*\| \Rightarrow$$
$$\Rightarrow \hat{f}_1(x_k) = \mathrm{O}\left(\|x_k - x^*\|\right). \tag{18}$$

In limit, the inequality above is nonstrict. We substitute values τ_k and ε_k into the convergence estimate using alias $t_k = \|x_k - x^*\|$:

$$t_{k+1} \leq \frac{\frac{3L_{\hat{F}} t_k^2}{2} + t_k \sqrt{\hat{f}_1(x_k) L_k + \frac{L_{\hat{F}}^2 t_k^2}{4}}}{\varsigma - L_{\hat{F}} \|x_k - x^*\|} < \left\{L_k \leq 2L_{\hat{F}}, \text{ estimate (18)}\right\} <$$
$$< t_k \underbrace{\left(\frac{\frac{3L_{\hat{F}} t_k}{2} + \sqrt{3M_{\hat{F}} L_{\hat{F}} t_k + \frac{L_{\hat{F}}^2 t_k^2}{4}}}{\varsigma - L_{\hat{F}} t_k} \right)}_{\in [0,1] \text{ by theorem conditions}} \leq t_k.$$

Let us describe possible values for t_k form the limitations above:

$$0 \leq \frac{3L_{\hat{F}} t_k}{2} + \sqrt{3M_{\hat{F}} L_{\hat{F}} t_k + \frac{L_{\hat{F}}^2 t_k^2}{4}} \leq \varsigma - L_{\hat{F}} t_k \Rightarrow 0 \leq \sqrt{3M_{\hat{F}} L_{\hat{F}} t_k + \frac{L_{\hat{F}}^2 t_k^2}{4}} \leq \varsigma - \frac{5L_{\hat{F}} t_k}{2} \Rightarrow$$
$$\Rightarrow t_k \leq \frac{2\varsigma}{5L_{\hat{F}}},$$

the first condition is obtained. We square the inequality above to get rest of conditions:

$$3M_{\hat{F}} L_{\hat{F}} t_k + \frac{L_{\hat{F}}^2 t_k^2}{4} \leq \left(\varsigma - \frac{5L_{\hat{F}} t}{2}\right)^2 \Rightarrow -6L_{\hat{F}}^2 t_k^2 + \left(3M_{\hat{F}} L_{\hat{F}} + 5L_{\hat{F}} \varsigma\right) t_k - \varsigma^2 \leq 0 \Rightarrow \{t_k \geq 0\} \Rightarrow$$
$$\Rightarrow 0 \leq t_k \leq \frac{1}{12L_{\hat{F}}} \left(\left(3M_{\hat{F}} + 5\varsigma\right) - \sqrt{\left(3M_{\hat{F}} + 5\varsigma\right)^2 - 24\varsigma^2} \right).$$

Combining restrictions for $t_k \geq 0$ we justify lower bound for the region of superlinear convergence:

$$\|x_k - x^*\| \leq \min\left\{\frac{2\varsigma}{5L_{\hat{F}}},\ \frac{1}{12L_{\hat{F}}}\left(\left(3M_{\hat{F}}+5\varsigma\right) - \sqrt{\left(3M_{\hat{F}}+5\varsigma\right)^2 - 24\varsigma^2}\right)\right\}.$$

□

Corollary 2.1. *The statement of theorem implicitly put restrictions for system of equations* (4)*:*

- *the nondegeneracy of* (4) *in the point of minimum* $\sigma_{\min}\left(\hat{F}'(x^*)\right) \geq \varsigma > 0$ *means* $\dim(E_2) \geq \dim(E_1)$*;*
- *the solvability of the system of equations* (4) $\hat{F}(x^*) = \mathbf{0}_m$ *is usually meet the following limitation:* $\dim(E_2) \leq \dim(E_1)$.

So, typically we can achieve local superlinear convergence solving the system of equations with $\dim(E_1) = \dim(E_2)$.

B.3 The proof of theorem 3

Theorem 3 states global sublinear and local linear convergence rates to approximate solution of problem (5).

Theorem 3. *Assume that assumptions 1 and 2 are held for Gauss–Newton method 1 with* $\tau_k = \hat{f}_1(x_k)$. *Then any sequence* $\{x_k\}_{k\in\mathbb{Z}_+}$ *has the property:*

$$\hat{f}_1(x_{k+1}) \leq \varepsilon_k + \begin{cases} \frac{\hat{f}_1(x_k)}{2} + \frac{L_{\hat{F}}}{\mu}\hat{f}_2(x_k) \leq \frac{3}{4}\hat{f}_1(x_k), \text{ if } \hat{f}_1(x_k) \leq \frac{\mu}{4L_{\hat{F}}}; \\ \hat{f}_1(x_k) - \frac{\mu}{16L_{\hat{F}}}, \text{ otherwise.} \end{cases}$$

Proof. Consider system of linear equations $\hat{F}(x) + \hat{F}'(x)h = \mathbf{0}_m,\ x \in \mathscr{F}$. There is $h \in E_1$: $\hat{F}(x) + \hat{F}'(x)h = \mathbf{0}_m, x \in \mathscr{F}$ according to PL condition and

$$h = -\hat{F}'(x)^*\left(\hat{F}'(x)\hat{F}'(x)^*\right)^{-1}\hat{F}(x).$$

Then, using assumption 2 we have

$$\|h\| = \left\|\hat{F}'(x)^*\left(\hat{F}'(x)\hat{F}'(x)^*\right)^{-1}\hat{F}(x)\right\| = \sqrt{\left\langle\left(\hat{F}'(x)\hat{F}'(x)^*\right)^{-1}\hat{F}(x),\ \hat{F}(x)\right\rangle} \leq \frac{\left\|\hat{F}(x)\right\|}{\sqrt{\mu}} = \\ = \frac{\hat{f}_1(x)}{\sqrt{\mu}}. \tag{19}$$

By definition of the local model for $x_{k+1},\ k \in \mathbb{Z}_+$:

$$\begin{aligned}
\hat{f}_1(x_{k+1}) &\le \psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1}) = \psi_{x_k,L_k,\hat{f}_1(x_k)}(T_{L_k,\hat{f}_1(x_k)}(x_k)) + \Big(\psi_{x_k,L_k,\hat{f}_1(x_k)}(x_{k+1}) - \\
&- \psi_{x_k,L_k,\hat{f}_1(x_k)}(T_{L_k,\hat{f}_1(x_k)}(x_k))\Big) \le \varepsilon_k + \psi_{x_k,L_k,\hat{f}_1(x_k)}(T_{L_k,\hat{f}_1(x_k)}(x_k)) = \varepsilon_k + \\
&+ \min_{y\in E_1}\left\{\frac{\hat{f}_1(x_k)}{2} + \frac{(\phi(x_k,x_k+y))^2}{2\hat{f}_1(x_k)} + \frac{L_k}{2}\|y\|^2\right\} \le \\
&\le \left\{\text{instead of } y \text{ we substitute } th_k = -t\hat{F}'(x_k)^*\left(\hat{F}'(x_k)\hat{F}'(x_k)^*\right)^{-1}\hat{F}(x_k),\ t\in[0,\ 1]\right\} \le \\
&\le \frac{\hat{f}_1(x_k)}{2} + \min_{t\in[0,\ 1]}\left\{\frac{1}{2\hat{f}_1(x_k)}\left\|\hat{F}(x_k) + t\hat{F}'(x_k)h_k\right\|^2 + \frac{t^2L_k}{2}\|h_k\|^2\right\} \le \\
&\le \{\text{inequality (19)}\} \le \varepsilon_k + \frac{\hat{f}_1(x_k)}{2} + \min_{t\in[0,\ 1]}\left\{\frac{\left\|(1-t)\hat{F}(x_k)\right\|^2}{2\hat{f}_1(x_k)} + \frac{t^2L_k}{2\mu}\hat{f}_2(x_k)\right\} \le \\
&\le \left\{\|\cdot\|^2 \text{ is convex}\right\} \le \varepsilon_k + \frac{\hat{f}_1(x_k)}{2} + \min_{t\in[0,\ 1]}\left\{\frac{1-t}{2}\hat{f}_1(x_k) + \frac{t^2L_k}{2\mu}\hat{f}_2(x_k)\right\} = \\
&= \varepsilon_k + \hat{f}_1(x_k) + \frac{\hat{f}_2(x_k)L_k}{\mu}\min_{t\in[0,\ 1]}\left\{\frac{-t\mu}{2\hat{f}_1(x_k)L_k} + \frac{t^2}{2}\right\} = \varepsilon_k + \\
&+ \hat{f}_1(x_k) - \frac{\hat{f}_2(x_k)L_k}{\mu}\max_{t\in[0,\ 1]}\left\{\frac{t\mu}{2\hat{f}_1(x_k)L_k} - \frac{t^2}{2}\right\} = \{(14),\ \text{lemma 3}\} = \varepsilon_k + \hat{f}_1(x_k) - \\
&- \frac{\hat{f}_2(x_k)L_k}{\mu}\varkappa\left(\frac{\mu}{2\hat{f}_1(x_k)L_k}\right) \le \{\text{monotone decrease over } L_k\} \le \\
&\le \varepsilon_k + \hat{f}_1(x_k) - \frac{2\hat{f}_2(x_k)L_{\hat{F}}}{\mu}\varkappa\left(\frac{\mu}{4\hat{f}_1(x_k)L_{\hat{F}}}\right).
\end{aligned}$$

We express using the explicit form of $\varkappa(\cdot)$ considering monotone decrease of $\frac{\hat{f}_2(x_k)L_k}{\mu}\varkappa\left(\frac{\mu}{2\hat{f}_1(x_k)L_k}\right)$ over L_k (corollary 3.2):

$$\hat{f}_1(x_{k+1}) \le \varepsilon_k + \begin{cases}\hat{f}_1(x_k) - \frac{\mu}{16L_{\hat{F}}},\ \text{if } \hat{f}_1(x_k) \ge \frac{\mu}{4L_{\hat{F}}};\\ \frac{\hat{f}_1(x_k)}{2} + \frac{\hat{f}_2(x_k)L_{\hat{F}}}{\mu} \le \frac{3}{4}\hat{f}_1(x_k),\ \text{if } \hat{f}_1(x_k) \le \frac{\mu}{4L_{\hat{F}}}.\end{cases}$$

□

Corollary 3.1. *An adaptive choice of $\varepsilon_k \ge 0$ allows us to solve* (5) *with arbitrary precision. As an example for such choice define the following sequence $\{\delta_k\}_{k\in\mathbb{Z}_+}$: $\frac{3}{4}\delta_k > \delta_{k+1} > 0$, $\delta_{-1} = \frac{8}{3}\delta_0$, $\lim\limits_{k\to+\infty}\delta_k = 0$ and additionally define*

$$\hat{f}_1(x_{-1}) \overset{\text{def}}{=} \frac{\mu}{4L_{\hat{F}}}.$$

We define as $N \in \mathbb{Z}_+ \cup \{-1\}$ the minimal number of the iteration for which the next inequalities hold (and set $N = -1$ if such iteration does not exist):

$$\hat{f}_1(x_N) \ge \frac{\mu}{4L_{\hat{F}}} \ge \hat{f}_1(x_{N+1}).$$

The next strategy of ε_k choice allows us to achieve an arbitrary precision of the approximate solution for (4)*:*

$$\varepsilon_k = \begin{cases} \delta_0 : \ \delta_0 < \frac{\mu}{16L_{\hat{F}}}, \text{ for } k = 0; \\ \delta_{k-1} - \delta_k, \text{ if } 0 < k \leq N+1; \\ \frac{3}{4}\delta_{k-1} - \delta_k, \text{ if } k > N+1. \end{cases}$$

So the strategy states decrease of the approximation error of x_{k+1} search forming the convergence estimates below:

$$\begin{cases} \hat{f}_1(x_k) \leq 2\delta_0 - \delta_{k-1} + \hat{f}_1(x_0) - \frac{k\mu}{16L_{\hat{F}}}, \text{ if } 0 < k \leq N+1; \\ \\ \hat{f}_1(x_k) \leq \left(\frac{3}{4}\right)^{k-N-1} \hat{f}_1(x_{N+1}) + \delta_N \left(\frac{3}{4}\right)^{k-N-1} - \delta_{k-1}, \text{ if } k > N+1. \end{cases}$$

Corollary 3.2. *If we have a constant level of the approximation error $\varepsilon_k = \varepsilon > 0$, we can formally deduce necessary number of iterations and the maximal value of error in the worst case scenario to get $\hat{f}_1(x_k) \leq \hat{\varepsilon}$:*

- *if $\hat{\varepsilon} \geq \frac{\mu}{4L_{\hat{F}}}$, then $k \geq \left\lceil \left(\frac{\mu}{16L_{\hat{F}}} - \varepsilon\right)^{-1} \left(\hat{f}_1(x_0) - \hat{\varepsilon}\right) \mathbb{1}_{\{\hat{f}_1(x_0) > \hat{\varepsilon}\}} \right\rceil$, $\varepsilon < \frac{\mu}{16L_{\hat{F}}}$;*
- *if $\hat{\varepsilon} < \frac{\mu}{4L_{\hat{F}}}$, then $k \geq \left\lceil \left(\frac{\mu}{16L_{\hat{F}}} - \varepsilon\right)^{-1} \left(\hat{f}_1(x_0) - \frac{\mu}{4L_{\hat{F}}}\right) \mathbb{1}_{\left\{\hat{f}_1(x_0) > \frac{\mu}{4L_{\hat{F}}}\right\}} + \log_{\frac{4}{3}} \left(\frac{\mu}{4r\hat{\varepsilon}L_{\hat{F}}}\right) \right\rceil$, $\varepsilon \leq \frac{(1-r)\hat{\varepsilon}}{4}$, $r \in (0,1)$.*

C The proof of auxiliary results for stochastic Gauss–Newton method

Lemma 5 states an important partial order for establishing linear convergence under PL condition.

Lemma 5 ([21])**.** *Suppose the linear operator $A : E_1 \to E_2$ with $\dim(E_1) = n$, $\dim(E_2) = m$, $m \leq n$ is row–nondegenerate:*

$$AA^* \succeq \mu I_m$$

for some $\mu > 0$. Then for any $\xi > 0$ we have

$$A\left(\xi I_n + A^*A\right)^{-t} A^* \succeq \frac{\mu}{(\xi + \mu)^t} I_m, \ t \in [0, \ 1].$$

The partial order $\succeq$ is defined on positive semi–definite cone.

Proof. Consider the singular value decomposition of the operator matrix A:

$$A = U\Lambda V^*, \ U^*U = I_m, \ V^*V = I_m,$$

where Λ is a diagonal matrix, $\Lambda \succeq \sqrt{\mu} I_m$ (by statement of this lemma). Define matrix W with columns as orthogonal complement of columns in V up to full basis in E_1:

$$VV^* + WW^* = I_n, \ W^*V = \mathbf{0}_{(n-m)\times m}.$$

Exploiting block structure as result of the identity $W^*V = \mathbf{0}_{(n-m)\times m}$ we get:

$$\begin{aligned}
A\left(\xi I_n + A^*A\right)^{-t} A^* &= U\Lambda V^* \left(\xi\left(VV^* + WW^*\right) + V\Lambda^2 V^*\right)^{-t} V\Lambda U^* = \\
&= U\Lambda V^* \left(V\left(\xi I_m + \Lambda^2\right)V^* + \xi WW^*\right)^{-t} V\Lambda U^* = \\
&= U\Lambda V^* \left(V\left(\xi I_m + \Lambda^2\right)^{-t} V^* + \frac{1}{\xi^t} WW^*\right) V\Lambda U^* = \\
&= U\Lambda\left(\xi I_m + \Lambda^2\right)^{-t}\Lambda U^* = \\
&= U\left(\xi\Lambda^{-\frac{2}{t}} + \Lambda^{2-\frac{2}{t}}\right)^{-t} U^* \succeq \frac{1}{\left(\xi\mu^{-\frac{1}{t}} + \mu^{1-\frac{1}{t}}\right)^t} I_m = \frac{\mu}{(\xi+\mu)^t} I_m.
\end{aligned}$$

□

The lemma below describes main properties of the partial order on positive semi–definite cone.

Lemma 6. *Let assumptions 4 and 7 hold. Then for every $t \geq 0$, $x \in E_1$, $B \subseteq \mathscr{B}$, $|B| = b \in \overline{1, \min\{m,n\}}$ the next relations are satisfied:*

$$\begin{cases}
\tau^t I_n \preceq \left(\hat{G}'(x,B)^*\hat{G}'(x,B) + \tau I_n\right)^t \preceq \left(M_{\hat{G}}^2 + \tau\right)^t I_n,\ \tau \geq 0; \\
\frac{1}{\left(M_{\hat{G}}^2+\tau\right)^t} I_n \preceq \left(\hat{G}'(x,B)^*\hat{G}'(x,B) + \tau I_n\right)^{-t} \preceq \frac{1}{\tau^t} I_n,\ \tau > 0; \\
(\mu+\tau)^t I_b \preceq \left(\hat{G}'(x,B)\hat{G}'(x,B)^* + \tau I_b\right)^t \preceq \left(M_{\hat{G}}^2 + \tau\right)^t I_b,\ \tau \geq 0; \\
\frac{1}{\left(M_{\hat{G}}^2+\tau\right)^t} I_b \preceq \left(\hat{G}'(x,B)\hat{G}'(x,B)^* + \tau I_b\right)^{-t} \preceq \frac{1}{(\mu+\tau)^t} I_b,\ \tau \geq 0.
\end{cases}$$

Partial order $\preceq$ is defined on positive semi–definite cone.

Proof. Assumption 4 bounds maximal singular value of $\hat{G}'(x,B)$ from above:

$$\begin{aligned}
\left\|\hat{G}'(x,B)\right\| = \sigma_{\max}(\hat{G}'(x,B)) \leq M_{\hat{G}} \Leftrightarrow & \hat{G}'(x,B)^*\hat{G}'(x,B) \preceq M_{\hat{G}}^2 I_n, \\
& \hat{G}'(x,B)\hat{G}'(x,B)^* \preceq M_{\hat{G}}^2 I_b.
\end{aligned}$$

Assumption 7 bounds minimal singular value of $\hat{G}'(x,B)^*$ from below:

$$\hat{G}'(x,B)\hat{G}'(x,B)^* \succeq \mu I_b \Leftrightarrow \sigma_{\min}(\hat{G}'(x,B)^*) \geq \sqrt{\mu}.$$

Symmetric matrices $\left(\hat{G}'(x,B)^*\hat{G}'(x,B) + \tau I_n\right)^t$ and $\left(\hat{G}'(x,B)\hat{G}'(x,B)^* + \tau I_b\right)^t$ have spectral decomposition (or eigendecomposition) with corresponding diagonal matrices Λ_1^t and Λ_2^t, with corresponding orthogonal matrices Q_1 and Q_2. For arbitrary $v \in E_1$ we have:

$$\left\langle \left(\hat{G}'(x,B)^*\hat{G}'(x,B) + \tau I_n\right)^t v,\ v \right\rangle = \left\langle Q_1 \left(\Lambda_1 + \tau I_n\right)^t \underbrace{Q_1^* v}_{\stackrel{\text{def}}{=} v_1},\ v \right\rangle =$$

$$= \begin{bmatrix}
\left\langle \underbrace{\left(\Lambda_1 + \tau I_n\right)^t}_{\sigma_{\max}(\Lambda_1) \leq M_{\hat{G}}^2} v_1,\ v_1 \right\rangle \leq \underbrace{\left(M_{\hat{G}}^2 + \tau\right)^t \|v_1\|^2,\ \forall v_1 \in E_1}_{\text{bounding Rayleigh quotient}}; \\
\left\langle \underbrace{\left(\Lambda_1 + \tau I_n\right)^t}_{\sigma_{\min}(\Lambda_1) \geq 0} v_1,\ v_1 \right\rangle \geq \underbrace{\tau^t \|v_1\|^2,\ \forall v_1 \in E_1}_{\substack{\text{bounding Rayleigh} \\ \text{quotient}}}.
\end{bmatrix}$$

Analogously for arbitrary $w \in E_3^*,\ \dim(E_3^*) = b$ we have:

$$\left\langle w,\ \left(\hat{G}'(x,B)\hat{G}'(x,B)^* + \tau I_b\right)^t w \right\rangle = \left\langle w,\ Q_2 \left(\Lambda_2 + \tau I_b\right)^t \underbrace{Q_2^* w}_{\stackrel{\text{def}}{=} w_1} \right\rangle =$$

$$= \begin{cases} \left\langle w_1,\ \underbrace{\left(\Lambda_2 + \tau I_b\right)^t}_{\sigma_{\max}(\Lambda_2) \le M_{\hat{G}}^2} w_1 \right\rangle \le \underbrace{\left(M_{\hat{G}}^2 + \tau\right)^t \|w_1\|^2,\ \forall w_1 \in E_3^*}_{\text{bounding Rayleigh quotient}}; \\ \left\langle w_1,\ \underbrace{\left(\Lambda_2 + \tau I_b\right)^t}_{\sigma_{\min}(\Lambda_2) \ge \mu} w_1 \right\rangle \ge \underbrace{(\mu + \tau)^t \|w_1\|^2,\ \forall w_1 \in E_3^*}_{\text{bounding Rayleigh quotient}}. \end{cases}$$

In both cases after replacement t for $-t$ we cause inversion of the spectrum forcing to interchange lower and upper estimates by Courant–Fischer–Weyl min–max principle. This means satisfaction for the next relation of partial order:

$$\begin{cases} \tau^t I_n \preceq \left(\hat{G}'(x,B)^* \hat{G}'(x,B) + \tau I_n\right)^t \preceq \left(M_{\hat{G}}^2 + \tau\right)^t I_n,\ \tau \ge 0; \\ \frac{1}{\left(M_{\hat{G}}^2 + \tau\right)^t} I_n \preceq \left(\hat{G}'(x,B)^* \hat{G}'(x,B) + \tau I_n\right)^{-t} \preceq \frac{1}{\tau^t} I_n,\ \tau > 0; \\ \tau^t I_b \preceq \left(\hat{G}'(x,B) \hat{G}'(x,B)^* + \tau I_b\right)^t \preceq \left(M_{\hat{G}}^2 + \tau\right)^t I_b,\ \tau \ge 0; \\ \frac{1}{\left(M_{\hat{G}}^2 + \tau\right)^t} I_b \preceq \left(\hat{G}'(x,B) \hat{G}'(x,B)^* + \tau I_b\right)^{-t} \preceq \frac{1}{\tau^t} I_b,\ \tau > 0. \end{cases} \tag{20}$$

□

The next proposition deduces lipschitzness of jacobians $\hat{G}'$ and $\hat{F}'$.

Lemma 7. *Suppose assumption 3 holds. Then $\hat{F}'$ and $\hat{G}'$ are Lipschitz continuous:*

$$\begin{cases} \left\| \hat{F}'(x) - \hat{F}'(y) \right\| \le L_{\hat{F}} \|x - y\|,\ \forall (x,y) \in E_1^2; \\ \left\| \hat{G}'(x,B) - \hat{G}'(y,B) \right\| \le L_{\hat{F}} \|x - y\|,\ \forall (x,y) \in E_1^2,\ \forall B \subseteq \mathscr{B}, |B| = b. \end{cases}$$

Analogously functions $\hat{f}_2$ and $\hat{g}_2$ are Lipschitz continuous:

$$\begin{cases} \left| \hat{f}_2(x) - \hat{f}_2(y) \right| \le l_{\hat{F}} \|x - y\|,\ \forall (x,y) \in E_1^2; \\ \left| \hat{g}_2(x,B) - \hat{g}_2(y,B) \right| \le l_{\hat{F}} \|x - y\|,\ \forall (x,y) \in E_1^2,\ \forall B \subseteq \mathscr{B}, |B| = b. \end{cases}$$

Proof. Consider batch of functions $\hat{G}$, for arbitrary $(x,y) \in E_1^2$ we express the following:

$$\left\| \hat{G}'(x,B) - \hat{G}'(y,B) \right\| = \sqrt{\frac{1}{b} \sum_{j=1}^{b} \left\| \nabla F_{i_j}(x) - \nabla F_{i_j}(y) \right\|^2} \le \{\text{assumptions 3}\} \le$$

$$\le \sqrt{\frac{1}{b} \sum_{j=1}^{b} L_{\hat{F}}^2 \|x - y\|^2} = L_{\hat{F}} \|x - y\|;$$

$$\left\| \hat{F}'(x) - \hat{F}'(y) \right\| = \sqrt{\mathbb{E}_q \left[\left\| \nabla F_\xi(x) - \nabla F_\xi(y) \right\|^2 \right]} \le \{\text{assumption 3}\} \le$$

$$\le \sqrt{\mathbb{E}_q \left[L_{\hat{F}}^2 \|x - y\|^2 \right]} = L_{\hat{F}} \|x - y\|,\ q \text{ defines a distribution over } \mathscr{B}.$$

By analogy, for arbitrary $(x,y) \in E_1^2$ we express inequalities for $\hat{g}_2$ and $\hat{f}_2$:

$$|\hat{g}_2(x,B) - \hat{g}_2(y,B)| = \left| \frac{1}{b}\sum_{j=1}^{b} \left(F_{i_j}(x)\right)^2 - \frac{1}{b}\sum_{j=1}^{b} \left(F_{i_j}(y)\right)^2 \right| \leq \frac{1}{b}\sum_{j=1}^{b} \left| \left(F_{i_j}(x)\right)^2 - \left(F_{i_j}(y)\right)^2 \right| \leq$$
$$\leq \{\text{assumption 3}\} \leq \frac{1}{b}\sum_{j=1}^{b} l_{\hat{F}} \|x-y\| = l_{\hat{F}} \|x-y\|;$$

$$\left|\hat{f}_2(x) - \hat{f}_2(y)\right| = \left| \mathbb{E}_q \left[\left(F_\xi(x)\right)^2 - \left(F_\xi(y)\right)^2 \right] \right| \leq \mathbb{E}_q \left[\left| \left(F_\xi(x)\right)^2 - \left(F_\xi(y)\right)^2 \right| \right] \leq$$
$$\leq \{\text{assumption 3}\} \leq \mathbb{E}_q \left[l_{\hat{F}} \|x-y\| \right] = l_{\hat{F}} \|x-y\|, \ \xi \sim q.$$

□

Corollary 7.1. *The statement of lemma also holds for infinite population $\mathscr{B}$.*

Lemma 8 deduces stochastic variant of the local model.

Lemma 8. *Let $(x,y) \in E_1^2$, $B \subseteq \mathscr{B}$, $L \geq L_{\hat{F}}$, $\tau > 0$, $\hat{g}_1(x,B) > 0$ almost sure and assumption 3 holds. Then*

$$\begin{cases} \hat{g}_1(y,B) \leq \hat{\psi}_{x,L,\tau}(y,B) = \frac{\tau}{2} + \frac{L}{2}\|y-x\|^2 + \frac{1}{2\tau}\left\| \hat{G}(x,B) + \hat{G}'(x,B)(y-x) \right\|^2; \\ \hat{f}_1(y) \leq \psi_{x,L,\tau}(y) \overset{\text{def}}{=} \hat{\psi}_{x,L,\tau}(y,\mathscr{B}) = \frac{\tau}{2} + \frac{L}{2}\|y-x\|^2 + \frac{1}{2\tau}\left\| \hat{F}(x) + \hat{F}'(x)(y-x) \right\|^2. \end{cases}$$

Proof. The proof uses structure of the proof from lemma 1 with $\hat{F} := \hat{G}$ under batch $B \subseteq \mathscr{B}$ for arbitrary $(x,y) \in E_1^2$, $L \geq L_{\hat{F}}$, $\tau > 0$. □

The next lemma expresses main properties of scaled descend direction update in the optimization procedure.

Lemma 9. *Suppose assumption 3 holds, $x_k \in E_1$, $\tau_k > 0$, $L_k \geq L_{\hat{F}}$, $B_k \subseteq \mathscr{B}$, $\eta_k > 0$, initialization x_0 is chosen randomly and independently from B_k, $k \in \mathbb{Z}_+$. Then*

$$\begin{cases} x_{k+1} = x_k - \eta_k \left(\hat{G}'(x_k,B_k)^* \hat{G}'(x_k,B_k) + \tau_k L_k I_n \right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k); \\ \hat{g}_1(x_k,B_k) - \hat{g}_1(x_{k+1},B_k) \geq \hat{g}_1(x_k,B_k) - \frac{\tau_k}{2} - \frac{\hat{g}_2(x_k,B_k)}{2\tau_k} + \\ \quad + \frac{\eta_k(2-\eta_k)}{2\tau_k} \left\langle \left(\hat{G}'(x_k,B_k)^* \hat{G}'(x_k,B_k) + \tau_k L_k I_n \right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k), \right. \\ \qquad \left. \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k) \right\rangle. \end{cases} \tag{21}$$

Proof. By definition of $\hat{\psi}_{x,L,\tau}(y,B)$:

$$\hat{g}_1(x_k,B_k) - \hat{g}_1(x_{k+1},B_k) \geq \hat{g}_1(x_k,B_k) - \hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k) = \hat{g}_1(x_k,B_k) - \frac{\tau_k}{2} -$$
$$- \frac{L_k}{2}\|x_{k+1}-x_k\|^2 - \frac{1}{2\tau_k}\left\| \hat{G}(x_k,B_k) + \hat{G}'(x_k,B_k)(x_{k+1}-x_k) \right\|^2. \tag{22}$$

Substitute expression of x_{k+1} into (22):

$$\hat{g}_1(x_k,B_k)-\hat{g}_1(x_{k+1},B_k)\geq \hat{g}_1(x_k,B_k)-\frac{\tau_k}{2}-\frac{1}{2\tau_k}\left\|\hat{G}(x_k,B_k)+\hat{G}'(x_k,B_k)(x_{k+1}-x_k)\right\|^2-$$
$$-\frac{L_k}{2}\|x_{k+1}-x_k\|^2=\hat{g}_1(x_k,B_k)-\frac{\tau_k}{2}-$$
$$-\frac{1}{2\tau_k}\left\|\hat{G}(x_k,B_k)-\eta_k\hat{G}'(x_k,B_k)\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\|^2-$$
$$-\frac{L_k}{2}\left\|\eta_k\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\|^2=\hat{g}_1(x_k,B_k)-\frac{\tau_k}{2}-\frac{\hat{g}_2(x_k,B_k)}{2\tau_k}+$$
$$+\frac{1}{2\tau_k}\left(2\left\langle\eta_k\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle-\right.$$
$$-\left\langle\eta_k\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\right.$$
$$\left.\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\eta_k\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle-$$
$$-L_k\tau_k\left\langle\eta_k\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\right.$$
$$\left.\left.\eta_k\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle\right)=$$
$$=\hat{g}_1(x_k,B_k)-\frac{\tau_k}{2}-\frac{\hat{g}_2(x_k,B_k)}{2\tau_k}+$$
$$+\frac{\eta_k(2-\eta_k)}{2\tau_k}\left\langle\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\ \hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle=$$
$$=\hat{g}_1(x_k,B_k)-\frac{\tau_k}{2}-\frac{\hat{g}_2(x_k,B_k)}{2\tau_k}+$$
$$+\frac{\eta_k(2-\eta_k)}{2\tau_k}\left\langle\hat{G}'(x_k,B_k)\left(L_k\tau_kI_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\ \hat{G}(x_k,B_k)\right\rangle.$$

□

Corollary 9.1. *If we take $\eta_k\in(0,2),\ \tau_k=\hat{g}_1(x_k,B_k)$, we automatically obtain local decrease on batch B_k:*

$$\hat{g}_1(x_k,B_k)-\hat{g}_1(x_{k+1},B_k)\geq$$
$$\geq\frac{\eta_k(2-\eta_k)}{2\hat{g}_1(x_k,B_k)}\left\langle\left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_kI_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\right.$$
$$\left.\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle\geq 0,$$

because matrix $\left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_kI_n\right)^{-1}$ is positive semi–definite.

Corollary 9.2. *Corollary 9.1 also holds for $\hat{g}_2$ in expectation:*

$$\hat{g}_1(x_k,B_k)-\hat{g}_1(x_{k+1},B_k)\geq$$
$$\geq\frac{\eta_k(2-\eta_k)}{2\hat{g}_1(x_k,B_k)}\left\langle\left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_kI_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\right.$$
$$\left.\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle\geq 0\Rightarrow\hat{g}_2(x_k,B_k)-\hat{g}_2(x_{k+1},B_k)\geq$$
$$\geq\hat{g}_2(x_k,B_k)-\hat{g}_1(x_k,B_k)\hat{g}_1(x_{k+1},B_k)\geq$$

$$\geq \frac{\eta_k(2-\eta_k)}{2}\left\langle \left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_kI_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\right.$$
$$\left.\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle \geq 0.$$

After averaging over the whole randomness we get:

$$\mathbb{E}\left[\hat{g}_2(x_k,B_k)-\hat{g}_2(x_{k+1},B_k)\right]=\mathbb{E}\left[\hat{f}_2(x_k)-\hat{g}_2(x_{k+1},B_k)\right]=\mathbb{E}\left[\hat{f}_2(x_k)\right]-\mathbb{E}\left[\hat{g}_2(x_{k+1},B_k)\right]\geq$$
$$\geq \mathbb{E}\left[\frac{\eta_k(2-\eta_k)}{2}\left\langle \left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_kI_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\right.\right.$$
$$\left.\left.\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle\right]\geq 0.$$

If we average the local decrease only over batches B_k, we can establish the following:

$$\sqrt{\mathbb{E}_{B_k}\left[\hat{g}_2(x_k,B_k)\right]}=\sqrt{\hat{f}_2(x_k)}=\hat{f}_1(x_k)\geq\sqrt{\mathbb{E}_{B_k}\left[\hat{g}_2(x_{k+1},B_k)\right]}\geq\{\textit{Jensen's inequality}\}\geq$$
$$\geq\mathbb{E}_{B_k}\left[\sqrt{\hat{g}_2(x_{k+1},B_k)}\right]\Rightarrow\hat{f}_1(x_k)\geq\mathbb{E}_{B_k}\left[\hat{g}_1(x_{k+1},B_k)\right],$$

because the value x_{k+1} depends on B_k.

Corollary 9.3 ([21])**.** *In deterministic settings ($B_k=\mathscr{B}$) we have analogous results:*

$$\begin{cases} x_{k+1}=x_k-\eta_k\left(\hat{F}'(x_k)^*\hat{F}'(x_k)+\tau_kL_kI_n\right)^{-1}\hat{F}'(x_k)^*\hat{F}(x_k);\\ \hat{f}_1(x_k)-\hat{f}_1(x_{k+1})\geq\hat{f}_1(x_k)-\dfrac{\tau_k}{2}-\dfrac{\hat{f}_2(x_k)}{2\tau_k}+\\ \qquad+\dfrac{\eta_k(2-\eta_k)}{2\tau_k}\left\langle\left(\hat{F}'(x_k)^*\hat{F}'(x_k)+\tau_kL_kI_n\right)^{-1}\hat{F}'(x_k)^*\hat{F}(x_k),\ \hat{F}'(x_k)^*\hat{F}(x_k)\right\rangle. \end{cases}$$

For $\eta_k\in(0,2),\ \tau_k=\hat{f}_1(x_k)$ we have

$$\hat{f}_1(x_k)-\hat{f}_1(x_{k+1})\geq$$
$$\geq\frac{\eta_k(2-\eta_k)}{2\hat{f}_1(x_k)}\left\langle\left(\hat{F}'(x_k)^*\hat{F}'(x_k)+\hat{f}_1(x_k)L_kI_n\right)^{-1}\hat{F}'(x_k)^*\hat{F}(x_k),\ \hat{F}'(x_k)^*\hat{F}(x_k)\right\rangle\geq 0,$$

because matrix $\left(\hat{F}'(x_k)^*\hat{F}'(x_k)+\hat{f}_1(x_k)L_kI_n\right)^{-1}$ is also positive semi–definite.

Lemmas 10 and 11 reveal main effects of batching in the optimization procedure relatively function value variance.

Lemma 10. *Suppose assumption 6 is satisfied. Under sampling without replacement of batches $B\subseteq\mathscr{B}$, $|B|=b$ from uniform distribution q over subsets B we have upper bound:*

$$\mathbb{E}_q\left[\left|\hat{g}_2(x,B)-\hat{f}_2(x)\right|^2\right]\leq\frac{\tilde{\sigma}^2}{b}\left(1-\frac{b}{m}\right),\ \forall x\in E_1,$$

for some finite $\tilde{\sigma}\geq\sigma$.

Proof. The expectation of $\hat{g}_2(x,B)$ over batch sample B can be represented using dependent Bernoulli random variables $Z_i \in \{0,1\}$, which encode exclusion of F_i using value 0 and inclusion of F_i into batch B using value 1:

$$\mathbb{E}_q\left[\hat{g}_2(x,B)\right] = \mathbb{E}_q\left[\frac{1}{b}\sum_{j=1}^{b}\left(F_{i_j}(x)\right)^2\right] = \frac{1}{b}\mathbb{E}\left[\sum_{i=1}^{m}\left(F_i(x)\right)^2 Z_i\right] = \frac{1}{b}\sum_{i=1}^{m}\left(F_i(x)\right)^2\mathbb{E}\left[Z_i\right] =$$
$$= \frac{1}{m}\sum_{i=1}^{m}\left(F_i(x)\right)^2 = \hat{f}_2(x),\ i_j \sim q,$$

because probability of picking F_i for sample B equals

$$\mathbb{P}\left(Z_i = 1\right) = \frac{C_{m-1}^{b-1}}{C_m^b} = \frac{(m-1)!}{(m-b)!(b-1)!}\frac{(m-b)!b!}{m!} = \frac{b}{m},\ i \in \overline{1,m}.$$

By definition of variance over finite population:

$$\mathbb{V}_q\left[\left(F_\xi(x)\right)^2\right] = \frac{1}{m}\sum_{i=1}^{m}\left(\left(F_i(x)\right)^2 - \hat{f}_2(x)\right)^2 = \frac{m-1}{m}\left(\sigma(x)\right)^2 \le \sigma^2,\ \xi \sim q,$$

$\sigma(x)$ — quasi–variance for sample B with $|B| = 1$ for arbitrary $x \in E_1$. By assumption 6: $\sigma(x) \le \sigma\sqrt{\frac{m}{m-1}}$. The variance of function g_2 value equals:

$$\mathbb{V}_q\left[\hat{g}_2(x,B)\right] = \mathbb{E}_q\left[\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|^2\right] = \mathbb{V}\left[\frac{1}{b}\sum_{i=1}^{m}\left(F_i(x)\right)^2 Z_i\right] =$$
$$= \frac{1}{b^2}\left(\sum_{i=1}^{m}\left(F_i(x)\right)^4\mathbb{V}\left[Z_i\right] + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(F_i(x)F_j(x)\right)^2\mathrm{Cov}(Z_i,Z_j)\right),$$

where summation over empty set considered to be equal zero. $Z_i,\ i \in \overline{1,m}$ are Bernoulli random variables, so $\mathbb{V}$ and Cov are defined in the following way:

$$\begin{cases}\mathbb{V}\left[Z_i\right] = \frac{b}{m}\left(1 - \frac{b}{m}\right);\\ \mathrm{Cov}(Z_i,Z_j) = \mathbb{E}\left[Z_iZ_j\right] - \mathbb{E}\left[Z_i\right]\mathbb{E}\left[Z_j\right] = \frac{C_{m-2}^{b-2}}{C_m^b} - \left(\frac{b}{m}\right)^2 = \frac{b(b-1)}{m(m-1)} - \left(\frac{b}{m}\right)^2.\end{cases}$$

Substitute these values into $\mathbb{V}_q\left[\hat{g}_2(x,B)\right]$:

$$\mathbb{V}_q\left[\hat{g}_2(x,B)\right] = \frac{1}{b^2}\left(\sum_{i=1}^{m}\left(F_i(x)\right)^4\mathbb{V}\left[Z_i\right] + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(F_i(x)F_j(x)\right)^2\mathrm{Cov}(Z_i,Z_j)\right) =$$
$$= \frac{1}{b^2}\left(\sum_{i=1}^{m}\left(F_i(x)\right)^4\frac{b}{m}\left(1-\frac{b}{m}\right) +\right.$$
$$\left. +2\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(F_i(x)F_j(x)\right)^2\left(\frac{b(b-1)}{m(m-1)} - \left(\frac{b}{m}\right)^2\right)\right) =$$
$$= \frac{(m-b)}{mb}\left(\frac{1}{m}\sum_{i=1}^{m}\left(F_i(x)\right)^4 - \frac{2}{m(m-1)}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(F_i(x)F_j(x)\right)^2\right) =$$

$$= \frac{1}{b}\left(1-\frac{b}{m}\right)\left(\frac{1}{m-1}\left(\sum_{i=1}^{m}(F_i(x))^2 - \hat{f}_2(x)\right)^2\right) = \frac{(\sigma(x))^2}{b}\left(1-\frac{b}{m}\right).$$

Define $\tilde{\sigma} \overset{\text{def}}{=} \sigma\sqrt{\frac{m}{m-1}}$, then

$$\mathbb{E}_q\left[\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|^2\right] = \frac{(\sigma(x))^2}{b}\left(1-\frac{b}{m}\right) \leq \frac{m\sigma^2}{b(m-1)}\left(1-\frac{b}{m}\right) = \frac{\tilde{\sigma}^2}{b}\left(1-\frac{b}{m}\right), \quad \forall x \in E_1.$$

□

Corollary 10.1. *The estimate obtained can be straightforwardly generalized for infinite population:*

$$\mathbb{E}_q\left[\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|^2\right] \leq \lim_{m\to+\infty}\left[\frac{\tilde{\sigma}^2}{b}\left(1-\frac{b}{m}\right)\right] = \frac{\tilde{\sigma}^2}{b}, \ \forall x \in E_1.$$

This estimate coincides with estimate for case of sampling with replacement:

$$\mathbb{V}_q\left[\hat{g}_2(x,B)\right] = \mathbb{V}_q\left[\frac{1}{b}\sum_{j=1}^{b}\left(F_{i_j}(x)\right)^2\right] = \frac{1}{b^2}\sum_{j=1}^{b}\mathbb{V}_q\left[\left(F_{i_j}(x)\right)^2\right] \leq \frac{\sigma^2}{b} \leq \frac{\tilde{\sigma}^2}{b},$$

while $\lim\limits_{m\to+\infty}[\tilde{\sigma}] = \sigma$.

Corollary 10.2. *The lemma conditions bounds* $\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|$ *for all* $x \in E_1$*:*

$$\mathbb{E}_q\left[\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|\right] = \mathbb{E}_q\left[\sqrt{\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|^2}\right] \leq \sqrt{\mathbb{E}_q\left[\left|\hat{g}_2(x,B) - \hat{f}_2(x)\right|^2\right]} \leq$$
$$\leq \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}.$$

Lemma 11. *Let assumptions 3 and 6 hold for sequence* $\{x_{k-1}\}_{k\in\mathbb{N}}$, $x_{k-1} \in E_1$, *obtained using one of the rules:* (7) *or* (8)*. Under independent sampling without replacement of* $B_{k-1} \subseteq \mathscr{B}$, $|B_{k-1}| = b$ *from uniform distribution over subsets* B_{k-1} *for each* $k \in \mathbb{N}$ *we have*

$$\mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] \leq 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right]\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}},$$

for some finite $\tilde{\sigma} \geq \sigma$*. Averaging is done over samples* B_{k-1}, $k \in \mathbb{N}$ *and initialization.*

Proof. Firstly, we express upper bound:

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] = \mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{f}_2(x_{k-1}) + \hat{f}_2(x_{k-1}) - \hat{g}_2(x_{k-1}, B_{k-1}) + \right.\right. \\
&\left.\left. + \hat{g}_2(x_{k-1}, B_{k-1}) - \hat{g}_2(x_k, B_{k-1})\right|\right] \leq \mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{f}_2(x_{k-1})\right|\right] + \\
&+ \mathbb{E}\left[\left|\hat{f}_2(x_{k-1}) - \hat{g}_2(x_{k-1}, B_{k-1})\right|\right] + \mathbb{E}\left[\left|\hat{g}_2(x_{k-1}, B_{k-1}) - \hat{g}_2(x_k, B_{k-1})\right|\right] \leq \\
&\leq \left\{\text{lipschitzness of } \hat{g}_2 \text{ and } \hat{f}_2\right\} \leq 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \mathbb{E}\left[\sqrt{\left|\hat{f}_2(x_{k-1}) - \hat{g}_2(x_{k-1}, B_{k-1})\right|^2}\right] \leq \\
&\leq 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \sqrt{\mathbb{E}\left[\left|\hat{f}_2(x_{k-1}) - \hat{g}_2(x_{k-1}, B_{k-1})\right|^2\right]} \leq \\
&\leq \{\text{lemma 10}\} \leq 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \sqrt{\frac{\tilde{\sigma}^2}{b}\left(1-\frac{b}{m}\right)} = 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}} =
\end{aligned}$$

$$= 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right]\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}, \tag{23}$$

because for $b = m$ we have $\hat{f}_2(x_k) = \hat{g}_2(x_k, \mathscr{B}) = \hat{g}_2(x_k, B_{k-1})$. Sometimes, it is convenient to express (23) in the following way:

$$\begin{aligned}\mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] &\leq 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right]\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}} \leq \\ &\leq \sqrt{\frac{1}{b}-\frac{1}{m}}\left(2l_{\hat{F}}\sqrt{m(m-1)}\mathbb{E}\left[\|x_k - x_{k-1}\|\right]\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\right).\end{aligned} \tag{24}$$

□

Corollary 11.1. *As in lemma 10, the proved estimate has a natural generalization for infinite population:*

$$\begin{aligned}\mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] &\leq \lim_{m\to+\infty}\left[2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right] = \\ &= 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \frac{\tilde{\sigma}}{\sqrt{b}},\ \forall k \in \mathbb{N}.\end{aligned}$$

Analogously this estimate coincides with the case of sampling with replacement and can be obtained using corollary 10.1:

$$\begin{aligned}\mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] &\leq \{(23)\} \leq 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \\ &+ \sqrt{\mathbb{E}\left[\left|\hat{f}_2(x_{k-1}) - \hat{g}_2(x_{k-1}, B_{k-1})\right|^2\right]} \leq \\ &\leq \{\textit{lemma 10, corollary 10.1}\} \leq 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right] + \frac{\tilde{\sigma}}{\sqrt{b}},\end{aligned}$$

while $\lim\limits_{m\to+\infty}[\tilde{\sigma}] = \sigma$.

Corollary 11.2. *For* $|B_{k-1}| = m,\ k \in \mathbb{N}$ *we have* $\mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] = 0$.

The lemma below represents local model relatively the step (7).

Lemma 12. *Let assumption 3 holds for sequence* $\{x_k\}_{k\in\mathbb{Z}_+}$, $x_k \in E_1$ *obtained using* (7) *with* $\tau_k > 0$, $L_k > 0$, $B_k \subseteq \mathscr{B}$, $\eta_k \in (0,2)$. *Then, for arbitrary* $y \in E_1$ *we have*

$$\begin{aligned}\hat{\psi}_{x_k,L_k,\tau_k}(y, B_k) = \hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1}, B_k) + \frac{L_k}{2}\|y - x_{k+1}\|^2 + \\ + \frac{1}{2\tau_k}\left\|\hat{G}'(x_k, B_k)(y - x_{k+1})\right\|^2 + \frac{1-\eta_k}{2\tau_k}\left\langle y - x_{k+1},\ \nabla_{x_k}\hat{g}_2(x_k, B_k)\right\rangle.\end{aligned}$$

Proof. Firstly, we rewrite $\hat{\psi}_{x_k,L_k,\tau_k}(y, B_k)$:

$$\begin{aligned}&\hat{\psi}_{x_k,L_k,\tau_k}(y, B_k) = \frac{\tau_k}{2} + \frac{L_k}{2}\|y - x_k\|^2 + \frac{1}{2\tau_k}\left\|\hat{G}(x_k, B_k) + \hat{G}'(x_k, B_k)(y - x_k)\right\|^2 = \frac{\tau_k}{2} + \\ &+ \frac{L_k}{2}\|(y - x_{k+1}) + (x_{k+1} - x_k)\|^2 + \frac{1}{2\tau_k}\left\|\hat{G}(x_k, B_k) + \hat{G}'(x_k, B_k)((y - x_{k+1}) + (x_{k+1} - x_k))\right\|^2 = \\ &= \frac{\tau_k}{2} + \frac{L_k}{2}\|y - x_{k+1}\|^2 + L_k\left\langle y - x_{k+1},\ x_{k+1} - x_k\right\rangle + \frac{L_k}{2}\|x_{k+1} - x_k\|^2 + \\ &+ \frac{1}{2\tau_k}\left\|\left(\hat{G}(x_k, B_k) + \hat{G}'(x_k, B_k)(x_{k+1} - x_k)\right) + \hat{G}'(x_k, B_k)(y - x_{k+1})\right\|^2 = \\ &= \left(\frac{\tau_k}{2} + \frac{L_k}{2}\|x_{k+1} - x_k\|^2 + \frac{1}{2\tau_k}\left\|\hat{G}(x_k, B_k) + \hat{G}'(x_k, B_k)(x_{k+1} - x_k)\right\|^2\right) + \frac{L_k}{2}\|y - x_{k+1}\|^2 +\end{aligned}$$

$$+\langle y-x_{k+1},\, L_k(x_{k+1}-x_k)\rangle+\frac{1}{\tau_k}\left\langle \hat{G}'(x_k,B_k)(y-x_{k+1}),\, \hat{G}(x_k,B_k)+\hat{G}'(x_k,B_k)(x_{k+1}-x_k)\right\rangle+$$
$$+\frac{1}{2\tau_k}\left\|\hat{G}'(x_k,B_k)(y-x_{k+1})\right\|^2=\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)+\frac{L_k}{2}\|y-x_{k+1}\|^2+$$
$$+\frac{1}{2\tau_k}\left\|\hat{G}'(x_k,B_k)(y-x_{k+1})\right\|^2+$$
$$+\left\langle y-x_{k+1},\underbrace{L_k(x_{k+1}-x_k)+\frac{1}{\tau_k}\hat{G}'(x_k,B_k)^*\left(\hat{G}(x_k,B_k)+\hat{G}'(x_k,B_k)(x_{k+1}-x_k)\right)}_{=\nabla_{x_{k+1}}\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)}\right\rangle=$$
$$=\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)+\frac{L_k}{2}\|y-x_{k+1}\|^2+\frac{1}{2\tau_k}\left\|\hat{G}'(x_k,B_k)(y-x_{k+1})\right\|^2+$$
$$+\frac{1}{2\tau_k}\left\langle y-x_{k+1},\, 2\left(\left(\tau_k L_k I_n+\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)\right)(x_{k+1}-x_k)+\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right)\right\rangle=$$
$$=\left\{x_{k+1}-x_k=-\eta_k\left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\tau_k L_k I_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\}=$$
$$=\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)+\frac{L_k}{2}\|y-x_{k+1}\|^2+\frac{1}{2\tau_k}\left\|\hat{G}'(x_k,B_k)(y-x_{k+1})\right\|^2+$$
$$+\frac{1}{2\tau_k}\left\langle y-x_{k+1},\,(1-\eta_k)\left(2\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right)\right\rangle=\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)+\frac{L_k}{2}\|y-x_{k+1}\|^2+$$
$$+\frac{1}{2\tau_k}\left\|\hat{G}'(x_k,B_k)(y-x_{k+1})\right\|^2+\frac{1-\eta_k}{2\tau_k}\left\langle y-x_{k+1},\,\nabla_{x_k}\hat{g}_2(x_k,B_k)\right\rangle.$$

□

Corollary 12.1. *For $\eta_k=1$ the representation obtained allows us to estimate closeness to the global minimum of $\hat{\psi}_{x_k,L_k,\tau_k}(\cdot,B_k)$, if we define $\hat{x}_{k+1}\in E_1$ as an approximate value of x_{k+1} computed with error $\varepsilon_k\geq 0$ and use the following representation of difference $\hat{\psi}_{x_k,L_k,\tau_k}(y,B_k)-\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)$:*

$$0\leq\hat{\psi}_{x_k,L_k,\tau_k}(\hat{x}_{k+1},B_k)-\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)=\frac{L_k}{2}\|\hat{x}_{k+1}-x_{k+1}\|^2+$$
$$+\frac{1}{2\tau_k}\left\|\hat{G}'(x_k,B_k)(\hat{x}_{k+1}-x_{k+1})\right\|^2\leq\varepsilon_k,\; x_{k+1}=\hat{T}_{L_k,\tau_k}(x_k,B_k).$$

Corollary 12.2. *In the deterministic case ($\hat{G}(x,\mathscr{B})=\hat{F}(x),\; x\in E_1$) we have an analogous representation for $\psi_{x_k,L_k,\tau_k}(y),\; y\in E_1$:*

$$\psi_{x_k,L_k,\tau_k}(y)=\psi_{x_k,L_k,\tau_k}(x_{k+1})+\frac{L_k}{2}\|y-x_{k+1}\|^2+\frac{1}{2\tau_k}\left\|\hat{F}'(x_k)(y-x_{k+1})\right\|^2+$$
$$+\frac{1-\eta_k}{2\tau_k}\left\langle y-x_{k+1},\,\nabla\hat{f}_2(x_k)\right\rangle.$$

Lemma 13 describes bounds of variation for the sequence $\{x_k\}_{k\in\mathbb{Z}_+}$ under considered update rules.

Lemma 13. *Let assumptions 3 and 4 hold for sequence $\{x_k\}_{k\in\mathbb{Z}_+}$ obtained using update rule* (7)*, $\tau_k=\hat{g}_1(x_k,B_k)$, $\eta_k\in(0,1]$, $L_k>0$. Then we have bounds of sequence variation:*

$$\|x_{k+1}-x_k\|\in\left[\frac{\eta_k\|\nabla_{x_k}\hat{g}_2(x_k,B_k)\|}{2\left(M_{\hat{G}}^2+\hat{g}_1(x_k,B_k)L_k\right)},\,\min\left\{\sqrt{\frac{2\hat{g}_1(x_k,B_k)}{L_k}},\,\frac{\eta_k M_{\hat{G}}}{L_k}\right\}\right],\; k\in\mathbb{Z}_+.$$

In case of update rule (8) *variation* $\|x_{k+1}-x_k\|$ *is bounded in the following way:*

$$\|x_{k+1}-x_k\| \in \left[\frac{\eta_k \|\nabla_{x_k}\hat{g}_2(x_k,B_k)\|}{2\left(M^2_{\hat{G}}+\tilde{\tau}_k L_k\right)},\ \frac{\eta_k M_{\hat{G}}\hat{g}_1(x_k,B_k)}{\tilde{\tau}_k L_k}\right],\ k\in\mathbb{Z}_+.$$

Proof. According to (7) we define $\tilde{B}_k = B_k$, $\tilde{\tau}_k = \tau_k$ to get

$$\begin{aligned}
&\|x_{k+1}-x_k\| = \left\|-\eta_k\left(\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k)+\tilde{\tau}_k L_k I_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\| = \\
&= \eta_k\left\|\underbrace{\left(\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k)+\tilde{\tau}_k L_k I_n\right)^{-1}}_{\text{symmetric matrix}}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\| = \\
&= \eta_k\left(\left\langle\left(\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k)+\tilde{\tau}_k L_k I_n\right)^{-2}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\ \hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle\right)^{\frac{1}{2}} = \\
&= \left\{\nabla_{x_k}\hat{g}_2(x_k,B_k) = 2\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\} = \\
&= \frac{\eta_k}{2}\sqrt{\left\langle\left(\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k)+\tilde{\tau}_k L_k I_n\right)^{-2}\nabla_{x_k}\hat{g}_2(x_k,B_k),\ \nabla_{x_k}\hat{g}_2(x_k,B_k)\right\rangle} \geq \\
&\geq \{\text{assumption 4, (20)}\} \geq \frac{\eta_k\|\nabla_{x_k}\hat{g}_2(x_k,B_k)\|}{2\left(M^2_{\hat{G}}+\tilde{\tau}_k L_k\right)} \geq \{\tilde{\tau}_k = \hat{g}_1(x_k,B_k)\} \geq \frac{\eta_k\|\nabla_{x_k}\hat{g}_2(x_k,B_k)\|}{2\left(M^2_{\hat{G}}+\hat{g}_1(x_k,B_k)L_k\right)}.
\end{aligned} \tag{25}$$

So, the formulas we got hold for the case of update rule (8) even for $\tilde{B}_k \neq B_k$. Now consider the upper bound:

$$\begin{aligned}
&\hat{\psi}_{x_k,L_k,\hat{g}_1(x_k,B_k)}(x_k,B_k) = \hat{g}_1(x_k,B_k) = \{\text{lemma 12}\} = \hat{\psi}_{x_k,L_k,\hat{g}_1(x_k,B_k)}(x_{k+1},B_k)+ \\
&+\frac{L_k}{2}\|x_k-x_{k+1}\|^2 + \\
&+\frac{1}{2\hat{g}_1(x_k,B_k)}\left\|\hat{G}'(x_k,B_k)(x_k-x_{k+1})\right\|^2+\frac{1-\eta_k}{2\hat{g}_1(x_k,B_k)}\langle x_k-x_{k+1},\ \nabla_{x_k}\hat{g}_2(x_k,B_k)\rangle = \\
&= \hat{\psi}_{x_k,L_k,\hat{g}_1(x_k,B_k)}(x_{k+1},B_k)+\frac{L_k}{2}\|x_k-x_{k+1}\|^2+\frac{1}{2\hat{g}_1(x_k,B_k)}\left\|\hat{G}'(x_k,B_k)(x_k-x_{k+1})\right\|^2 + \\
&+\frac{(1-\eta_k)\eta_k}{4\hat{g}_1(x_k,B_k)}\left\langle\left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_k I_n\right)^{-1}\nabla_{x_k}\hat{g}_2(x_k,B_k),\ \nabla_{x_k}\hat{g}_2(x_k,B_k)\right\rangle \geq 0.
\end{aligned} \tag{26}$$

The expression above leads to

$$\begin{aligned}
&\hat{g}_1(x_k,B_k) \geq \hat{g}_1(x_k,B_k)-\hat{\psi}_{x_k,L_k,\hat{g}_1(x_k,B_k)}(x_{k+1},B_k) \geq \frac{L_k}{2}\|x_{k+1}-x_k\|^2 \Rightarrow \\
&\Rightarrow \|x_{k+1}-x_k\| \leq \sqrt{\frac{2\hat{g}_1(x_k,B_k)}{L_k}}.
\end{aligned}$$

There is also exists another upper bound with defined $\tilde{B}_k = B_k$, $\tilde{\tau}_k = \tau_k$:

$$\|x_{k+1}-x_k\| = \left\|-\eta_k\left(\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k)+\tilde{\tau}_k L_k I_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\| =$$

$$= \eta_k \left(\left\langle \underbrace{\left(\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k) + \tilde{\tau}_k L_k I_n\right)^{-2}}_{\tilde{B}_k \text{ can be sampled independently from } B_k} \hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\ \hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k) \right\rangle \right)^{\frac{1}{2}} \leq$$

$$\leq \frac{\eta_k}{\tilde{\tau}_k L_k} \left\| \hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k) \right\| \leq \frac{\eta_k}{\tilde{\tau}_k L_k} \left\| \hat{G}'(x_k,B_k)^* \right\| \left\| \hat{G}(x_k,B_k) \right\| \leq \frac{\eta_k M_{\hat{G}} \hat{g}_1(x_k,B_k)}{\tilde{\tau}_k L_k} =$$

$$= \{\tau_k = \hat{g}_1(x_k,B_k)\} = \frac{\eta_k M_{\hat{G}}}{L_k}. \tag{27}$$

Expressions in (27) are applicable for the rule (8) allowing us to deduce upper bound on $\|x_{k+1} - x_k\|$ in case of (8). □

Corollary 13.1. *For $\tau_k \in \left[\tilde{\tau},\ \tilde{\mathscr{T}}\right]$, $\tilde{\tau} \in \left(0,\ \tilde{\mathscr{T}}\right]$, $L_k \in \left[L,\ \tilde{\gamma}L_{\hat{F}}\right]$, $L \in \left(0,\ \tilde{\gamma}L_{\hat{F}}\right]$, $\tilde{\gamma} \geq 1$ and under assumption 5 the value $\|x_{k+1} - x_k\|$ obtained using update rule* (7) *is bounded:*

$$\|x_{k+1} - x_k\| \in \left[\frac{\eta_k \|\nabla_{x_k} \hat{g}_2(x_k,B_k)\|}{2\left(M_{\hat{G}}^2 + \tilde{\gamma}\tilde{\mathscr{T}}L_{\hat{F}}\right)},\ \min\left\{ \sqrt{\frac{1}{L}\left(\tilde{\mathscr{T}} + \frac{P_{\hat{g}_1}^2}{\tilde{\tau}}\right)},\ \frac{\eta_k M_{\hat{G}} P_{\hat{g}_1}}{\tau_k L_k} \right\} \right].$$

Lower bound is obtained from (25) *using monotone decrease over $\tau_k L_k$. Upper bound $\frac{\eta_k M_{\hat{G}} P_{\hat{g}_1}}{\tau_k L_k}$ is deduced from* (27) *using assumptions 4 and 5. Upper bound*

$$\sqrt{\frac{1}{L}\left(\tilde{\mathscr{T}} + \frac{P_{\hat{g}_1}^2}{\tilde{\tau}}\right)}$$

is expressed via (26) *for local model $\hat{\psi}_{x_k,L_k,\tau_k}(\cdot,B_k)$ under assumption 5:*

$$\frac{\tilde{\mathscr{T}}}{2} + \frac{P_{\hat{g}_1}^2}{2\tilde{\tau}} \geq \hat{\psi}_{x_k,L_k,\tau_k}(x_k,B_k) = \frac{\tau_k}{2} + \frac{\hat{g}_2(x_k,B_k)}{2\tau_k} \geq \frac{L_k}{2}\|x_{k+1} - x_k\|^2 \geq \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Corollary 13.2. *Under assumption 5 and $L_k \in [L,\ \gamma L_{\hat{F}}]$, $L \in (0,\ \gamma L_{\hat{F}}]$, $\gamma \geq 1$ we can bound the value $\|x_{k+1} - x_k\|$ obtained using update rule* (7) *in the following way:*

$$\|x_{k+1} - x_k\| \in \left[\frac{\eta_k \|\nabla_{x_k} \hat{g}_2(x_k,B_k)\|}{2\left(M_{\hat{G}}^2 + \gamma P_{\hat{g}_1} L_{\hat{F}}\right)},\ \min\left\{ \sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{\eta_k M_{\hat{G}}}{L} \right\} \right],\ k \in \mathbb{Z}_+.$$

Corollary 13.3. *In the deterministic setting $\hat{G}(x_k,B_k) = \hat{F}(x_k)$, $\tilde{B}_k = B_k = \mathscr{B}$ and $\tilde{\tau}_k = \tau_k = \hat{f}_1(x_k)$ we have bounded variation for the sequence $\{x_k\}_{k\in\mathbb{Z}_+}$ built using 9.3:*

$$\|x_{k+1} - x_k\| \in \left[\frac{\eta_k \left\|\nabla \hat{f}_2(x_k)\right\|}{2\left(M_{\hat{F}}^2 + \hat{f}_1(x_k)L_k\right)},\ \min\left\{ \sqrt{\frac{2\hat{f}_1(x_k)}{L_k}},\ \frac{\eta_k M_{\hat{G}}}{L_k} \right\} \right],\ k \in \mathbb{Z}_+.$$

Corollary 13.4. *Under constant step scale $\eta_k = \eta = \text{const}$, $0 < L_k \leq \gamma L_{\hat{F}}$, $\gamma \geq 1$, $k \in \mathbb{Z}_+$ and under assumptions 3, 4 and 5 the lower bound for $\|x_{k+1} - x_k\|$ obtained using update rule* (7) *is proportional to the norm of gradient of optimization criterion and can be used as stopping criterion to achieve level*

$\varepsilon > 0$ *of the gradient norm:*

$$\mathbb{E}\left[\left\|\nabla \hat{f}_2(x_k)\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|\nabla \hat{f}_2(x_k)\right\|^2\right]} \leq \sqrt{\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla_{x_k} \hat{g}_2(x_k, B_k)\right\|^2\right]\right]} =$$
$$= \sqrt{\mathbb{E}\left[\left\|\nabla_{x_k} \hat{g}_2(x_k, B_k)\right\|^2\right]} \leq \frac{2\left(M_{\hat{G}}^2 + \gamma P_{\hat{g}_1} L_{\hat{F}}\right)}{\eta} \sqrt{\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right]} \leq \varepsilon,$$

it means

$$\mathbb{E}\left[\|x_{k+1} - x_k\|\right] \leq \sqrt{\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right]} \leq \frac{\varepsilon\eta}{2\left(M_{\hat{G}}^2 + \gamma P_{\hat{g}_1} L_{\hat{F}}\right)},$$

where we average over the whole randomness of the optimization procedure. In cases of update rule (8) *with* $\tilde{\tau}_k \leq \tilde{\mathscr{T}}$ *this condition transforms in the following way:*

$$\mathbb{E}\left[\|x_{k+1} - x_k\|\right] \leq \sqrt{\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right]} \leq \frac{\varepsilon\eta}{2\left(M_{\hat{G}}^2 + \gamma \tilde{\mathscr{T}} L_{\hat{F}}\right)}.$$

Lemma 14 presents the lipschitzness of gradients of bounded functions $\hat{f}_2$ and $\hat{g}_2$.

Lemma 14. *Let assumptions 3, 4, 5 hold. Then function* $\hat{g}_2$ *has Lipschitz gradient with an upper estimate of the Lipschitz constant* $l_{\hat{g}_2} = 2\left(M_{\hat{G}}^2 + L_{\hat{F}} P_{\hat{g}_1}\right)$.

Proof. We compute $l_{\hat{g}_2}$ — an upper estimate for the best (the lowest) Lipschitz constant for arbitrary $(x, y) \in E_1^2$ and $B \subseteq \mathscr{B}$:

$$\left\|\nabla_y \hat{g}_2(y, B) - \nabla_x \hat{g}_2(x, B)\right\| = \left\|2\hat{G}'(y, B)^* \hat{G}(y, B) - 2\hat{G}'(x, B)^* \hat{G}(x, B)\right\| =$$
$$= 2\left\|\left(\hat{G}'(y, B)^* \hat{G}(y, B) - \hat{G}'(x, B)^* \hat{G}(y, B)\right) + \left(\hat{G}'(x, B)^* \hat{G}(y, B) - \hat{G}'(x, B)^* \hat{G}(x, B)\right)\right\| \leq$$
$$\leq 2\left(\left\|\hat{G}'(y, B)^* \hat{G}(y, B) - \hat{G}'(x, B)^* \hat{G}(y, B)\right\| + \left\|\hat{G}'(x, B)^* \hat{G}(y, B) - \hat{G}'(x, B)^* \hat{G}(x, B)\right\|\right) \leq$$
$$\leq 2\left(\left\|\hat{G}'(y, B)^* - \hat{G}'(x, B)^*\right\| \left\|\hat{G}(y, B)\right\| + \left\|\hat{G}'(x, B)^*\right\| \left\|\hat{G}(y, B) - \hat{G}(x, B)\right\|\right) \leq$$
$$\leq 2\left(L_{\hat{F}} \|y - x\| P_{\hat{g}_1} + M_{\hat{G}}^2 \|y - x\|\right) \leq \left(2\left(L_{\hat{F}} P_{\hat{g}_1} + M_{\hat{G}}^2\right)\right) \|y - x\| \Rightarrow l_{\hat{g}_2} = 2\left(L_{\hat{F}} P_{\hat{g}_1} + M_{\hat{G}}^2\right).$$

We use the lipschitzness of multidimensional map $\hat{G}$ from above:

$$\left\|\hat{G}(y, B) - \hat{G}(x, B)\right\| = \left\|\int_0^1 \hat{G}'(x + t(y - x), B)(y - x)\, \mathrm{d}t\right\| \leq$$
$$\leq \int_0^1 \left\|\hat{G}'(x + t(y - x), B)\right\| \|y - x\|\, \mathrm{d}t \leq M_{\hat{G}} \|y - x\|.$$

□

Corollary 14.1. *For* $B = \mathscr{B}$ *function* $\hat{f}_2$ *has Lipschitz gradient with the Lipschitz constant estimate* $l_{\hat{f}_2} \stackrel{\text{def}}{=} 2\left(M_{\hat{F}}^2 + L_{\hat{F}} P_{\hat{f}_1}\right)$.

Lemma 15 justifies local model used in doubly stochastic step analysis.

Lemma 15. *Suppose assumptions 3, 4, 5 hold. Then there is exists the following stochastic local model for function $\hat{g}_2$:*

$$\hat{g}_2(y,B) \leq \hat{\varphi}_{x,l}(y,B) = \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \frac{l}{2}\|y-x\|^2,\ \forall l \geq l_{\hat{g}_2},$$
$$\forall (x,y) \in E_1^2,\ \forall B \subseteq \mathscr{B}.$$

Proof. Consider an upper estimate for $\hat{g}_2(x,B)$ under arbitrary $(x,y) \in E_1^2$ and $B \subseteq \mathscr{B}$:

$$\begin{aligned}
\hat{g}_2(y,B) &= \hat{g}_2(y,B) - \hat{g}_2(x,B) - \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle \leq \\
&\leq \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + |\hat{g}_2(y,B) - \hat{g}_2(x,B) - \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle| = \\
&= \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \\
&+ \left| \int_0^1 \left\langle \nabla_{x+t(y-x)} \hat{g}_2(x+t(y-x),B),\, y-x \right\rangle \mathrm{d}t - \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle \right| = \\
&= \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \\
&+ \left| \int_0^1 \left\langle \nabla_{x+t(y-x)} \hat{g}_2(x+t(y-x),B) - \nabla_x \hat{g}_2(x,B),\, y-x \right\rangle \mathrm{d}t \right| \leq \\
&\leq \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \\
&+ \int_0^1 \left| \left\langle \nabla_{x+t(y-x)} \hat{g}_2(x+t(y-x),B) - \nabla_x \hat{g}_2(x,B),\, y-x \right\rangle \right| \mathrm{d}t \leq \\
&\leq \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \\
&+ \int_0^1 \left\| \nabla_{x+t(y-x)} \hat{g}_2(x+t(y-x),B) - \nabla_x \hat{g}_2(x,B) \right\| \|y-x\| \mathrm{d}t \leq \\
&\leq \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \int_0^1 t l_{\hat{g}_2} \|y-x\|^2 \mathrm{d}t = \\
&= \underbrace{\hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \frac{l_{\hat{g}_2}}{2}\|y-x\|^2}_{=\hat{\varphi}_{x,l_{\hat{g}_2}}(y,B)} \leq \\
&\leq \hat{g}_2(x,B) + \langle \nabla_x \hat{g}_2(x,B),\, y-x\rangle + \frac{l}{2}\|y-x\|^2,\ l \geq l_{\hat{g}_2}.
\end{aligned}$$

□

Corollary 15.1. *In case of $B = \mathscr{B}$ the local model for $\hat{g}_2$ morphs into the local model for $\hat{f}_2$:*

$$\hat{f}_2(y) \leq \varphi_{x,l}(y) \stackrel{\text{def}}{=} \hat{f}_2(x) + \left\langle \nabla \hat{f}_2(x),\, y-x \right\rangle + \frac{l}{2}\|y-x\|^2,\ \forall l \geq l_{\hat{f}_2},\ \forall (x,y) \in E_1^2.$$

Lemma 16 bounds of the gradient norm under WGC and PL condition.

Lemma 16. *Let assumptions 4 and 7 are satisfied. Then the squared norm of gradient of $\hat{g}_2$ is bounded from both sides with $\hat{g}_2$:*

$$4\mu\hat{g}_2(x,B) \leq \|\nabla_x\hat{g}_2(x,B)\|^2 \leq 4M_{\hat{G}}^2\hat{g}_2(x,B),\ \forall x \in E_1,\ \forall B \subseteq \mathscr{B}.$$

Proof. Conditions 4 and 7 state the following inequalities:

$$\begin{aligned} &4\mu\hat{g}_2(x,B) \leq \{\text{assumption 7}\} \leq 4\left\|\hat{G}'(x,B)^*\hat{G}(x,B)\right\|^2 = \|\nabla_x\hat{g}_2(x,B)\|^2 \leq \\ &\leq 4\left\|\hat{G}'(x,B)\right\|^2\left\|\hat{G}(x,B)\right\|^2 \leq \{\text{assumption 4}\} \leq \\ &\leq 4M_{\hat{G}}^2\hat{g}_2(x,B),\ \forall x \in E_1, \forall B \subseteq \mathscr{B} \Rightarrow \mu \leq M_{\hat{G}}^2,\ 4\mu\hat{g}_2(x,B) \leq \|\nabla_x\hat{g}_2(x,B)\|^2. \end{aligned} \tag{28}$$

□

Corollary 16.1. *The averaged over batches B squared norm of the gradient of function $\hat{g}_2$ is also bounded from both sides:*

$$4\mu\hat{f}_2(x) \leq \mathbb{E}_B\left[\|\nabla_x\hat{g}_2(x,B)\|^2\right] \leq 4M_{\hat{G}}^2\hat{f}_2(x),\ \forall x \in E_1.$$

D The proof of results for stochastic Gauss–Newton method

The general Gauss–Newton method with scaled step uses the update rule based on direct minimization of the local model $\hat{\psi}_{x_k,L_k,\tau_k}(y,B_k)$ over $y \in E_1$ (7), where the minimal value of $\hat{\psi}_{x_k,L_k,\tau_k}(x_{k+1},B_k)$ is obtained at $\eta_k = 1$. This framework is described via settings (9), and conceptual scheme of the framework forms up algorithm 2.

D.1 The proof of theorem 4

Theorem 4 proves sublinear global convergence rate to approximate stationary point in mean.

Theorem 4. *Suppose assumptions 3, 4, 5, 6 are satisfied. Consider Stochastic Gauss–Newton method 2 with $\tau_k = \hat{g}_1(x_k,B_k)$, $\eta_k \in [\eta,1]$, $\eta \in (0,1]$ and some finite $\tilde{\sigma} \geq \sigma$. Then:*

$$\mathbb{E}\left[\min_{i\in\overline{0,k-1}}\left\|\nabla\hat{f}_2(x_i)\right\|^2\right] \leq \frac{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}{\eta(2-\eta)}\left(\frac{\mathbb{E}\left[\hat{f}_2(x_0)\right]}{k} + 2l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \right.$$
$$\left. + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right),\ k \in \mathbb{N}. \tag{29}$$

Expectation operator $\mathbb{E}[\cdot]$ averages over all randomness in optimization procedure.

Proof. According to update rule for x_k (7), (21):

$$\begin{aligned} &\hat{g}_1(x_k,B_k) - \hat{g}_1(x_{k+1},B_k) \geq \hat{g}_1(x_k,B_k) - \frac{\tau_k}{2} - \frac{\hat{g}_2(x_k,B_k)}{2\tau_k} + \\ &+ \frac{\eta_k(2-\eta_k)}{2\tau_k}\left\langle\left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k) + \tau_k L_k I_n\right)^{-1}\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\right. \\ &\left.\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)\right\rangle \geq \\ &\geq \left\{\tau_k = \hat{g}_1(x_k,B_k),\ \eta_k \geq \eta,\ \nabla_{x_k}\hat{g}_2(x_k,B_k) = 2\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k),\ \text{corollary 9.1}\right\} \geq \end{aligned}$$

$$\geq \frac{\eta(2-\eta)}{8\hat{g}_1(x_k,B_k)}\left\langle \left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_kI_n\right)^{-1}\nabla_{x_k}\hat{g}_2(x_k,B_k),\right.$$
$$\left.\nabla_{x_k}\hat{g}_2(x_k,B_k)\right\rangle \geq 0 \Rightarrow$$
$$\Rightarrow \hat{g}_2(x_k,B_k)-\hat{g}_2(x_{k+1},B_k)\geq \hat{g}_2(x_k,B_g)-\hat{g}_1(x_k,B_k)\hat{g}_1(x_{k+1},B_k)\geq$$
$$\geq \frac{\eta(2-\eta)}{8}\left\langle \left(\hat{G}'(x_k,B_k)^*\hat{G}'(x_k,B_k)+\hat{g}_1(x_k,B_k)L_kI_n\right)^{-1}\nabla_{x_k}\hat{g}_2(x_k,B_k),\ \nabla_{x_k}\hat{g}_2(x_k,B_k)\right\rangle \geq$$
$$\geq \left\{\text{using assumptions 4 and 5, (20)}\right\} \geq \frac{\eta(2-\eta)\left\|\nabla_{x_k}\hat{g}_2(x_k,B_k)\right\|^2}{8\left(M_{\hat{G}}^2+P_{\hat{g}_1}L_k\right)} \geq \left\{L_k\leq \gamma L_{\hat{F}}\right\}\geq$$
$$\geq \frac{\eta(2-\eta)\left\|\nabla_{x_k}\hat{g}_2(x_k,B_k)\right\|^2}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}.$$

We sum the inequalities above for the first k iterations and average it using $\mathbb{E}[\cdot]$ operator:

$$\mathbb{E}\left[\sum_{i=0}^{k-1}\left(\hat{g}_2(x_i,B_i)-\hat{g}_2(x_{i+1},B_i)\right)\right]=\mathbb{E}\left[\sum_{i=0}^{k-1}\left(\hat{f}_2(x_i)-\hat{g}_2(x_{i+1},B_i)\right)\right]\geq$$
$$\geq \mathbb{E}\left[\sum_{i=0}^{k-1}\frac{\eta(2-\eta)\left\|\nabla_{x_i}\hat{g}_2(x_i,B_i)\right\|^2}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}\right]=$$
$$=\frac{\eta(2-\eta)}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}\sum_{i=0}^{k-1}\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla_{x_i}\hat{g}_2(x_i,B_i)\right\|^2\right]\right]\geq \frac{\eta(2-\eta)}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}\sum_{i=0}^{k-1}\mathbb{E}\left[\left\|\nabla\hat{f}_2(x_i)\right\|^2\right]\geq$$
$$\geq \frac{k\eta(2-\eta)}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}\min_{i\in\overline{0,k-1}}\left(\mathbb{E}\left[\left\|\nabla\hat{f}_2(x_i)\right\|^2\right]\right)\geq \frac{k\eta(2-\eta)}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}\mathbb{E}\left[\min_{i\in\overline{0,k-1}}\left\|\nabla\hat{f}_2(x_i)\right\|^2\right].$$

Now we rewrite the obtained inequality:

$$\frac{k\eta(2-\eta)}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}\mathbb{E}\left[\min_{i\in\overline{0,k-1}}\left\|\nabla\hat{f}_2(x_i)\right\|^2\right]\leq$$
$$\leq \mathbb{E}\left[\hat{f}_2(x_0)+\sum_{i=1}^{k-1}\left(\hat{f}_2(x_i)-\hat{g}_2(x_i,B_{i-1})\right)-\hat{g}_2(x_k,B_{k-1})\right]\leq$$
$$\leq \mathbb{E}\left[\hat{f}_2(x_0)\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\hat{f}_2(x_i)-\hat{g}_2(x_i,B_{i-1})\right]=\left|\mathbb{E}\left[\hat{f}_2(x_0)\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\hat{f}_2(x_i)-\hat{g}_2(x_i,B_{i-1})\right]\right|\leq$$
$$\leq \mathbb{E}\left[\hat{f}_2(x_0)\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\left|\hat{f}_2(x_i)-\hat{g}_2(x_i,B_{i-1})\right|\right].$$

According to lemma 11 the expression above is bounded:

$$\frac{k\eta(2-\eta)}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}\mathbb{E}\left[\min_{i\in\overline{0,k-1}}\left\|\nabla\hat{f}_2(x_i)\right\|^2\right]\leq \mathbb{E}\left[\hat{f}_2(x_0)\right]+\sum_{i=1}^{k-1}\mathbb{E}\left[\left|\hat{f}_2(x_i)-\hat{g}_2(x_i,B_{i-1})\right|\right]\leq$$
$$\leq \mathbb{E}\left[\hat{f}_2(x_0)\right]+\sum_{i=1}^{k-1}\left(2l_{\hat{F}}\mathbb{E}\left[\|x_i-x_{i-1}\|\right]\mathbb{1}_{\{b<m\}}+\tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right)\leq$$
$$\leq \left\{\text{lemma 13, corollary 13.2, } L_k\geq L,\ \eta_k\leq 1\right\}\leq$$

$$\leq \mathbb{E}\left[\hat{f}_2(x_0)\right] + (k-1)\left(2l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right) = \mathbb{E}\left[\hat{g}_2(x_0,B_0)\right] +$$

$$+ (k-1)\left(2l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right) \leq$$

$$\leq \mathbb{E}\left[\hat{f}_2(x_0)\right] + k\left(2l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right).$$

So, dividing by $\frac{k\eta(2-\eta)}{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}$ we get the desired estimate (29):

$$\mathbb{E}\left[\min_{i\in 0,k-1}\left\|\nabla\hat{f}_2(x_i)\right\|^2\right] \leq \frac{8\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}{\eta(2-\eta)}\left(\frac{\mathbb{E}\left[\hat{f}_2(x_0)\right]}{k} + 2l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \right.$$

$$\left. + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right),\ k\in\mathbb{N}.$$

□

Corollary 4.1. *The proved estimate has irreducible term*

$$\frac{16l_{\hat{F}}\left(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}}\right)}{\eta(2-\eta)}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}},$$

so it is more convenient to use inequality (24) *to get an upper estimate for the batch size. And to achieve an* $\hat{\varepsilon}>0$ *level for minimal value of norm of gradient in expectation we consider the system of inequalities below:*

$$\begin{cases} \frac{8(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}})\mathbb{E}[\hat{g}_2(x_0,B_0)]}{k\eta(2-\eta)} \leq (1-r)\hat{\varepsilon}^2; \\ \frac{8(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}})}{\eta(2-\eta)}\left(2l_{\hat{F}}\sqrt{m(m-1)}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}}+\tilde{\sigma}\right)\sqrt{\frac{1}{b}-\frac{1}{m}} \leq r\hat{\varepsilon}^2. \end{cases} \tag{30}$$

Inequalities (30) *put the following restrictions for the number of iterations and for the batch size:*

$$\begin{cases} k = \left\lceil \frac{8(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}})\mathbb{E}[\hat{g}_2(x_0,B_0)]}{\hat{\varepsilon}^2(1-r)\eta(2-\eta)} \right\rceil,\ r\in(0,1); \\ b = \min\left\{m,\ \left\lceil \frac{\frac{64(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}})^2}{\eta^2(2-\eta)^2}\left(2l_{\hat{F}}\sqrt{m(m-1)}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}+\tilde{\sigma}\right)^2}{\hat{\varepsilon}^4r^2+\frac{64(M_{\hat{G}}^2+\gamma P_{\hat{g}_1}L_{\hat{F}})^2}{m\eta^2(2-\eta)^2}\left(2l_{\hat{F}}\sqrt{m(m-1)}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}+\tilde{\sigma}\right)^2} \right\rceil \right\}; \end{cases} \tag{31}$$

these estimates asymptotically look like

$$k = \mathrm{O}\left(\frac{1}{\hat{\varepsilon}^2}\right),\ b = \min\left\{m,\ \mathrm{O}\left(\frac{1}{\hat{\varepsilon}^4}\right)\right\}.$$

D.2 The proof of theorem 5

Theorem 5 proves linear global convergence rate to approximate solution of problem (5) point in mean.

Theorem 5. *Suppose assumptions 3, 4, 5, 6, 7 are satisfied. Consider Stochastic Gauss–Newton method 2 with* $\tau_k = \hat{g}_1(x_k, B_k)$, $\eta_k \in [\eta, 1]$, $\eta \in (0,1]$ *and some finite* $\tilde{\sigma} \geq \sigma$. *Then:*

$$\begin{cases} \mathbb{E}\left[\left\|\nabla \hat{f}_2(x_k)\right\|^2\right] \leq 4M_{\hat{G}}^2 \Delta_{k,b}; \\ \mathbb{E}\left[\hat{f}_2(x_k)\right] \leq \hat{f}_2^* + \Delta_{k,b}; \\ \Delta_{k,b} \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{f}_2(x_0)\right] \exp\left(-\dfrac{k\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} P_{\hat{g}_1} + \mu\right)}\right) + 4\left(l_{\hat{F}} \min\left\{\sqrt{\dfrac{2P_{\hat{g}_1}}{L}},\ \dfrac{M_{\hat{G}}}{L}\right\} \mathbb{1}_{\{b<m\}} + \right. \\ \left. + \tilde{\sigma}\sqrt{\dfrac{1}{b} - \dfrac{1}{m}}\right)\left(\dfrac{\gamma L_{\hat{F}} P_{\hat{g}_1} + \mu}{\eta(2-\eta)\mu}\right),\ k \in \mathbb{Z}_+,\ b \in \overline{1,\ \min\{m,n\}}. \end{cases} \tag{32}$$

Expectation operator $\mathbb{E}[\cdot]$ *averages over all randomness in optimization procedure.*

Proof. According to update rule for x_k (7), (21) we have the following (corollary 9.1):

$$\hat{g}_1(x_k, B_k) - \hat{g}_1(x_{k+1}, B_k) \geq$$
$$\geq \frac{\eta_k(2-\eta_k)}{2\hat{g}_1(x_k,B_k)} \left\langle \left(\hat{G}'(x_k,B_k)^* \hat{G}'(x_k,B_k) + \hat{g}_1(x_k,B_k) L_k I_n\right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k),\right.$$
$$\left.\hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k)\right\rangle \geq 0 \Rightarrow \hat{g}_2(x_k,B_k) - \hat{g}_2(x_{k+1},B_k) \geq \hat{g}_2(x_k,B_k) - \hat{g}_1(x_k,B_k)\hat{g}_1(x_{k+1},B_k) \geq$$
$$\geq \left\{\eta_k \geq \eta,\ L_k \leq \gamma L_{\hat{F}}\right\} \geq$$
$$\geq \frac{\eta(2-\eta)}{2} \left\langle \left(\hat{G}'(x_k,B_k)^* \hat{G}'(x_k,B_k) + \gamma \hat{g}_1(x_k,B_k) L_{\hat{F}} I_n\right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k),\right.$$
$$\left.\hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k)\right\rangle =$$
$$= \frac{\eta(2-\eta)}{2} \left\langle \hat{G}'(x_k,B_k)\left(\hat{G}'(x_k,B_k)^* \hat{G}'(x_k,B_k) + \gamma \hat{g}_1(x_k,B_k) L_{\hat{F}} I_n\right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k),\right.$$
$$\left.\hat{G}(x_k,B_k)\right\rangle \geq \{\text{lemma 5}\} \geq \frac{\eta(2-\eta)\left\|\hat{G}(x_k,B_k)\right\|^2 \mu}{2\left(\gamma L_{\hat{F}} \hat{g}_1(x_k,B_k) + \mu\right)} = \hat{g}_2(x_k,B_k) \frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} \hat{g}_1(x_k,B_k) + \mu\right)} \Rightarrow$$
$$\Rightarrow \hat{g}_2(x_{k+1},B_k) \leq \hat{g}_2(x_k,B_k)\left(1 - \frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} \hat{g}_1(x_k,B_k) + \mu\right)}\right).$$

We add $-\hat{f}_2^* \leq 0$ to the inequality above:

$$\hat{g}_2(x_{k+1},B_k) - \hat{f}_2^* \leq \left(\hat{g}_2(x_k,B_k) - \hat{f}_2^*\right) - \hat{g}_2(x_k,B_k) \frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} \hat{g}_1(x_k,B_k) + \mu\right)} \leq$$
$$\leq \left(\hat{g}_2(x_k,B_k) - \hat{f}_2^*\right)\left(1 - \frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} \hat{g}_1(x_k,B_k) + \mu\right)}\right) = \left(\hat{g}_2(x_k,B_k) - \hat{f}_2(x_k) + \hat{f}_2(x_k) - \right.$$
$$\left. - \hat{g}_2(x_k,B_{k-1}) + \hat{g}_2(x_k,B_{k-1}) - \hat{f}_2^*\right)\left(1 - \frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} \hat{g}_1(x_k,B_k) + \mu\right)}\right) \leq$$
$$\leq \left(\left|\hat{g}_2(x_k,B_k) - \hat{f}_2(x_k) + \hat{f}_2(x_k) - \hat{g}_2(x_k,B_{k-1})\right| + \right.$$
$$\left. + \hat{g}_2(x_k,B_{k-1}) - \hat{f}_2^*\right)\left(1 - \frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}} \hat{g}_1(x_k,B_k) + \mu\right)}\right) \leq$$
$$\leq \left(\left|\hat{g}_2(x_k,B_k) - \hat{f}_2(x_k)\right| + \left|\hat{f}_2(x_k) - \hat{g}_2(x_k,B_{k-1})\right| + \right.$$

$$+\left(\hat{g}_2(x_k,B_{k-1})-\hat{f}_2^*\right)\Big)\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}\hat{g}_1(x_k,B_k)+\mu\right)}\right). \tag{33}$$

Now we average the inequality above using operator $\mathbb{E}[\cdot]$:

$$\begin{aligned}
&\mathbb{E}\left[\hat{g}_2(x_{k+1},B_k)-\hat{f}_2^*\right]\leq\mathbb{E}\left[\left(\left|\hat{g}_2(x_k,B_k)-\hat{f}_2(x_k)\right|+\left|\hat{f}_2(x_k)-\hat{g}_2(x_k,B_{k-1})\right|+\right.\right.\\
&\left.\left.+\left(\hat{g}_2(x_k,B_{k-1})-\hat{f}_2^*\right)\right)\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}\hat{g}_1(x_k,B_k)+\mu\right)}\right)\right]\leq\{\text{according to assumption 5}\}\leq\\
&\leq\left(\underbrace{\mathbb{E}\left[\left|\hat{g}_2(x_k,B_k)-\hat{f}_2(x_k)\right|\right]}_{\text{is bounded with batch variance}}+\underbrace{\mathbb{E}\left[\left|\hat{f}_2(x_k)-\hat{g}_2(x_k,B_{k-1})\right|\right]}_{\text{is bounded according to lemma 11}}+\right.\\
&\left.+\mathbb{E}\left[\hat{g}_2(x_k,B_{k-1})-\hat{f}_2^*\right]\right)\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right).
\end{aligned}$$

The expression above is bounded according to lemmas 10 (corollary 10.2) and 11:

$$\begin{aligned}
&\mathbb{E}\left[\hat{g}_2(x_{k+1},B_k)-\hat{f}_2^*\right]\leq\left(\tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}+2l_{\hat{F}}\mathbb{E}\left[\|x_k-x_{k-1}\|\right]\mathbb{1}_{\{b<m\}}+\tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}+\right.\\
&\left.+\mathbb{E}\left[\hat{g}_2(x_k,B_{k-1})-\hat{f}_2^*\right]\right)\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)\leq\{\text{corollary 13.2, }\eta_k\leq 1\}\leq\\
&\leq\left(2\left(l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}}+\tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right)+\right.\\
&\left.+\mathbb{E}\left[\hat{g}_2(x_k,B_{k-1})-\hat{f}_2^*\right]\right)\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right).
\end{aligned} \tag{34}$$

Formula (34) represents a recurrent dependency over iterations $k\in\mathbb{N}$:

$$\begin{cases}
a_k & \stackrel{\text{def}}{=}\mathbb{E}\left[\hat{g}_2(x_k,B_{k-1})-\hat{f}_2^*\right];\\
c_k & \stackrel{\text{def}}{=}c=2\left(l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}}+\tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right);\\
q & \stackrel{\text{def}}{=}\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)\leq\exp\left(-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)\in(0,1).
\end{cases} \tag{35}$$

So, for sequence $\{a_k\}_{k\in\mathbb{N}}$ defined in (35) we have

$$\begin{cases}
a_0 & \stackrel{\text{def}}{=}\mathbb{E}\left[\hat{g}_2(x_0,B_0)\right];\\
a_1 & \leq a_0q\leq a_0q+cq;\\
a_{k+1} & \leq(a_k+c)\,q,\ k\in\mathbb{N}.
\end{cases}$$

The bound for a_1 is straightforwardly deduced from (33):

$$\begin{aligned}
&\hat{g}_2(x_{k+1},B_k)-\hat{f}_2^*\leq\hat{g}_2(x_{k+1},B_k)\leq\hat{g}_2(x_k,B_k)\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}\hat{g}_1(x_k,B_k)+\mu\right)}\right)\leq\\
&\leq\{\text{assumption 5}\}\leq\hat{g}_2(x_k,B_k)\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)\Rightarrow\{k=0\}\Rightarrow\mathbb{E}\left[\hat{g}_2(x_1,B_0)-\hat{f}_2^*\right]\leq\\
&\leq\mathbb{E}\left[\hat{g}_2(x_0,B_0)\right]\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)\leq\mathbb{E}\left[\hat{g}_2(x_0,B_0)\right]\exp\left(-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right).
\end{aligned}$$

Now, we can bound a_k form above using the following sum:

$$a_k \leq \underbrace{(a_{k-1}+c)q \leq ((a_{k-2}+c)q+c)q}_{\text{partial sum for geometric series}} \leq \cdots \leq a_0 q^k + c\sum_{i=1}^{k-1} q^i = a_0 q^k + cq\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}},$$
$$k \in \mathbb{Z}_+. \tag{36}$$

We link a_k bound with $\hat{f}_2(x_k)$ upper bound:

$$\begin{aligned}
&\hat{f}_2(x_k) - \hat{f}_2^* = \hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1}) + \hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^* \leq \left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right| + \\
&+ \hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^* \Rightarrow \mathbb{E}\left[\hat{f}_2(x_k) - \hat{f}_2^*\right] \leq \mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] + \\
&+ \mathbb{E}\left[\hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^*\right] \leq \{\text{according to lemma 11 and corollary 13.2, } \eta_k \leq 1\} \leq \\
&\leq 2l_{\hat{F}} \min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}}, \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}} + \mathbb{E}\left[\hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^*\right] = \{(35)\} = \\
&= 2l_{\hat{F}} \min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}}, \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}} + a_k \leq \{(36)\} \leq \\
&\leq 2l_{\hat{F}} \min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}}, \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}} + a_0 q^k + cq\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}} \leq \\
&\leq 2\left(l_{\hat{F}} \min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}}, \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right) + a_0 q^k + cq\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}} = \\
&= a_0 q^k + c\left(1 + q\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}}\right),\ k \in \mathbb{Z}_+.
\end{aligned}$$

And we express $\mathbb{E}\left[\hat{f}_2(x_k)\right]$ estimate from the inequality above:

$$\mathbb{E}\left[\hat{f}_2(x_k)\right] \leq \hat{f}_2^* + a_0 q^k + c\left(1 + q\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}}\right),\ k \in \mathbb{Z}_+. \tag{37}$$

We use WGC to bound the mean squared norm of the gradient:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla \hat{f}_2(x_k)\right\|^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[\|\nabla_{x_k}\hat{g}_2(x_k, B_k)\|^2\right]\right] = \mathbb{E}\left[\left\|2\hat{G}'(x_k, B_k)^*\hat{G}(x_k, B_k)\right\|^2\right] \leq \\
&\leq 4\mathbb{E}\left[\left\|\hat{G}'(x_k, B_k)^*\right\|^2\left\|\hat{G}(x_k, B_k)\right\|^2\right] \leq \{\text{assumption 4}\} \leq 4M_{\hat{G}}^2\mathbb{E}\left[\hat{g}_2(x_k, B_k)\right].
\end{aligned}$$

Consider the next expression:

$$\begin{aligned}
&\hat{g}_2(x_k, B_k) - \hat{f}_2^* = \hat{g}_2(x_k, B_k) - \hat{f}_2(x_k) + \hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1}) + \hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^* \leq \\
&\leq \left|\hat{g}_2(x_k, B_k) - \hat{f}_2(x_k)\right| + \left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right| + \left(\hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^*\right) \Rightarrow \\
&\Rightarrow \mathbb{E}\left[\hat{g}_2(x_k, B_k) - \hat{f}_2^*\right] \leq \mathbb{E}\left[\left|\hat{g}_2(x_k, B_k) - \hat{f}_2(x_k)\right|\right] + \mathbb{E}\left[\left|\hat{f}_2(x_k) - \hat{g}_2(x_k, B_{k-1})\right|\right] + \\
&+ \mathbb{E}\left[\hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^*\right] \leq \{\text{corollary 10.2 and lemma 11}\} \leq \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}} + \\
&+ 2l_{\hat{F}}\mathbb{E}\left[\|x_k - x_{k-1}\|\right]\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}} + \mathbb{E}\left[\hat{g}_2(x_k, B_{k-1}) - \hat{f}_2^*\right] \leq \\
&\leq \{\text{corollary 13.2, } \eta_k \leq 1\} \leq 2\left(l_{\hat{F}} \min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}}, \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}} + \tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right) +
\end{aligned}$$

$$+\mathbb{E}\left[\hat{g}_2(x_k,B_{k-1})-\hat{f}_2^*\right]=c+a_k\leq\{(36)\}\leq a_0q^k+c\left(1+q\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}}\right),\ k\in\mathbb{Z}_+. \tag{38}$$

Using the arbitrariness of $\hat{f}_2^*\geq 0$ in (33) and (38) we can set $\hat{f}_2^*=0$ in (33) and (38) to estimate $\mathbb{E}\left[\left\|\nabla\hat{f}_2(x_k)\right\|^2\right]$:

$$\mathbb{E}\left[\left\|\nabla\hat{f}_2(x_k)\right\|^2\right]\leq 4M_{\hat{G}}^2\left(a_0q^k+c\left(1+q\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}}\right)\right),\ k\in\mathbb{Z}_+. \tag{39}$$

Simplifying (37) and (39) we obtain the desired result:

$$\begin{aligned}
&a_0q^k+c\left(1+q\left(\frac{1-q^{k-1}}{1-q}\right)\mathbb{1}_{\{k>0\}}\right)\leq a_0q^k+c\left(1+\frac{q}{1-q}\right)=\\
&=\mathbb{E}\left[\hat{g}_2(x_0,B_0)\right]\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)^k+2\left(l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}}+\right.\\
&\left.+\tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right)\left(1+\left(1-\frac{\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)\frac{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}{\eta(2-\eta)\mu}\right)\leq\\
&\leq\mathbb{E}\left[\hat{f}_2(x_0)\right]\exp\left(-\frac{k\eta(2-\eta)\mu}{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}\right)+4\left(l_{\hat{F}}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}\mathbb{1}_{\{b<m\}}+\right.\\
&\left.+\tilde{\sigma}\sqrt{\frac{1}{b}-\frac{1}{m}}\right)\left(\frac{\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu}{\eta(2-\eta)\mu}\right)=\Delta_{k,b},\ k\in\mathbb{Z}_+,\ b\in\overline{1,\ \min\{m,\ n\}}.
\end{aligned}$$

So, the estimates (37) and (39) using $\Delta_{k,b}$ represent (32):

$$\begin{cases}\mathbb{E}\left[\hat{f}_2(x_k)\right]\leq\hat{f}_2^*+\Delta_{k,b};\\ \mathbb{E}\left[\left\|\nabla\hat{f}_2(x_k)\right\|^2\right]\leq 4M_{\hat{G}}^2\Delta_{k,b}.\end{cases}$$

□

Corollary 5.1. *Analogously to corollary 4.1 we establish the following conditions on the batch size and on the number of iterations using approach from* (30) *and* (31)*:*

$$\begin{cases}k=\left\lceil\frac{2\left(\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu\right)}{\eta(2-\eta)\mu}\ln\left(\frac{4M_{\hat{G}}^2\mathbb{E}[\hat{g}_2(x_0,B_0)]}{\hat{\varepsilon}^2(1-r)}\right)\right\rceil,\ r\in(0,1);\\ b=\min\left\{m,\ n,\ \left\lceil\frac{256M_{\hat{G}}^4\left(l_{\hat{F}}\sqrt{m(m-1)}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}+\tilde{\sigma}\right)^2\left(\frac{\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu}{\eta(2-\eta)\mu}\right)^2}{\hat{\varepsilon}^4r^2+\frac{256M_{\hat{G}}^4}{m}\left(l_{\hat{F}}\sqrt{m(m-1)}\min\left\{\sqrt{\frac{2P_{\hat{g}_1}}{L}},\ \frac{M_{\hat{G}}}{L}\right\}+\tilde{\sigma}\right)^2\left(\frac{\gamma L_{\hat{F}}P_{\hat{g}_1}+\mu}{\eta(2-\eta)\mu}\right)^2}\right\rceil\right\};\end{cases}$$

or asymptotically:

$$k=\mathrm{O}\left(\ln\left(\frac{1}{\hat{\varepsilon}}\right)\right),\ b=\min\left\{m,\ n,\ \mathrm{O}\left(\frac{1}{\hat{\varepsilon}^4}\right)\right\}.$$

In such conditions we can be sure that

$$\mathbb{E}\left[\left\|\nabla\hat{f}_2(x_k)\right\|^2\right]\leq\hat{\varepsilon}^2.$$

However, the batch size limitation $b\leq\min\{m,\ n\}$ restricts from achieving arbitrary low $\hat{\varepsilon}\geq 0$ value of the gradient norm in mean with linear convergence speed in the worst case.

D.3 The proof of theorem 6

The doubly stochastic step is described in (8). This step is originally based on the two batch estimate of local model $\hat{\psi}_{x_k,L_k,\tilde{\tau}_k}(y,\tilde{B}_k)$:

$$\hat{\psi}_{x_k,L_k,\tilde{\tau}_k}(y,\tilde{B}_k) = \frac{\tilde{\tau}_k}{2} + \frac{L_k}{2}\|y-x_k\|^2 + \frac{1}{2\tilde{\tau}_k}\left\|\hat{G}(x_k,\tilde{B}_k) + \hat{G}'(x_k,\tilde{B}_k)(y-x_k)\right\|^2 =$$
$$= \left(\frac{\tilde{\tau}_k}{2} + \frac{\hat{g}_2(x_k,\tilde{B}_k)}{2\tilde{\tau}_k}\right) + \left\langle \frac{\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}(x_k,\tilde{B}_k)}{\tilde{\tau}_k},\ y-x_k\right\rangle +$$
$$+ \frac{1}{2}\left\langle \left(\frac{\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k)}{\tilde{\tau}_k} + L_k I_n\right)(y-x_k),\ y-x_k\right\rangle \Rightarrow$$

$$\Rightarrow \hat{\psi}_{x_k,L_k,\tilde{\tau}_k}(y,\tilde{B}_k) \approx \left(\frac{\tilde{\tau}_k}{2} + \frac{\hat{g}_2(x_k,\tilde{B}_k)}{2\tilde{\tau}_k}\right) + \left\langle \frac{\hat{G}'(x_k,B_k)^*\hat{G}(x_k,B_k)}{\tilde{\tau}_k},\ y-x_k\right\rangle +$$
$$+ \frac{1}{2}\left\langle \left(\frac{\hat{G}'(x_k,\tilde{B}_k)^*\hat{G}'(x_k,\tilde{B}_k)}{\tilde{\tau}_k} + L_k I_n\right)(y-x_k),\ y-x_k\right\rangle.$$

So, the rule from (8) can be viewed as some scaled step of Newton method to optimize the two batch estimate of $\hat{\psi}_{x_k,L_k,\tilde{\tau}_k}(y,\tilde{B}_k)$. Note that estimate of $\nabla_y\hat{\psi}_{x_k,L_k,\tilde{\tau}_k}(y,\tilde{B}_k)$ evaluated at $y=x_k$ is unbiased w.r.t. B_k. The stochastic Gauss–Newton framework with doubly stochastic step possesses settings (10). And formal description of this framework is presented in algorithm 3. Theorem 6 elaborates linear convergence rates in mean to solution of problem (5) in the worst case scenario. The whole procedure 3 is justified as the optimizer for local model $\varphi_{x_k,l_k}(y),\ y := x_{k+1}$ in mean:

$$\hat{f}_2(y) \le \underbrace{\varphi_{x_k,l_k}(y)}_{\text{by corollary 15.1}} = \hat{f}_2(x_k) + \left\langle \nabla\hat{f}_2(x_k),\ y-x_k\right\rangle + \frac{l_k}{2}\|y-x_k\|^2,$$
$$l_k \ge l_{\hat{f}_2} = \underbrace{2\left(M_{\hat{F}}^2 + L_{\hat{F}}P_{\hat{f}_1}\right)}_{\text{by corollary 14.1}},\ (x,y)\in E_1^2.$$

Theorem 6. *Suppose that assumptions 3, 4, 5, 7 are satisfied. Consider Stochastic Gauss–Newton method 3 with $\tilde{\tau}_k \ge \tilde{\tau} > 0$, $L_k \ge L > 0$. Then, for sequence*

$$\eta_k = \frac{\mu\left(\tilde{\tau}_k L_k\right)^2}{\left(M_{\hat{G}}^2 + \tilde{\tau}_k L_k\right)\left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)M_{\hat{G}}^2},\ k\in\mathbb{Z}_+$$

the next estimate holds

$$\mathbb{E}\left[\hat{f}_2(x_k)\right] \le \mathbb{E}\left[\hat{f}_2(x_0)\right]\exp\left(-\frac{k}{\left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)M_{\hat{G}}^2}\left(\frac{\mu\tilde{\tau}L}{M_{\hat{G}}^2 + \tilde{\tau}L}\right)^2\right),\ k\in\mathbb{Z}_+.$$

In case of $\eta_k = 1,\ k\in\mathbb{Z}_+$ convergence estimate is no better than

$$\begin{cases} \mathbb{E}\left[\hat{f}_2(x_k)\right] \le \mathbb{E}\left[\hat{f}_2(x_0)\right]\exp\left(-\frac{k\mu^2}{M_{\hat{G}}^2}\left(\frac{2}{\mu + \left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)c} - \frac{1}{\left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)c^2}\right)\right); \\ c \overset{\text{def}}{=} \frac{1}{3}\left(1 + 7\sqrt[3]{\frac{2}{47+3\sqrt{93}}} + \sqrt[3]{\frac{47+3\sqrt{93}}{2}}\right),\ k\in\mathbb{Z}_+. \end{cases} \tag{40}$$

Expectation operator $\mathbb{E}[\cdot]$ averages over all randomness in optimization procedure.

Proof. Function $\hat{f}_2$ has the Lipschitz gradient with the Lipschitz constant estimate

$$l_{\hat{f}_2} = 2\left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)$$

from corollary 14.1. Consider local model for function $\hat{f}_2$ at k–th iteration evaluated at $x_{k+1},\ k \in \mathbb{Z}_+$ (corollary 15.1):

$$\hat{f}_2(x_{k+1}) \leq \hat{f}_2(x_k) + \left\langle \nabla \hat{f}_2(x_k),\ x_{k+1} - x_k \right\rangle + \frac{l_{\hat{f}_2}}{2}\|x_{k+1} - x_k\|^2.$$

Update rule for x_k is defined as follows:

$$x_{k+1} = x_k - \eta_k \underbrace{\left(\hat{G}'(x_k, \tilde{B}_k)^* \hat{G}'(x_k, \tilde{B}_k) + \tilde{\tau}_k L_k I_n\right)^{-1}}_{\stackrel{\text{def}}{=} 2H_k} \underbrace{\hat{G}'(x_k, B_k)^* \hat{G}(x_k, B_k)}_{\stackrel{\text{def}}{=} \frac{1}{2}\nabla_{x_k}\hat{g}_2(x_k, B_k)} =$$
$$= x_k - \eta_k H_k \nabla_{x_k}\hat{g}_2(x_k, B_k),\ \eta_k > 0.$$

$B_k, \tilde{B}_k \subseteq \mathscr{B}$ — independently sampled batches at k–th iteration. Substitute this update rule into the local model:

$$\hat{f}_2(x_{k+1}) \leq \hat{f}_2(x_k) - \eta_k \left\langle \nabla \hat{f}_2(x_k),\ H_k \nabla_{x_k}\hat{g}_2(x_k, B_k) \right\rangle + \frac{\eta_k^2 l_{\hat{f}_2}}{2}\|H_k \nabla_{x_k}\hat{g}_2(x_k, B_k)\|^2 =$$
$$= \hat{f}_2(x_k) - \eta_k \left\langle \nabla \hat{f}_2(x_k), H_k \nabla_{x_k}\hat{g}_2(x_k, B_k) \right\rangle + \frac{\eta_k^2 l_{\hat{f}_2}}{2}\left\langle H_k^2 \nabla_{x_k}\hat{g}_2(x_k, B_k), \nabla_{x_k}\hat{g}_2(x_k, B_k) \right\rangle.$$

For matrix H_k the next relations hold (lemma 6):

$$\frac{1}{\left(2\left(M_{\hat{G}}^2 + \tilde{\tau}_k L_k\right)\right)^t} I_n \preceq H_k^t \preceq \frac{1}{\left(2\tilde{\tau}_k L_k\right)^t} I_n,\ k \in \mathbb{Z}_+,\ t \geq 0.$$

Now we average local model using the expectation operator:

$$\mathbb{E}\left[\hat{f}_2(x_{k+1})\right] \leq \mathbb{E}\left[\hat{f}_2(x_k)\right] - \eta_k \mathbb{E}\left[\left\langle \nabla \hat{f}_2(x_k), H_k \nabla \hat{f}_2(x_k) \right\rangle\right] +$$
$$+ \frac{\eta_k^2 l_{\hat{f}_2}}{2}\mathbb{E}\left[\left\langle H_k^2 \nabla_{x_k}\hat{g}_2(x_k, B_k), \nabla_{x_k}\hat{g}_2(x_k, B_k) \right\rangle\right] \leq$$
$$\leq \mathbb{E}\left[\hat{f}_2(x_k)\right] - \frac{\eta_k \mathbb{E}\left[\left\|\nabla \hat{f}_2(x_k)\right\|^2\right]}{2\left(M_{\hat{G}}^2 + \tilde{\tau}_k L_k\right)} + \frac{\eta_k^2 l_{\hat{f}_2}}{8\left(\tilde{\tau}_k L_k\right)^2}\mathbb{E}\left[\|\nabla_{x_k}\hat{g}_2(x_k, B_k)\|^2\right].$$

We use WGC and PL condition from (28) for the inequality above:

$$\mathbb{E}\left[\hat{f}_2(x_{k+1})\right] \leq \mathbb{E}\left[\hat{f}_2(x_k)\right] - \frac{\eta_k \mathbb{E}\left[\left\|\nabla \hat{f}_2(x_k)\right\|^2\right]}{2\left(M_{\hat{G}}^2 + \tilde{\tau}_k L_k\right)} + \frac{\eta_k^2 l_{\hat{f}_2}}{8\left(\tilde{\tau}_k L_k\right)^2}\mathbb{E}\left[\|\nabla_{x_k}\hat{g}_2(x_k, B_k)\|^2\right] \leq$$
$$\leq \mathbb{E}\left[\hat{f}_2(x_k)\right] - \frac{2\eta_k \mu \mathbb{E}\left[\hat{f}_2(x_k)\right]}{M_{\hat{G}}^2 + \tilde{\tau}_k L_k} + \frac{\eta_k^2 M_{\hat{G}}^2 l_{\hat{f}_2}}{2\left(\tilde{\tau}_k L_k\right)^2}\mathbb{E}\left[\hat{f}_2(x_k)\right] =$$
$$= \mathbb{E}\left[\hat{f}_2(x_k)\right]\left(1 - \frac{2\eta_k \mu}{M_{\hat{G}}^2 + \tilde{\tau}_k L_k} + \frac{l_{\hat{f}_2}}{2}\left(\frac{\eta_k M_{\hat{G}}}{\tilde{\tau}_k L_k}\right)^2\right) =$$
$$= \mathbb{E}\left[\hat{f}_2(x_k)\right]\left(1 - \frac{2\eta_k \mu}{M_{\hat{G}}^2 + \tilde{\tau}_k L_k} + \left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)\left(\frac{\eta_k M_{\hat{G}}}{\tilde{\tau}_k L_k}\right)^2\right).$$

We compute the optimal step scale for each iteration:

$$\eta_k^2\left(\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)\left(\frac{M_{\hat{G}}}{\tilde{\tau}_k L_k}\right)^2\right)-\eta_k\left(\frac{2\mu}{M_{\hat{G}}^2+\tilde{\tau}_k L_k}\right)\to\min_{\eta_k>0}\Rightarrow$$
$$\Rightarrow\eta_k=\frac{\mu\left(\tilde{\tau}_k L_k\right)^2}{\left(M_{\hat{G}}^2+\tilde{\tau}_k L_k\right)\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)M_{\hat{G}}^2},\ \tilde{\tau}_k>0,\ L_k>0,\ k\in\mathbb{Z}_+.$$

Such step scale leads to linear convergence speed with an arbitrary batch size:

$$\mathbb{E}\left[\hat{f}_2(x_{k+1})\right]\le\mathbb{E}\left[\hat{f}_2(x_k)\right]\underbrace{\left(1-\left(\frac{\mu\tilde{\tau}_k L_k}{M_{\hat{G}}^2+\tilde{\tau}_k L_k}\right)^2\frac{1}{\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)M_{\hat{G}}^2}\right)}_{\in(0,\,1)\text{ because }0<\mu\le\min\{M_{\hat{F}}^2,\,M_{\hat{G}}^2\}\text{ and }\tilde{\tau}_k L_k>0}\le$$
$$\le\mathbb{E}\left[\hat{f}_2(x_0)\right]\exp\left(\frac{-(k+1)}{\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)M_{\hat{G}}^2}\left(\frac{\mu\tilde{\tau}L}{M_{\hat{G}}^2+\tilde{\tau}L}\right)^2\right),\ k\in\mathbb{Z}_+.$$

Now look closer at convergence estimate. Define function

$$\alpha(t)\stackrel{\text{def}}{=}1-\left(\frac{\mu t}{M_{\hat{G}}^2+t}\right)^2\frac{1}{\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)M_{\hat{G}}^2},$$

and find its minimal value and points of the minimal value $t=\tilde{\tau}_k L_k$ to estimate the best decrease $\mathbb{E}\left[\hat{f}_2(x_{k+1})\right]$:

$$\mathbb{E}\left[\hat{f}_2(x_{k+1})\right]\le\alpha(\tilde{\tau}_k L_k)\mathbb{E}\left[\hat{f}_2(x_k)\right].$$

The search of minimum of $\alpha(t)$ is equivalent to the search of maximum of the function below:

$$\beta(t)\stackrel{\text{def}}{=}\frac{\mu t}{M_{\hat{G}}^2+t}.$$

Function $\beta(t)$ has non negative first derivative and non positive second derivative on $\mathbb{R}_+$:

$$\beta'(t)=\frac{\mu}{M_{\hat{G}}^2+t}\left(1-\frac{t}{M_{\hat{G}}^2+t}\right)\ge 0;$$
$$\beta''(t)=\frac{2\mu}{\left(M_{\hat{G}}^2+t\right)^2}\left(\frac{t}{M_{\hat{G}}^2+t}-1\right)\le 0.$$

It means, the greater t we have, the less $\alpha(t)$ we get:

$$1=\alpha(0)\ge\alpha(t)\ge\lim_{t\to+\infty}\alpha(t)=1-\frac{\mu^2}{\left(L_{\hat{F}}P_{\hat{f}_1}+M_{\hat{F}}^2\right)M_{\hat{G}}^2}>0,\ t\in\mathbb{R}_+,$$

because $\mu \leq \min\{M_{\hat{G}}^2,\, M_{\hat{F}}^2\}$ (by assumption 4). Consider change of the update rule relatively $t = \tilde{\tau}_k L_k$:

$$\eta_k H_k \nabla_{x_k} \hat{g}_2(x_k, B_k) = \frac{\mu\,(\tilde{\tau}_k L_k)^2 \left(\hat{G}'(x_k,\tilde{B}_k)^* \hat{G}'(x_k,\tilde{B}_k) + \tilde{\tau}_k L_k I_n\right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k)}{\left(M_{\hat{G}}^2 + \tilde{\tau}_k L_k\right)\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2} \Rightarrow$$

$$\Rightarrow \lim_{t\to+\infty} \frac{\mu t}{M_{\hat{G}}^2 + t} \frac{\left(\frac{\hat{G}'(x_k,\tilde{B}_k)^* \hat{G}'(x_k,\tilde{B}_k)}{t} + I_n\right)^{-1} \hat{G}'(x_k,B_k)^* \hat{G}(x_k,B_k)}{\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2} =$$

$$= \left(\frac{\mu}{2\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2}\right) \nabla_{x_k} \hat{g}_2(x_k, B_k).$$

So, the faster estimate we get, the closer stochastic Gauss–Newton method update to stochastic gradient method update.

If we set the value $\eta_k = 1$, we can find the unique optimal $t = \tilde{\tau}_k L_k,\ k \in \mathbb{Z}_+$ from the local decrease estimate

$$\mathbb{E}\left[\hat{f}_2(x_{k+1})\right] \leq \mathbb{E}\left[\hat{f}_2(x_k)\right] \left(1 - \frac{2\mu}{M_{\hat{G}}^2 + t} + \left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) \left(\frac{M_{\hat{G}}}{t}\right)^2\right).$$

To prove that we directly find the optimal convergence rate:

$$\underbrace{\frac{\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2}{t^2} - \frac{2\mu}{M_{\hat{G}}^2 + t}}_{\stackrel{\text{def}}{=}\zeta(t)} \to \min_{t>0}.$$

We express optimality conditions of the second order for defined function $\zeta(t)$:

$$\begin{cases} \zeta'(t) = \frac{-2\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2}{t^3} + \frac{2\mu}{\left(M_{\hat{G}}^2 + t\right)^2} = 0; \\ \zeta''(t) = \frac{6\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2}{t^4} - \frac{4\mu}{\left(M_{\hat{G}}^2 + t\right)^3} > 0. \end{cases}$$

Condition $\zeta'(t) = 0$ leads to cubic equation

$$\mu t^3 - \left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2 t^2 - 2\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^4 t - \left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^6 = 0,$$

with unique real root obtained from the general formula for roots of cubic equation:

$$t^* = \frac{\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) M_{\hat{G}}^2}{3\mu} \left(1 + 7\sqrt[3]{\frac{2}{47 + 3\sqrt{93}}} + \sqrt[3]{\frac{47 + 3\sqrt{93}}{2}}\right),$$

for this value we have $\zeta''(t^*) > 0$. Moreover, for t^* we have linear convergence with estimate (40), and linear convergence rate lies in $(0, 1)$:

$$\begin{cases} 0 < 1 - \frac{\mu^2}{M_{\hat{G}}^2} \left(\frac{2}{\mu + \left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) c} - \frac{1}{\left(L_{\hat{F}} P_{\hat{f}_1} + M_{\hat{F}}^2\right) c^2}\right) < 1; \\ c = \frac{1}{3}\left(1 + 7\sqrt[3]{\frac{2}{47+3\sqrt{93}}} + \sqrt[3]{\frac{47+3\sqrt{93}}{2}}\right) \in (2,\ 3). \end{cases}$$

Inequalities $\mu \leq \min\left\{M_{\hat{G}}^2,\, M_{\hat{F}}^2\right\}$ and $3 > c > 2$ force for convergence rate to be within $(0,\, 1)$. If we compare this rate with the convergence rate for non fixed, adaptive η_k, the second one turns out to be less in the limit case, when taking $t \to +\infty$, and, thus, possesses faster convergence in the worst case scenario. □

Corollary 6.1. *Unlike corollaries 4.1 and 5.1, we can state the convergence condition relatively function $\hat{f}_2$ value:*

$$\mathbb{E}\left[\hat{f}_2(x_k)\right] \leq \hat{\varepsilon}^2.$$

So, for adaptive η_k we have the following minimal number of iterations:

$$k = \left\lceil \frac{M_{\hat{G}}^2\left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)}{\mu^2}\left(\frac{M_{\hat{G}}^2}{\tilde{\tau}L} + 1\right)^2 \ln\left(\frac{\mathbb{E}\left[\hat{g}_2(x_0, B_0)\right]}{\hat{\varepsilon}^2}\right)\right\rceil = \mathrm{O}\left(\ln\left(\frac{1}{\hat{\varepsilon}}\right)\right).$$

The same asymptotics we have for the case of $\eta_k = 1$ with optimal value of $\tilde{\tau}_k L_k$:

$$k = \left\lceil \frac{M_{\hat{G}}^2}{\mu^2}\left(\frac{2}{\mu + \left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)c} - \frac{1}{\left(L_{\hat{F}}P_{\hat{f}_1} + M_{\hat{F}}^2\right)c^2}\right)^{-1} \ln\left(\frac{\mathbb{E}\left[\hat{g}_2(x_0, B_0)\right]}{\hat{\varepsilon}^2}\right)\right\rceil =$$
$$= \mathrm{O}\left(\ln\left(\frac{1}{\hat{\varepsilon}}\right)\right).$$

And because of independence from the batch sizes $\tilde{b}$ and b we can assume a constant asymptotics for batches, achieving the lowest polylogarithmic complexity cost of the number of oracle calls within our work for $m \leq n$ in the worst case:

- $bk = \min\left\{\mathrm{O}\left(\frac{m}{\hat{\varepsilon}^2}\right),\, \mathrm{O}\left(\frac{1}{\hat{\varepsilon}^6}\right)\right\}$ *from corollary 4.1;*
- $bk = \min\left\{\mathrm{O}\left(m\ln\left(\frac{1}{\hat{\varepsilon}}\right)\right),\, \mathrm{O}\left(\frac{1}{\hat{\varepsilon}^4}\ln\left(\frac{1}{\hat{\varepsilon}}\right)\right)\right\}$ *from corollary 5.1;*
- $\left(\tilde{b} + b\right)k = \mathrm{O}\left(\ln\left(\frac{1}{\hat{\varepsilon}}\right)\right)$ *from theorem 6.*

E Details of the experiments for Gauss–Newton method

This section provides the details of our experiments, including hyperparameters descriptions, data generating procedures and experiment configurations. We describe experiments in both deterministic and stochastic settings.

We run experiments on three benchmark tasks based on unconstrained minimization task. The original task means the following minimization of doubly smooth function f:

$$\min_{x \in E_1}\{f(x)\}.$$

For our case we consider an aggregated form of this task, solving the system of nonlinear equations to obtain a stationary point, not necessarily minimum point:

$$\nabla f(x) \equiv F(x) = \mathbf{0}_m,\ x \in E_1.$$

And for such system we have the optimizable merit $\|F(x)\|$. Clearly, the obtained system is *square* in terms of dimensions $m = n$. Using $\|F(x)\|$ we test three distinct functions $f(x)$, $x \overset{\text{def}}{=} (x^1, \ldots, x^n)^*$, $x \in E_1$:

- Nesterov–Skokov function [9]: $f_{NS}(x) := \frac{1}{4}(x^1 - 1)^2 + \sum\limits_{i=1}^{n-1} \left(x^{i+1} - 2\left(x^i\right)^2 + 1\right)^2$;
- Hat function: $f_H(x) := \left(\|x\|^2 - 1\right)^2$;
- PL function: $f_{PL}(x) = \|x\|^2 + 3\sum\limits_{i=1}^{n} \sin^2\left(x^i\right)$.

Function f_{NS} is non–convex and has one of the hardest surface for optimization because of its fluctuating landscape created using superpositions of Chebyshev polynomials of first kind $P_2\left(x^i\right) = 2\left(x^i\right)^2 - 1$, function has unique minimum point $x^* = (1, \ldots, 1)$. Function f_H is non–convex, has quadratic growth property and all its minima are global minima with $\|x^*\| = 1$. Function PL is non–convex, it is bounded by paraboloids from both sides and also satisfies quadratic growth property, this function has unique global minimum $x^* = \mathbf{0}_n$.

We fix random seed for reproducibility of the experiments. For numerical stability reasons we clip absolute values for all variates to stay within Chebyshev ball with radius 10^{12} centered at origin. For all symmetric matrix inversion operations we also clip matrix spectra by 10^{-6} from below and by 10^{12} from above. For efficient and stable computation of x_{k+1} we consider matrix factorizations described in the next subsection.

E.1 Fast binary search of the local Lipschitz constant

The most expensive operation in the designed algorithms is matrix inversion, so we use matrix factorization with an asymptotic cost of the one unoptimized iteration to have linear w.r.t. $\min\{m, n\}$ in asymptotics matrix inversion at each inner iteration. Firstly, we perform factorization of the update direction towards x_{k+1}. For simplicity we consider deterministic case, however we can extend the factorization to stochastic setting by substitution of the local model. The value $\min\{m, n\}$ points out the necessity to consider two cases: $m > n$ and $m \leq n$.

In the first case we use eigendecomposition of the matrix $\hat{F}'(x_k)^* \hat{F}'(x_k)$:

$$\begin{aligned}
&\hat{F}'(x_k)^* \hat{F}'(x_k) = Q_n \Lambda_n Q_n^*,\ Q_n^* Q_n = I_n,\ \Lambda_n \text{ is a diagonal matrix;} \\
&\qquad \hat{F}'(x_k)^* \hat{F}'(x_k) + \tau_k L_k I_n = Q_n \Lambda_n Q_n^* + \tau_k L_k I_n = Q_n \left(\Lambda_n + \tau_k L_k I_n\right) Q_n^* \Rightarrow \\
&\qquad \Rightarrow \left(\hat{F}'(x_k)^* \hat{F}'(x_k) + \tau_k L_k I_n\right)^{-1} = Q_n \left(\Lambda_n + \tau_k L_k I_n\right)^{-1} Q_n^*,\ m > n \Rightarrow \\
&\qquad \Rightarrow x_{k+1} = x_k - \eta_k Q_n \left(\Lambda_n + \tau_k L_k I_n\right)^{-1} Q_n^* \hat{F}'(x_k)^* \hat{F}(x_k).
\end{aligned}$$

For the expressions above we have $\mathrm{O}(n)$ complexity of the matrix inversion $\left(\Lambda_n + \tau_k L_k I_n\right)^{-1}$. The eigendecomposition has complexity cost $\mathrm{O}(n^3)$ achievable using the divide–and–conquer algorithm for tridiagonalization with Householder reflections [13, 7, 4]. Note that orthogonal matrix Q_n and diagonal matix Λ_n occupy $\mathrm{O}(n^2 + n)$ memory. Fixed vector $Q_n^* \hat{F}'(x_k)^* \hat{F}(x_k)$ can be computed using

$\mathrm{O}(n^2+nm)$ operations, matrix multiplication of Q_n and $(\Lambda_n+\tau_k L_k I_n)^{-1} Q_n^* \hat{F}'(x_k)^* \hat{F}(x_k)$ uses $\mathrm{O}(n^2)$ operations. The whole number of inner iterations is bounded by

$$\left\lceil \log_2 \left(\frac{\gamma L_{\hat{F}}}{L} \right) \right\rceil + 1,\ \gamma \geq 2,\ L \in \left(0,\ L_{\hat{F}}\right],$$

because the local Lipschitz constant is estimated via some binary search–like procedure. Also, we have at most only two inner iterations after hitting $L_{k-1} \in \left[L_{\hat{F}},\ 2L_{\hat{F}}\right]$ at the k-th step. So, we have the overall cost of the optimized step:

$$\mathrm{O}\left(n^3+n^2+mn+n(n+1)\left(\left\lceil \log_2 \left(\frac{\gamma L_{\hat{F}}}{L} \right) \right\rceil + 1 \right) \right),$$

which stays conceptually the same in stochastic settings with m substituted with b, assuming $\tilde{b} \leq b$.

For the second case we use Sherman–Morrison–Woodbury formula and an eigendecomposition to have matrix inversion with the const $\mathrm{O}(m)$. We perform the eigendecomposition for symmetric matrix $\hat{F}'(x_k)\hat{F}'(x_k)^*$ using $\mathrm{O}(m^3)$ operations and $\mathrm{O}(m^2+m)$ memory:

$$\begin{aligned}
\hat{F}'(x_k)\hat{F}'(x_k)^* &= Q_m \Lambda_m Q_m^*,\ Q_m^* Q_m = I_m,\ \Lambda_m \text{ is a diagonal matrix;} \\
&\left(\hat{F}'(x_k)^* \hat{F}'(x_k) + \tau_k L_k I_n \right)^{-1} = \frac{1}{\tau_k L_k} I_n - \\
&- \frac{1}{\tau_k L_k} \hat{F}'(x_k)^* \left(\tau_k L_k I_m + \hat{F}'(x_k)\hat{F}'(x_k)^* \right)^{-1} \hat{F}'(x_k) = \\
&= \frac{1}{\tau_k L_k} I_n - \frac{1}{\tau_k L_k} \hat{F}'(x_k)^* Q_m \left(\tau_k L_k I_m + \Lambda_m \right)^{-1} Q_m^* \hat{F}'(x_k),\ m \leq n \Rightarrow \\
\Rightarrow x_{k+1} &= x_k - \eta_k \left(\frac{1}{\tau_k L_k} I_n - \right. \\
&\left. - \frac{1}{\tau_k L_k} \hat{F}'(x_k)^* Q_m \left(\tau_k L_k I_m + \Lambda_m \right)^{-1} Q_m^* \hat{F}'(x_k) \right) \hat{F}'(x_k)^* \hat{F}(x_k) = \\
&= x_k - \frac{\eta_k}{\tau_k L_k} \left(\hat{F}'(x_k)^* \hat{F}(x_k) - \hat{F}'(x_k)^* Q_m \left(\tau_k L_k I_m + \Lambda_m \right)^{-1} \Lambda_m Q_m^* \hat{F}(x_k) \right) = \\
&= x_k - \frac{\eta_k}{\tau_k L_k} \hat{F}'(x_k)^* \left(\hat{F}(x_k) - Q_m \left(\tau_k L_k I_m + \Lambda_m \right)^{-1} \Lambda_m Q_m^* \hat{F}(x_k) \right).
\end{aligned}$$

We compute vector $\Lambda_m Q_m^* \hat{F}(x_k)$ using $\mathrm{O}\left(m^2+m\right)$ operations, and x_{k+1} is computed with the cost $\mathrm{O}\left(m^2+mn+m+n\right)$. So, we have the following cost of the step:

$$\mathrm{O}\left(m^3+m^2+m+\left(m^2+mn+m+n\right)\left(\left\lceil \log_2 \left(\frac{\gamma L_{\hat{F}}}{L} \right) \right\rceil + 1 \right) \right),$$

which also stays conceptually the same in stochastic settings with m substituted with b, assuming $\tilde{b} \leq b$. But for doubly stochastic step with $\tilde{b} = b$ we have another form of the fast update:

$$\begin{aligned}
x_{k+1} &= x_k - \eta_k \left(\frac{1}{\tilde{\tau}_k L_k} I_n - \right. \\
&\left. - \frac{1}{\tilde{\tau}_k L_k} \hat{G}'(x_k, \tilde{B}_k)^* Q_b \left(\tilde{\tau}_k L_k I_b + \Lambda_b \right)^{-1} Q_b^* \hat{G}'(x_k, \tilde{B}_k) \right) \hat{G}'(x_k, B_k)^* \hat{G}(x_k, B_k) = \\
&= x_k - \frac{\eta_k}{\tilde{\tau}_k L_k} \left(\hat{G}'(x_k, B_k)^* \hat{G}(x_k, B_k) - \right. \\
&\left. - \hat{G}'(x_k, \tilde{B}_k)^* Q_b \left(\tilde{\tau}_k L_k I_b + \Lambda_b \right)^{-1} Q_b^* \hat{G}'(x_k, \tilde{B}_k) \hat{G}'(x_k, B_k)^* \hat{G}(x_k, B_k) \right); \\
&\hat{G}'(x_k, \tilde{B}_k)\hat{G}'(x_k, \tilde{B}_k)^* = Q_b \Lambda_b Q_b^*,\ Q_b^* Q_b = I_b,\ \Lambda_b \text{ is a diagonal matrix,}
\end{aligned}$$

with the overall computational complexity of the step:

$$\mathrm{O}\left(b^3 + b^2(n+1) + bn + (bn+b+n)\left(\left\lceil \log_2\left(\frac{\gamma L_{\hat{F}}}{L}\right)\right\rceil + 1\right)\right),$$

if we use binary search for L_k, otherwise we have the following complexity:

$$\mathrm{O}\left(b^3 + b^2(n+1) + bn + b + n\right),$$

which is also cheaper for $n \gg 1$ than straightforward computation because we assumed $\tilde{b} = b \leq n$. Besides, the described factorizations allow us to clip spectrum of diagonal matrices to achieve numerically stable matrix inversion.

E.2 The performance of deterministic Gauss–Newton method

For the experiment we average every combination of the setting over 5 runs. For each run we sample initial value x_0 from standard normal multidimensional distribution. For deterministic Gauss–Newton method we use the exact oracle with $\eta_k = 1$ and set $\tau_k = \hat{f}_1(x_k)$ and $\varepsilon_k = 0$. We also use inequality $\tau_k \leq 10^{-6}$ as an early stopping criterion and define $L_0 = 1$. The maximal number of outer iterations equals 10^2. We test benchmark functions on different values of n: 10, 10^2 and 10^3. All depicted uncertainty intervals have two standard pointwise deviations width.

Figure 1 shows us sublinear convergence on function f_{NS}, while figure 3 shows us linear convergence with a major slowdown near the end of optimization procedure achieving a saddle point due to trigonometric fluctuations. Meanwhile, figure 2 shows us typical local superlinear convergence. All tested benchmark functions are unbounded but the experiments show us that it is sufficient to stay within the region of bounded values to achieve convergence rates proved for bounded functionals.

E.3 The performance of stochastic Gauss–Newton method with scaled step

For stochastic settings we average every combination of hyperparameters over the same set of initial points used in deterministic Gauss–Newton method. We fix $n = 10^3$ and use constant step size $\eta_k = \eta \in (0,\ 1]$, we also set $L_0 = 1$ and $\tau_k = \hat{g}_1(x_k, B_k)$. The maximal number of outer iterations equals 10^2. For experimental runs we use the following ranges of hyperparameters:

- batch size $b \in \{1, 10, 10^2, 10^3\}$;
- step scale $\eta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$.

Every run stands for the combination of hyperparameters taken as an element of cartesian product of the sets above and depicted uncertainty intervals have two standard pointwise deviations width. The stochastic Gauss–Newton method uses the same early stopping criterion as the deterministic method: $\tau_k \leq 10^{-6}$.

For stochastic setting we have preservance of convergence types from deterministic setting as averaged line show. Figure 4 stands for processes with sublinear convergence, figure 5 describes processes with local superlinear convergence, while figure 6 establishes linear convergence. Obviously, these figures state the increasing of the batch size leads to speedup of the convergence for conventional optimization tasks achieving better interpolation. The increase of the step scale up to 1 also causes such effects.

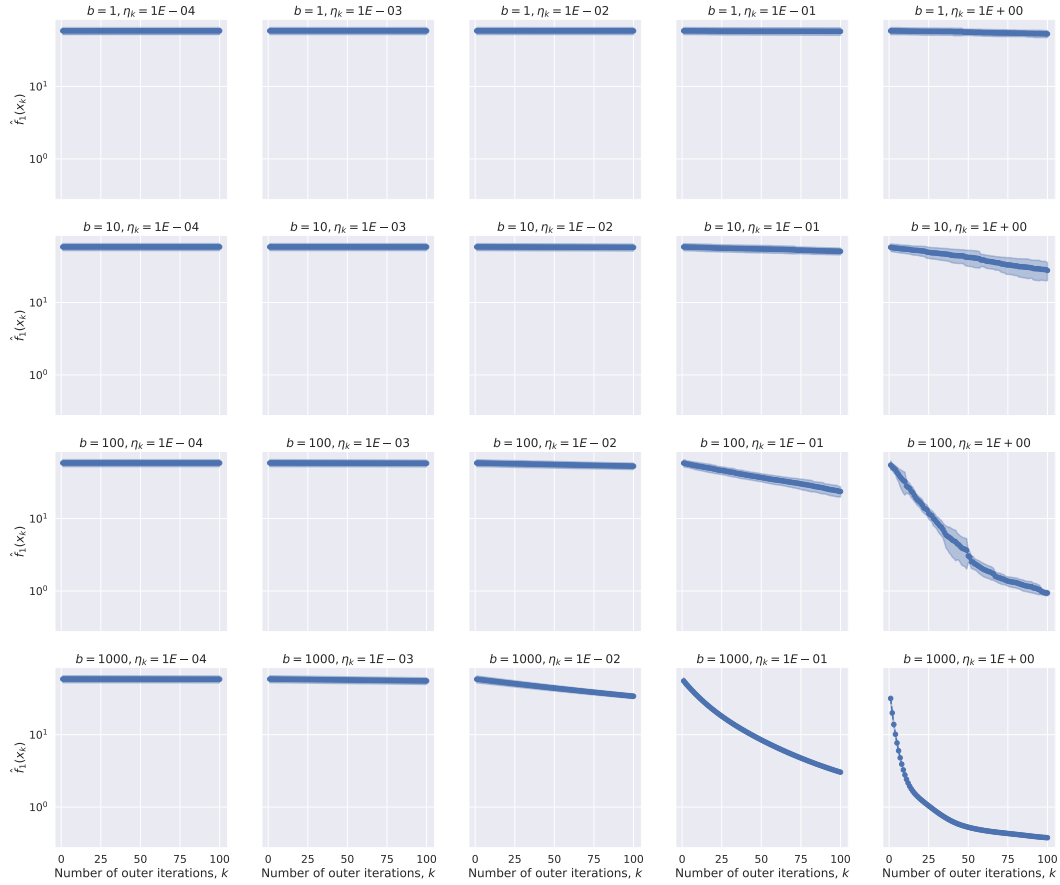

Figure 4: The performance of stochastic Gauss–Newton method with scaled step on Nesterov–Skokov function

E.4 The performance of doubly stochastic step usage

For these stochastic settings we also average every combination of hyperparameters over the same set of initial points used in deterministic Gauss–Newton method. We bound the maximal number of outer iterations by 10^2. We use constant values of $\tilde{\tau}_k L_k = \tilde{\tau} L$ and $\eta_k = \eta \tilde{\tau} L$ to simulate conditions similar to conditions from theorem 6. In doubly stochastic case we also set $n = 10^3$ and for experimental runs we use the following ranges of hyperparameters:

- batch size $b = \tilde{b} \in \{1, 10, 10^2, 10^3\}$;
- step scale $\eta \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$;
- value $\tilde{\tau} L \in \{10^{-4}, 1, 10^3, 10^6, +\infty\}$.

Every run stands for the combination of hyperparameters taken as an element of cartesian product of the sets above.

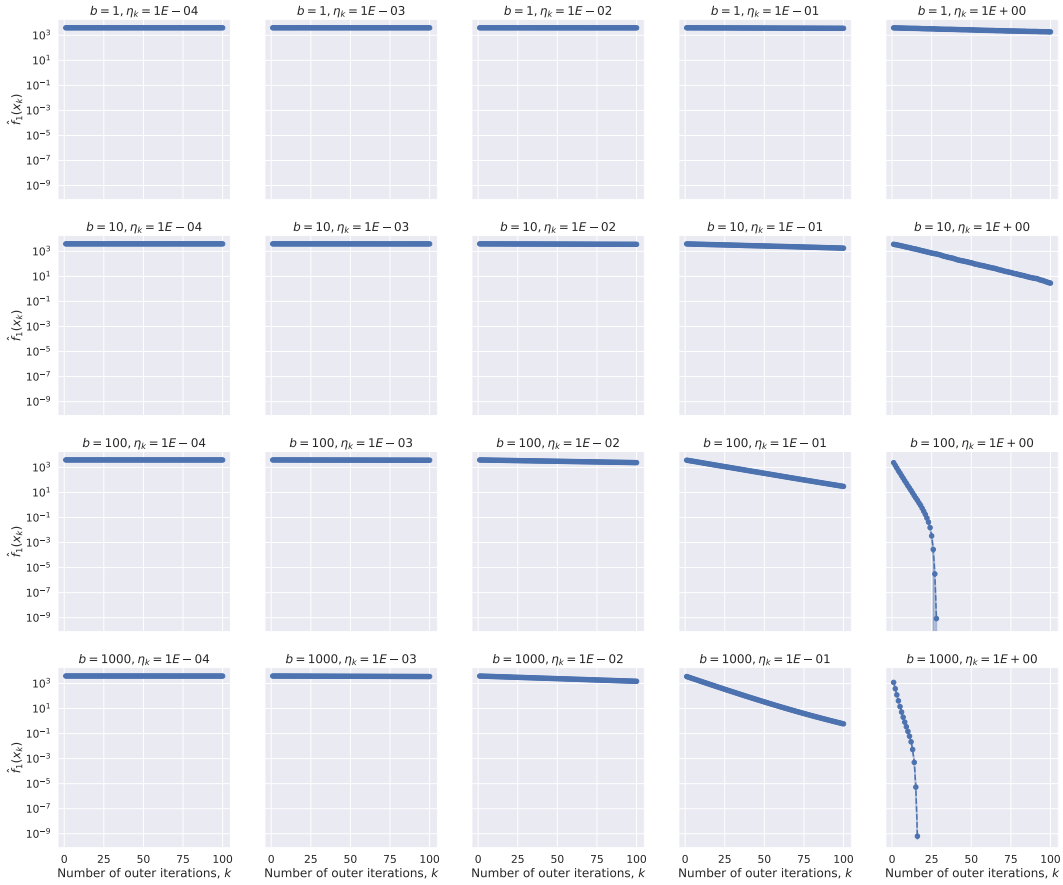

Figure 5: The performance of stochastic Gauss–Newton method with scaled step on Hat function

Unlike to the previous case, doubly stochastic setting requires to know the global Lipschitz constant of function $\hat{f}_2$, and figures 7, 8 and 9 show that the lack of such information leads to slower convergence and even to divergence. In the presented figures column with $\tau_k L_k = +\infty$, $\tilde{\tau}_k = \tau_k$ stands for the gradient descend method. Experiments with doubly stochastic step show that gradient and stochastic gradient methods perform no better than corresponding Gauss–Newton methods, especially with the increase of batch size. And only for small values of η_k these methods possess similar quality under small batch size: 1 and 10. Another observation states the "harder" function to optimize, the more quality gap between gradient methods and Gauss–Newton methods. Such criterion allows us to order functions with increasing of "the hardness" of unconstrained minimization problem to find stationary point: $f_{PL} \preceq f_H \preceq f_{NS}$.

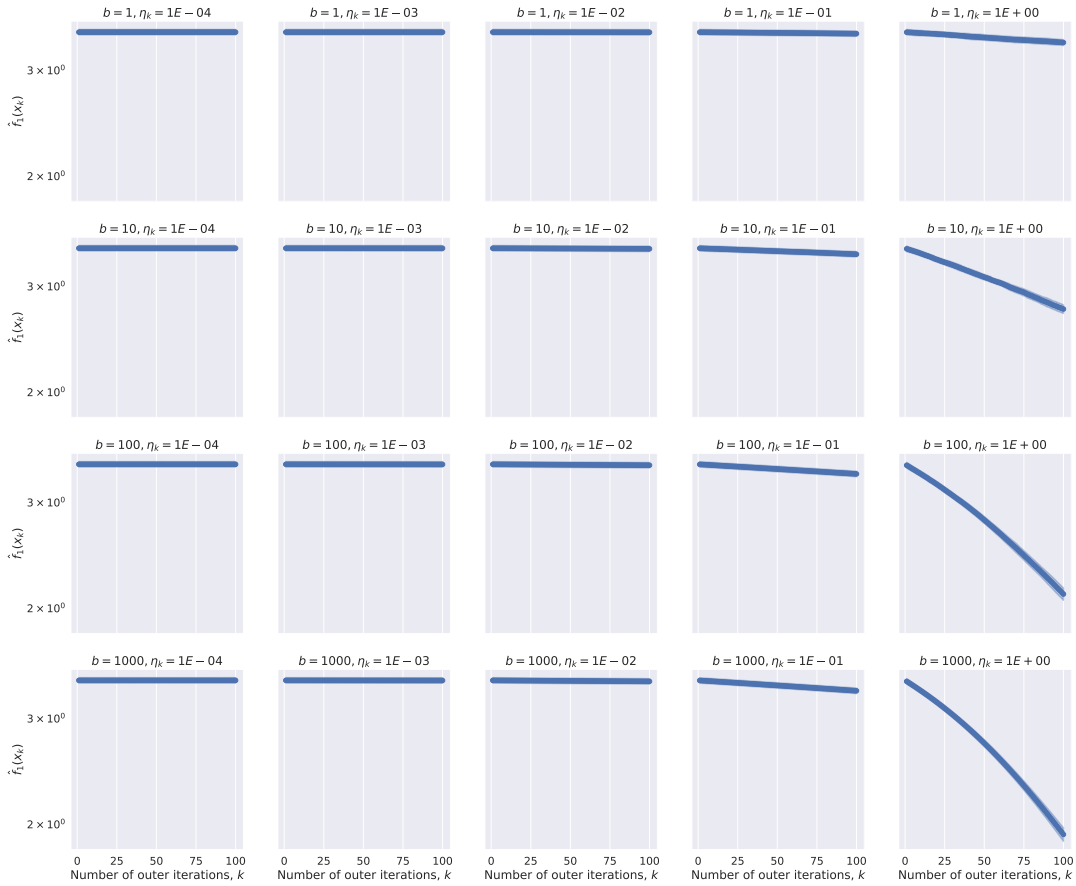

Figure 6: The performance of stochastic Gauss–Newton method with scaled step on PL function

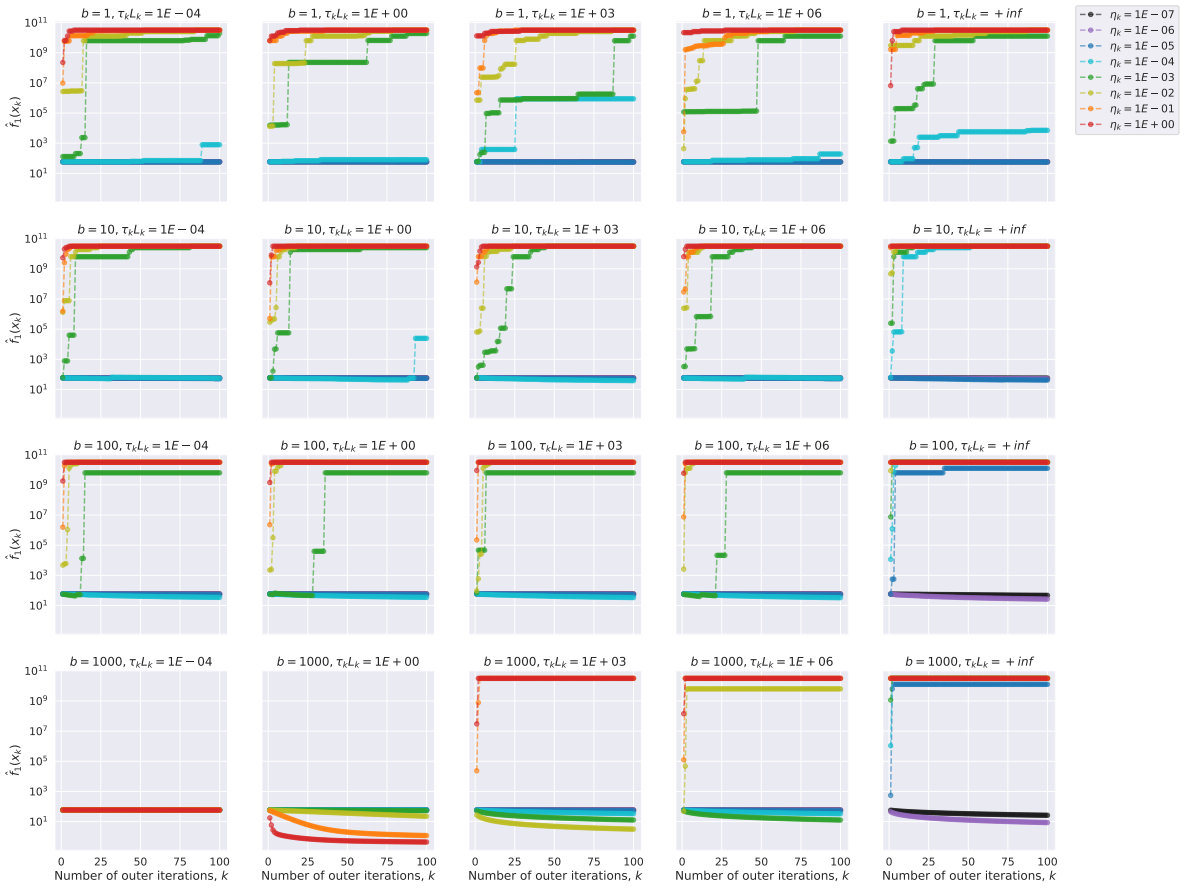

Figure 7: The performance of stochastic Gauss–Newton method with doubly stochastic step on Nesterov–Skokov function

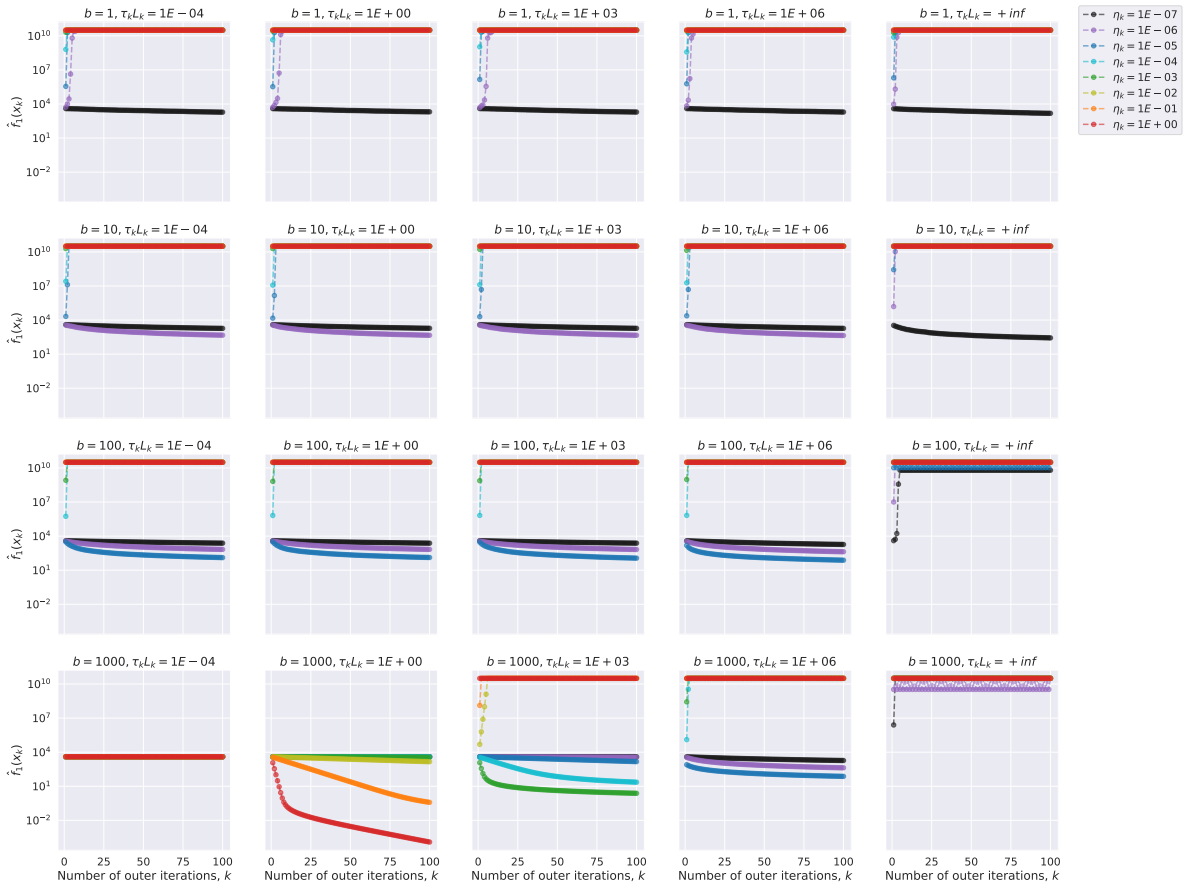

Figure 8: The performance of stochastic Gauss–Newton method with doubly stochastic step on Hat function

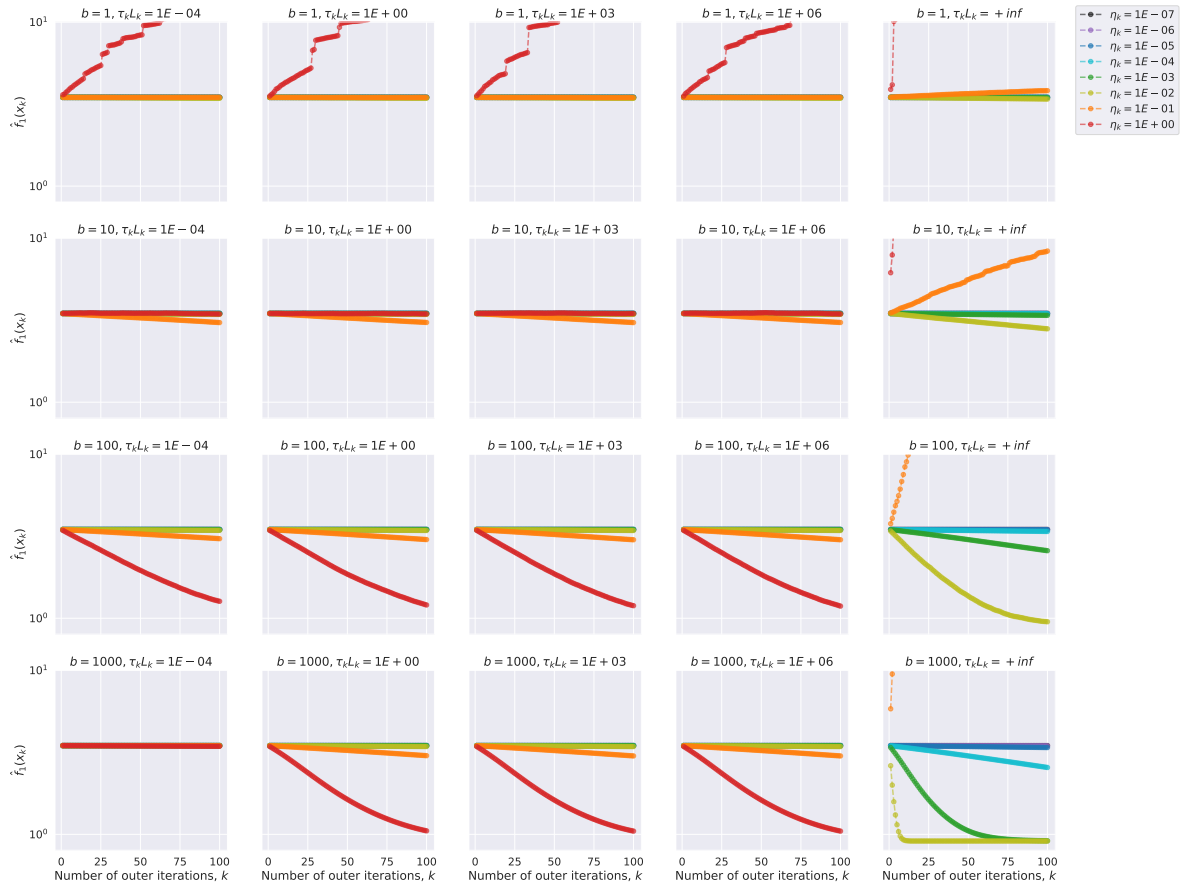

Figure 9: The performance of stochastic Gauss–Newton method with doubly stochastic step on PL function