Supplementary material for

A Superlinearly-Convergent Proximal Newton-type Method for the Optimization of Finite Sums

Contents

1 Notation

In what follows we work only with Euclidean norms:

$$\|x\| := \sqrt{x^\top x}, \qquad \text{and} \qquad \|x\|_H := \sqrt{x^\top H x}, \qquad x \in \mathbf{R}^d,$$

where H is a symmetric positive definite matrix.

We also use the following two proximal mappings:

$$\operatorname{prox}_h(x) := \operatorname*{argmin}_{y \in \mathbf{R}^d} \left[h(y) + \frac{1}{2} \|y - x\|^2 \right],$$

$$\operatorname{prox}_h^H(x) := \operatorname*{argmin}_{y \in \mathbf{R}^d} \left[h(y) + \frac{1}{2} \|y - x\|_H^2 \right].$$

2 Auxiliary lemmas

Lemma 1. *Let $w_1, \ldots, w_n \in \mathbf{R}^d$ be any vectors. Then*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} w_i \right\| \leq \left(\frac{1}{n} \sum_{i=1}^{n} \|w_i\|^2 \right)^{1/2}.$$

Proof. Denote $w := [w_1 \ldots w_n]^\top \in \mathbf{R}^{nd}$ and $E := [I \ldots I]^\top \in \mathbf{R}^{nd \times d}$, where $I \in \mathbf{R}^{d \times d}$ is the corresponding identity matrix. Then

$$\left\| \frac{1}{n} \sum_{i=1}^{n} w_i \right\| = \frac{1}{n} \left\| E^\top w \right\| \leq \frac{1}{n} \|E\| \, \|w\| = \left(\frac{1}{n} \sum_{i=1}^{n} \|w_i\|^2 \right)^{1/2},$$

because $\|E\| = \lambda_{\max}^{1/2}(E^\top E) = \lambda_{\max}^{1/2}(nI) = n^{1/2}$. □

Lemma 2. *Suppose the gradients ∇f_i are Lipschitz-continuous:*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_f \|x - y\|, \qquad i = 1, \ldots, n.$$

Then, for any minimizer x^* of $\phi(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x)$, we have the following two inequalities:

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n} v_k^i - \operatorname{prox}_h \left(\frac{1}{n} \sum_{i=1}^{n} v_k^i - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(v_k^i) \right) \right\| &\leq \frac{L_f + 2}{n} \sum_{i=1}^{n} \left\| v_k^i - x^* \right\| \\ &\leq (L_f + 2) \left(\frac{1}{n} \sum_{i=1}^{n} \left\| v_k^i - x^* \right\|^2 \right)^{1/2}. \end{aligned}$$

Proof. Since x^* is a minimizer of ϕ, it satisfies $x^* = \operatorname{prox}_h(x^* - (1/n) \sum_{i=1}^{n} \nabla f_i(x^*))$. Using this expression, the non-expansiveness of $\operatorname{prox}_h(\cdot)$ and the Lipschitz-continuity of ∇f_i, we get the following chain of inequalities:

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^{n} v_k^i - \operatorname{prox}_h \left(\frac{1}{n} \sum_{i=1}^{n} v_k^i - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(v_k^i) \right) \right\| \\ &\qquad \leq \left\| \frac{1}{n} \sum_{i=1}^{n} v_k^i - x^* \right\| + \left\| x^* - \operatorname{prox}_h \left(\frac{1}{n} \sum_{i=1}^{n} v_k^i - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(v_k^i) \right) \right\| \\ &\qquad \leq \frac{2}{n} \sum_{i=1}^{n} \left\| v_k^i - x^* \right\| + \frac{1}{n} \left\| \nabla f_i(v_k^i) - \nabla f_i(x^*) \right\| \\ &\qquad \leq \frac{L_f + 2}{n} \sum_{i=1}^{n} \left\| v_k^i - x^* \right\|. \end{aligned}$$

Thus, the first inequality of the lemma is proved. The other inequality follows from Lemma 1. □

3 Local convergence rate: simple case

In this section we consider the situation when $h \equiv 0$, i.e. we apply NIM for minimizing the function

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Recall that at each iteration NIM works with the following model of the objective:

$$m_k(x) = \frac{1}{n} \sum_{i=1}^{n} \left[f_i(v_k^i) + \nabla f_i(v_k^i)^\top (x - v_k^i) + \frac{1}{2} (x - v_k^i)^\top \nabla^2 f_i(v_k^i)(x - v_k^i) \right],$$

where v_k^i are some points that are updated in iterations (one point at every iteration). Since the model does not contain the term $h(x)$ (it is zero), we can write down the minimum of m_k in the closed form:

$$\bar{x}_k := \operatorname*{argmin}_{x \in \mathbf{R}^d} m_k(x) = H_k^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \nabla^2 f_i(v_k^i) v_k^i - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(v_k^i) \right),$$

where $H_k := \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f_i(v_k^i)$. In this section we focus on the simple case when we are able to calculate $\bar{x}_k$ exactly, so there is no additional error associated with inexact model minimization.

3.1 Theorem statement

In what follows we prove the following theorem on the local convergence rate of NIM.

Theorem 1 (local convergence rate)**.** *Suppose the Hessians $\nabla^2 f_i$ are Lipschitz-continuous:*

$$\left\| \nabla^2 f_i(x) - \nabla^2 f_i(y) \right\| \le M_f \left\| x - y \right\|, \qquad i = 1, \ldots, n,$$

for all $x, y \in \mathbf{R}^d$. Let $\{x_k\}_{k \ge n}$ be the sequence of iterates generated by NIM with the unit step size $\alpha_k \equiv 1$ and cyclic order of component selection. Assume x^* is a minimizer of f with positive definite Hessian:

$$\nabla^2 f(x^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f_i(x^*) \ge \mu_f I, \qquad \mu_f > 0, \tag{1}$$

and all the initial points $x_0, \ldots, x_{n-1}$ are close enough to x^*:

$$\|x_i - x^*\| \le R, \qquad i = 0, \ldots, n-1. \tag{2}$$

Then the sequence $\{x_k\}_{k \ge 0}$ converges to x^* at an R-superlinear rate, i. e. there exists $\{z_k\}_{k \ge 0}$ such that

$$\begin{aligned} \|x_k - x^*\| &\le z_k, & k &\ge 0 \\ z_{k+1} &\le q_k z_k, & k &\ge n, \end{aligned}$$

where $q_k \to 0$ and $z_0 = \cdots = z_{n-1} := \max_{0 \le i \le n-1} \|x_i - x^*\|$.

The expressions for R and q_k are as follows:

$$R := \frac{\mu_f}{2M_f}, \qquad q_k := \left(1 - \frac{3}{4n}\right)^{2^{\lfloor k/n \rfloor - 1}}.$$

3.2 Main estimate

When considering local convergence, we assume that NIM uses the unit step length $\alpha \equiv 1$ at every iteration and the order of updating the points v_k^i is cyclic:

$$\begin{aligned} x_{k+1} &= \bar{x}_k, \\ v_{k+1}^i &= \begin{cases} x_{k+1} & \text{if } i = k \bmod n + 1 \\ v_k^i & \text{otherwise.} \end{cases} \end{aligned}$$

Note that the cyclic order of updating means the points v_k^i, $i = 1, \ldots, n$, are exactly the last n iterates $x_k, x_{k-1}, \ldots, x_{k-n+1}$ (but possibly in different order).

Lemma 3 (main estimate)**.** *Let $k \ge n-1$ be the number of the current iteration. Assume the last n points $x_k, \ldots, x_{k-n+1}$ are close enough to x^*,*

$$\|x_{k-i} - x^*\| \le \frac{\mu_f}{2M_f}, \qquad i = 0, \ldots, n-1. \tag{3}$$

Then for the next generated point x_{k+1}, the following bound holds:

$$\|x_{k+1} - x^*\| \le \frac{M_f}{\mu_f} \left(\frac{1}{n} \sum_{i=0}^{n-1} \|x_{k-i} - x^*\|^2 \right). \tag{4}$$

Proof. Recall the iteration of NIM:

$$x_{k+1} = H_k^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)v_k^i - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(v_k^i)\right),$$

where $H_k := \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)$.

Since x^* is a minimizer of f, we have $0 = \nabla f(x^*) = \sum_{i=1}^{n}\nabla f_i(x^*)$. Using this equality, we get

$$\|x_{k+1} - x^*\| \le \left\|H_k^{-1}\right\| \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)[v_k^i - x^*] - \frac{1}{n}\sum_{i=1}^{n}[\nabla f_i(v_k^i) - \nabla f_i(x^*)]\right\|.$$

Note that $\left\|H_k^{-1}\right\| = 1/\lambda_{\min}(H_k)$, so we obtain

$$\|x_{k+1} - x^*\| \le \frac{1}{\lambda_{\min}(H_k)} \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)(v_k^i - x^*) - \frac{1}{n}\sum_{i=1}^{n}[\nabla f_i(v_k^i) - \nabla f_i(x^*)]\right\|.$$

Now we use the Taylor formula for gradients and Lipschitz-continuity of $\nabla^2 f_i$:

$$\begin{aligned}
&\frac{1}{\lambda_{\min}(H_k)} \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)(v_k^i - x^*) - \frac{1}{n}\sum_{i=1}^{n}[\nabla f_i(v_k^i) - \nabla f_i(x^*)]\right\| \\
&\qquad = \frac{1}{\lambda_{\min}(H_k)} \left\|\frac{1}{n}\sum_{i=1}^{n}\int_0^1 [\nabla^2 f_i(v_k^i) - \nabla^2 f_i(v_k^i + \tau(x^* - v_k^i))](v_k^i - x^*)d\tau\right\| \\
&\qquad \le \frac{M_f}{2\lambda_{\min}(H_k)}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|v_k^i - x^*\right\|^2\right).
\end{aligned}$$

Thus,

$$\|x_{k+1} - x^*\| \le \frac{M_f}{2\lambda_{\min}(H_k)}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|v_k^i - x^*\right\|^2\right). \tag{5}$$

Let us estimate $\lambda_{\min}(H_k)$. First, let us bound the difference between H_k and $\nabla^2 f(x^*)$:

$$\left\|H_k - \nabla^2 f(x^*)\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n}[\nabla^2 f_i(v_k^i) - \nabla^2 f_i(x^*)]\right\| \le \frac{M_f}{n}\sum_{i=1}^{n}\left\|v_k^i - x^*\right\|.$$

Since the order of component selection is cyclic, the points $v_k^i,\ i = 1, \ldots, n$, are exactly the last n iterates $x_{k-i},\ i = 0, \ldots, n-1$, (possibly rearranged). Therefore,

$$\left\|H_k - \nabla^2 f(x^*)\right\| = \frac{M_f}{n}\sum_{i=0}^{n-1}\|x_{k-i} - x^*\| \le \frac{\mu_f}{2},$$

where the inequality follows from (3). Using this bound together with (1), we get

$$\lambda_{\min}(H_k) \ge \lambda_{\min}(\nabla^2 f(x^*)) - \left\|H_k - \nabla^2 f(x^*)\right\| \ge \frac{\mu_f}{2}.$$

Thus, $\lambda_{\min}(H_k) \ge \mu_f/2$. Using this in (5) and replacing the sums involving v_k^i with the sums involving x_{k-i}, we get (4). □

3.3 Convergence rate of the sequence

Let us investigate the convergence properties of the following (recurrent) sequence arising in (4):

$$z_k := A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right), \qquad k \ge n, \tag{6}$$

where $A > 0$ is some constant.

First, let us understand the conditions under which the sequence $\{z_k\}_{k\geq 0}$ (monotonically) converges to zero. Note that from (6) it follows that

$$z_k \leq A \max_{1\leq i\leq n} \{z_{k-i}\}^2 = \left(A \max_{1\leq i\leq n} \{z_{k-i}\}\right) \max_{1\leq i\leq n} \{z_{k-i}\},$$

and the equality may hold when $z_{k-1} = \cdots = z_{k-n}$. Therefore to guarantee the (monotonic) convergence of $\{z_k\}_{k\geq 0}$ to zero, we must enforce the following condition on the initial elements $z_0, \ldots, z_{n-1}$:

$$Az_i < 1, \qquad i = 0, \ldots, n-1.$$

In particular, we may require that

$$z_i \leq \frac{1}{2A}, \qquad i = 0, \ldots, n-1. \tag{7}$$

This will guarantee that

$$z_k \leq \frac{1}{2} \max_{1\leq i\leq n} \{z_{k-i}\}, \qquad k \geq n. \tag{8}$$

In what follows we always assume that the initial elements of $\{z_k\}_{k\geq 0}$ satisfy (7).

In view of (8), the sequence $\{z_k\}_{k\geq 0}$ converges to zero. However, this convergence may be non-monotonic. For example, if $z_0 > 0$ (but small enough), and $z_1 = \cdots = z_{n-1} = 0$, then $z_n > z_{n-1}$. Nevertheless, it turns out that if the initial elements of $\{z_k\}_{k\geq 0}$ are initialized with the same number, $z_0 = \cdots = z_{n-1}$, then the sequence $\{z_k\}_{k\geq 0}$ is monotonic.

Lemma 4 (monotonicity)**.** *Let $z_0 = \cdots = z_{n-1}$. Then $\{z_k\}_{k\geq 0}$ is monotonic: $z_{k+1} \leq z_k$ for all $k \geq 0$.*

Proof. According to (8) we have $z_n \leq z_{n-1}$. Thus, we know that $z_0 = \cdots = z_{n-1} \geq z_n$.

We proceed by induction. Suppose we know that $z_0 \geq \cdots \geq z_k$ for some $k \geq n$. We will prove that this implies $z_k \geq z_{k+1}$. Indeed, according to the induction hypothesis, $z_k \leq z_{k-n}$. Therefore,

$$\sum_{i=1}^{n} z_{k+1-i}^2 = z_k^2 + \sum_{i=1}^{n-1} z_{k-i}^2 \leq z_{k-n}^2 + \sum_{i=1}^{n-1} z_{k-i}^2 = \sum_{i=1}^{n} z_{k-i}^2.$$

Using this and the definition (6) of $\{z_k\}_{k\geq n}$, we have

$$z_{k+1} = A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right) \leq A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right) = z_k.$$

□

From now on, in addition to (7), we assume that $z_0 = \cdots = z_{n-1}$. Due to the monotonicity of $\{z_k\}_{k\geq 0}$, inequality (8) now becomes

$$z_k \leq \frac{1}{2} z_{k-n}, \qquad k \geq n. \tag{9}$$

Using the monotonicity, we can prove that the convergence rate of $\{z_k\}_{k\geq 0}$ is at least linear.

Lemma 5 (linear convergence)**.** *The convergence rate of $\{z_k\}_{k\geq 0}$ is at least linear:*

$$z_{k+1} \leq \left(1 - \frac{3}{4n}\right) z_k, \qquad k \geq n. \tag{10}$$

Proof. Note that

$$\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2 = \frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2 - \frac{1}{n}[z_{k-n}^2 - z_k^2] = \left(1 - \frac{z_{k-n}^2 - z_k^2}{\sum_{i=1}^{n} z_{k-i}^2}\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right).$$

Let us find a lower bound for the fraction. Using $\sum_{i=1}^{n} z_{k-i}^2 \leq n z_{k-n}^2$ and (9), we have

$$\frac{z_{k-n}^2 - z_k^2}{\sum_{i=1}^{n} z_{k-i}^2} \geq \frac{1 - 1/4}{n} = \frac{3}{4n}.$$

Thus, we have proved:

$$\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2 \leq \left(1 - \frac{3}{4n}\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right).$$

Using this inequality and the definition (6) of $\{z_k\}_{k\geq n}$, we obtain

$$z_{k+1} = A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right) \leq \left(1 - \frac{3}{4n}\right) A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right) \leq \left(1 - \frac{3}{4n}\right) z_k.$$

□

The next lemma shows that the convergence constant in (10) improves after every n iterations.

Lemma 6 (improving the constant)**.** *Suppose that, starting from number $k_0 \geq n$, the sequence $\{z_k\}_{k\geq 0}$ converges linearly to zero with constant q_0:*

$$z_{k+1} \leq q_0 z_k, \qquad k \geq k_0. \tag{11}$$

Then, starting from number $k_0 + n$, the constant q_0 can be replaced with a smaller constant:

$$z_{k+1} \leq q_0^2 z_k, \qquad k \geq k_0 + n.$$

Proof. Let $k \geq k_0 + n$. Using the definition (6) of $\{z_k\}_{k\geq n}$ and bound (11), we have the following chain of inequalities:

$$z_{k+1} = A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right) \leq q_0^2 A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right) \leq q_0^2 z_k.$$

□

Let us summarize the results we have established.

Lemma 7. *Let $\{z_k\}_{k\geq 0}$ be a recurrent sequence defined in* (6)*. Suppose the initial elements $z_0, \ldots, z_{n-1}$ of this sequence are chosen equal to the same number small enough:*

$$z_0 = \cdots = z_{n-1} \leq R.$$

Then the sequence $\{z_k\}_{k\geq 0}$ converges monotonically to zero at a Q-superlinear rate:

$$z_{k+1} \leq q_k z_k, \qquad k \geq n.$$

The expressions for R and q_k are as follows:

$$R := \frac{1}{2A}, \qquad q_k := \left(1 - \frac{3}{4n}\right)^{2^{\lfloor k/n \rfloor - 1}}.$$

Proof. Denote $q := (1 - 3/4n)$. According to (10) and Lemma 6 we can write the following sequence of inequalities:

$$\begin{aligned} z_{k+1} &\leq q z_k, && k \geq n, \\ z_{k+1} &\leq q^2 z_k, && k \geq 2n, \\ z_{k+1} &\leq q^4 z_k, && k \geq 3n, \\ &\cdots \end{aligned}$$

Combining all these inequalities together, we get

$$z_{k+1} \leq q^{2^{\lfloor k/n \rfloor - 1}} z_k, \qquad k \geq n.$$

□

3.4 Proof of the theorem

Now we can give the proof of the theorem on the local convergence rate of NIM.

Proof. Consider the sequence $\{z_k\}_{k\geq 0}$ defined in (6) with $A := M_f/\mu_f$. Let us set the initial elements $z_0, \ldots, z_{n-1}$ of this sequence to the same value:

$$z_0 = \cdots = z_{n-1} := \max_{0\leq i\leq n-1} \|x_i - x^*\| \leq R.$$

According to Lemma 7, the sequence $\{z_k\}_{k\geq 0}$ converges monotonically and Q-superlinearly to zero with constants q_k. In particular, it means that $\{z_k\}_{k\geq 0}$ always stays bounded:

$$z_k \leq R \leq \frac{\mu_f}{2M_f}, \qquad k \geq 0. \tag{12}$$

Due to the initial condition (2), we can apply Lemma 3 for $k = n-1$. Since, by construction, the values $\|x_i - x^*\|$ are bounded above by z_i for $i = 0, \ldots, n-1$, we have $\|x_n - x^*\| \leq z_n$. In view of (12), it means that the new iterate x_n does not leave the R-vicinity of x^*. Therefore, we can apply Lemma 3 again but for $k = n$. Using the same reasoning, we conclude that $\|x_{n+1} - x^*\| \leq z_{n+1} \leq R$, and so on. Thus, Lemma 3 holds for all $k \geq n-1$ and the sequence $\{z_k\}_{k\geq 0}$ majorizes $\{\|x_k - x^*\|\}_{k\geq 0}$. □

4 Local convergence rate: general case

In this section we consider the more general situation than in Section 3—the case when the objective function is given in the composite form:

$$\phi(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + h(x). \tag{13}$$

In this case NIM uses the following model:

$$m_k(x) = \frac{1}{n}\sum_{i=1}^{n}\left[f_i(v_k^i) + \nabla f_i(v_k^i)^\top (x - v_k^i) + \frac{1}{2}(x - v_k^i)^\top \nabla^2 f_i(v_k^i)(x - v_k^i)\right] + h(x),$$

where v_k^i are some points that are updated in iterations (one point at every iteration). Using the prox operator, we can write down the minimum of the model m_k as follows:

$$\bar{x}_k := \operatorname*{argmin}_{x\in\mathbf{R}^d} m_k(x) = \operatorname{prox}_h^{H_k}\left(H_k^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)v_k^i - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(v_k^i)\right)\right),$$

where $H_k := \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)$. We assume that the subproblem $\bar{x}_k = \operatorname{argmin}_x m_k(x)$ may be solved inexactly, i.e. instead of $\bar{x}_k$ we actually get some $\hat{x}_k$. We now explain which conditions $\hat{x}_k$ must satisfy using the notions from Nesterov (2013).

Let us treat m_k as the composite function: $m_k(x) =: s(x) + h(x)$. Denote

$$T_L(x) := \operatorname*{argmin}_{y\in\mathbf{R}^d}\left[\nabla s(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2 + h(y)\right],$$
$$g_L(x) := L(x - T_L(x)).$$

Then we require that $\hat{x}_k = T_L(y_k)$ with y_k satisfying

$$\|g_L(y_k)\| \leq \left\|\frac{1}{n}\sum_{i=1}^{n} v_k^i - \operatorname{prox}_h\left(\frac{1}{n}\sum_{i=1}^{n} v_k^i - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(v_k^i)\right)\right\|^{1+\gamma}, \tag{14}$$

where $\gamma \in (0, 1]$ is some constant and L is any number such that $L \geq L_0 \equiv 1$.

4.1 Theorem statement

In what follows we prove the following theorem on the local convergence rate of NIM for composite functions.

Theorem 2 (local convergence rate)**.** *Suppose the Hessians $\nabla^2 f_i$ are Lipschitz-continuous:*

$$\left\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\right\| \le M_f \|x - y\|, \qquad i = 1, \dots, n,$$

for all $x, y \in \mathbf{R}^d$. Let $\{x_k\}_{k \ge n}$ be the sequence of iterates generated by NIM with the unit step size $\alpha_k \equiv 1$ and cyclic order of component selection. Assume x^* is a minimizer of (13) *with positive definite Hessian:*

$$\nabla^2 f(x^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 f_i(x^*) \ge \mu_f I, \qquad \mu_f > 0,$$

and all the initial points $x_0, \dots, x_{n-1}$ are close enough to x^*:

$$\|x_i - x^*\| \le R, \qquad i = 0, \dots, n-1. \tag{15}$$

Then the sequence $\{x_k\}_{k \ge 0}$ converges to x^* at an R-superlinear rate, i. e. there exists $\{z_k\}_{k \ge 0}$ such that

$$\begin{aligned} \|x_k - x^*\| &\le z_k, & k &\ge 0 \\ z_{k+1} &\le q_k z_k, & k &\ge n, \end{aligned}$$

where $q_k \to 0$ and $z_0 = \dots = z_{n-1} := \max_{0 \le i \le n-1} \|x_i - x^*\|$.

If the subproblem is solved exactly, then

$$R := \frac{\mu_f}{2M_f}, \qquad q_k := \left(1 - \frac{3}{4n}\right)^{2^{\lfloor k/n \rfloor - 1}}.$$

Otherwise, if it is solved inexactly using the termination condition (14)*, then*

$$R := \min\left\{ \frac{\mu_f}{2M_f}, \left(\frac{\mu_f^3}{128(2 + L_f)^{5+2\gamma}}\right)^{1/(2\gamma)} \right\}, \qquad q_k := \left(1 - \frac{7}{16n}\right)^{(1+\gamma)^{\lfloor k/n \rfloor}/2},$$

where L_f is the Lipschitz constant of ∇f_i:

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L_f \|x - y\|, \qquad i = 1, \dots, n.$$

for all $x, y \in \mathbf{R}^d$.

4.2 Main estimate

When considering local convergence, we assume that NIM uses the unit step length $\alpha \equiv 1$ at every iteration and the order of updating the points v_k^i is cyclic:

$$\begin{aligned} x_{k+1} &= \hat{x}_k, \\ v_{k+1}^i &= \begin{cases} x_{k+1} & \text{if } i = k \bmod n + 1 \\ v_k^i & \text{otherwise.} \end{cases} \end{aligned}$$

Note that the cyclic order of updating means the points $v_k^i,\ i = 1, \dots, n$, are exactly the last n iterates $x_k, x_{k-1}, \dots, x_{k-n+1}$ (but possibly in different order).

Lemma 8. *Let x^* be a minimizer of* (13)*. Then the stopping criterion* (14) *for solving the subproblem guarantees the following bound for $\|e_k\| = \|\hat{x}_k - \operatorname{argmin}_x m_k(x)\|$:*

$$\|e_k\| \le \lambda_{\min}^{-1}(H_k)(2 + L_f)^{2+\gamma} \left(\frac{1}{n} \sum_{i=1}^{n} \left\|v_k^i - x^*\right\|^2\right)^{(1+\gamma)/2}.$$

Proof. The function m_k is strongly convex with constant $\lambda_{\min}(H_k)$ and the gradient ∇s of its smooth (quadratic) part is Lipschitz continuous with constant L_f. Therefore, by Lemma 3 from Nesterov (2013), we have

$$\|\hat{x}_k - \bar{x}_k\| \leq \lambda_{\min}^{-1}(H_k)\left(1 + \frac{L_f}{L}\right)\|g_L(y)\|.$$

Recall that the constants L in (14) satisfy $L \geq L_0 \equiv 1$ (by construction). Thus,

$$\|\hat{x}_k - \bar{x}_k\| \leq \lambda_{\min}^{-1}(H_k)(1 + L_f)\|g_L(y)\|.$$

To finish the proof, it remains to apply inequality (14) together with Lemma 2. □

Lemma 9 (main estimate)**.** *Let $k \geq n-1$ be the number of the current iteration. Assume the last n points $x_k, \ldots, x_{k-n+1}$ are close enough to x^*,*

$$\|x_{k-i} - x^*\| \leq \frac{\mu_f}{2M_f}, \qquad i = 0, \ldots, n-1. \tag{16}$$

Then for the next generated point x_{k+1}, the following bound holds:

$$\|x_{k+1} - x^*\| \leq \frac{M_f}{\mu_f}\left(\frac{1}{n}\sum_{i=0}^{n-1}\|x_{k-i} - x^*\|^2\right) + E\left(\frac{1}{n}\sum_{i=0}^{n-1}\|x_{k-i} - x^*\|^2\right)^{(1+\gamma)/2}, \tag{17}$$

where $E = 0$ when the subproblem is solved exactly and $E = \sqrt{\frac{8(2+L_f)^{5+2\gamma}}{\mu_f^3}}$ when it is solved using FGM with stopping criterion (14)*.*

Proof. Denote $H_k := (1/n)\sum_{i=1}^n \nabla^2 f_i(v_k^i)$.

According to the iteration of NIM, we have: $x_{k+1} = \operatorname{argmin}_x m_k(x) + e_k$, where e_k corresponds to the error in solving the subproblem. Using the definition of the scaled proximal mapping, we can rewrite this as follows:

$$x_{k+1} = \operatorname{prox}_h^{H_k}\left(H_k^{-1}\left(\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(v_k^i)v_k^i - \frac{1}{n}\sum_{i=1}^n \nabla f_i(v_k^i)\right)\right) + e_k.$$

Since x^* is a solution of (13), we have $x^* = \operatorname{prox}_h^{H_k}(x^* - (1/n)\sum_{i=1}^n \nabla f_i(x^*))$. Using this equality and the non-expansiveness property of $\operatorname{prox}_h^{H_k}(\cdot)$, we get

$$\begin{aligned}\|x_{k+1} - x^*\|_{H_k} &\leq \left\|H_k^{-1}\left(\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(v_k^i)[v_k^i - x^*] - \frac{1}{n}\sum_{i=1}^n [\nabla f_i(v_k^i) - \nabla f_i(x^*)]\right)\right\|_{H_k} + \|e_k\|_{H_k} \\ &= \left\|\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(v_k^i)[v_k^i - x^*] - \frac{1}{n}\sum_{i=1}^n [\nabla f_i(v_k^i) - \nabla f_i(x^*)]\right\|_{H_k^{-1}} + \|e_k\|_{H_k}.\end{aligned}$$

Using the bounds $\lambda_{\min}^{1/2}(B)\|w\| \leq \|w\|_B \leq \lambda_{\max}^{1/2}(B)\|w\|$ in the previous inequality, we obtain

$$\|x_{k+1} - x^*\| \leq \frac{1}{\lambda_{\min}(H_k)}\left\|\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(v_k^i)(v_k^i - x^*) - \frac{1}{n}\sum_{i=1}^n [\nabla f_i(v_k^i) - \nabla f_i(x^*)]\right\| + \sqrt{\frac{\lambda_{\max}(H_k)}{\lambda_{\min}(H_k)}}\|e_k\|.$$

To bound the first term, we use the Taylor formula for gradients and Lipschitz-continuity of $\nabla^2 f_i$:

$$\begin{aligned}&\frac{1}{\lambda_{\min}(H_k)}\left\|\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(v_k^i)(v_k^i - x^*) - \frac{1}{n}\sum_{i=1}^n [\nabla f_i(v_k^i) - \nabla f_i(x^*)]\right\| \\ &\qquad = \frac{1}{\lambda_{\min}(H_k)}\left\|\frac{1}{n}\sum_{i=1}^n \int_0^1 [\nabla^2 f_i(v_k^i) - \nabla^2 f_i(v_k^i + \tau(x^* - v_k^i))](v_k^i - x^*)d\tau\right\| \\ &\qquad \leq \frac{M_f}{2\lambda_{\min}(H_k)}\left(\frac{1}{n}\sum_{i=1}^n \left\|v_k^i - x^*\right\|^2\right).\end{aligned}$$

To bound the second term, we use Lemma 8:

$$\sqrt{\frac{\lambda_{\max}(H_k)}{\lambda_{\min}(H_k)}}\,\|e_k\| \le \frac{(2+L_f)^{2+\gamma}}{\lambda_{\min}(H_k)}\sqrt{\frac{\lambda_{\max}(H_k)}{\lambda_{\min}(H_k)}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|v_k^i - x^*\right\|^2\right)^{(1+\gamma)/2}.$$

Thus,

$$\|x_{k+1}-x^*\| \le \frac{M_f}{2\lambda_{\min}(H_k)}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|v_k^i - x^*\right\|^2\right) + \frac{(2+L_f)^{2+\gamma}}{\lambda_{\min}(H_k)}\sqrt{\frac{\lambda_{\max}(H_k)}{\lambda_{\min}(H_k)}}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|v_k^i - x^*\right\|^2\right)^{(1+\gamma)/2}. \tag{18}$$

Let us estimate the eigenvalues of H_k. Since each function f_i has Lipschitz-continuous gradient with constant L_f, each Hessian $\nabla^2 f_i(v_k^i)$ is bounded above by $L_f I$, i.e. $\lambda_{\max}(H_k) \le L_f$. To find the lower bound for $\lambda_{\min}(H_k)$, we first bound the difference between H_k and $\nabla^2 f(x^*)$:

$$\left\|H_k - \nabla^2 f(x^*)\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n}[\nabla^2 f_i(v_k^i) - \nabla^2 f_i(x^*)]\right\| \le \frac{M_f}{n}\sum_{i=1}^{n}\left\|v_k^i - x^*\right\|.$$

Since the order of component selection is cyclic, the points $v_k^i,\ i=1,\dots,n$, are exactly the last n iterates $x_{k-i},\ i=0,\dots,n-1$, (possibly rearranged). Therefore,

$$\left\|H_k - \nabla^2 f(x^*)\right\| = \frac{M_f}{n}\sum_{i=0}^{n-1}\|x_{k-i}-x^*\| \le \frac{\mu_f}{2},$$

where the inequality follows from (16). Using this bound together with (**??**), we get

$$\lambda_{\min}(H_k) \ge \lambda_{\min}(\nabla^2 f(x^*)) - \left\|H_k - \nabla^2 f(x^*)\right\| \ge \frac{\mu_f}{2}.$$

Thus, $\lambda_{\min}(H_k) \ge \mu_f/2$ and $\lambda_{\max}(H_k) \le L_f$. Applying this in (18) and replacing the sums involving v_k^i with the sums involving x_{k-i}, we get (17). □

4.3 Convergence rate of the sequence

Let us investigate the convergence properties of the following (recurrent) sequence arising in (17):

$$z_k := A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right) + E\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right)^{(1+\gamma)/2}, \qquad k \ge n, \tag{19}$$

where $A > 0$, $E \ge 0$, $0 < \gamma \le 1$ are some constants.

First, let us understand the conditions under which the sequence $\{z_k\}_{k\ge 0}$ (monotonically) converges to zero. Note that from (19) it follows that

$$z_k \le A\max_{1\le i\le n}\{z_{k-i}\}^2 + E\max_{1\le i\le n}\{z_{k-i}\}^{1+\gamma} = \left(A\max_{1\le i\le n}\{z_{k-i}\} + E\max_{1\le i\le n}\{z_{k-i}\}^{\gamma}\right)\max_{1\le i\le n}\{z_{k-i}\},$$

and the equality may hold when $z_{k-1} = \cdots = z_{k-n}$. Therefore to guarantee the (monotonic) convergence of $\{z_k\}_{k\ge 0}$ to zero, we must enforce the following condition on the initial elements $z_0,\dots,z_{n-1}$:

$$Az_i + Ez_i^{\gamma} < 1, \qquad i = 0,\dots,n-1.$$

In particular, we may require that

$$\begin{aligned} &\text{(when } E=0\text{):} \quad && z_i \le \frac{1}{2A}, && i=0,\dots,n-1.\\ &\text{(when } E>0\text{):} \quad && z_i \le \min\left\{\frac{1}{2A},\ \left(\frac{1}{4E}\right)^{1/\gamma}\right\}, && i=0,\dots,n-1. \end{aligned} \tag{20}$$

This will guarantee that

$$\begin{aligned} &\text{(when } E=0\text{):} \qquad z_k \le \frac{1}{2}\max_{1\le i\le n}\{z_{k-i}\}, \qquad k\ge n. \\ &\text{(when } E>0\text{):} \qquad z_k \le \frac{3}{4}\max_{1\le i\le n}\{z_{k-i}\}, \qquad k\ge n. \end{aligned} \tag{21}$$

In what follows we always assume that the initial elements of $\{z_k\}_{k\ge 0}$ satisfy (20).

In view of (21), the sequence $\{z_k\}_{k\ge 0}$ converges to zero. However, this convergence may be non-monotonic. For example, if $z_0 > 0$ (but small enough), and $z_1 = \cdots = z_{n-1} = 0$, then $z_n > z_{n-1}$. Nevertheless, it turns out that if the initial elements of $\{z_k\}_{k\ge 0}$ are initialized with the same number, $z_0 = \cdots = z_{n-1}$, then the sequence $\{z_k\}_{k\ge 0}$ is monotonic.

Lemma 10 (monotonicity)**.** *Let $z_0 = \cdots = z_{n-1}$. Then $\{z_k\}_{k\ge 0}$ is monotonic: $z_{k+1} \le z_k$ for all $k \ge 0$.*

Proof. According to (21) we have $z_n \le z_{n-1}$. Thus, we know that $z_0 = \cdots = z_{n-1} \ge z_n$.

We proceed by induction. Suppose we know that $z_0 \ge \cdots \ge z_k$ for some $k \ge n$. We will prove that this implies $z_k \ge z_{k+1}$. Indeed, according to the induction hypothesis, $z_k \le z_{k-n}$. Therefore,

$$\sum_{i=1}^{n} z_{k+1-i}^2 = z_k^2 + \sum_{i=1}^{n-1} z_{k-i}^2 \le z_{k-n}^2 + \sum_{i=1}^{n-1} z_{k-i}^2 = \sum_{i=1}^{n} z_{k-i}^2.$$

Using this and the definition (19) of $\{z_k\}_{k\ge n}$, we have

$$\begin{aligned} z_{k+1} &= A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right) + E\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right)^{(1+\gamma)/2} \\ &\le A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right) + E\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right)^{(1+\gamma)/2} = z_k. \end{aligned}$$

□

From now on, in addition to (20), we assume that $z_0 = \cdots = z_{n-1}$. Due to the monotonicity of $\{z_k\}_{k\ge 0}$, inequality (21) now becomes

$$\begin{aligned} &\text{(when } E=0\text{):} \qquad z_k \le \frac{1}{2} z_{k-n}, \qquad k\ge n. \\ &\text{(when } E>0\text{):} \qquad z_k \le \frac{3}{4} z_{k-n}, \qquad k\ge n. \end{aligned} \tag{22}$$

Using the monotonicity, we can prove that the convergence rate of $\{z_k\}_{k\ge 0}$ is at least linear.

Lemma 11 (linear convergence)**.** *The convergence rate of $\{z_k\}_{k\ge 0}$ is at least linear:*

$$\begin{aligned} &\textit{(when } E=0\textit{):} \qquad z_{k+1} \le \left(1-\frac{3}{4n}\right) z_k, \qquad k\ge n. \\ &\textit{(when } E>0\textit{):} \qquad z_{k+1} \le \left(1-\frac{7}{16n}\right)^{(1+\gamma)/2} z_k, \qquad k\ge n. \end{aligned} \tag{23}$$

Proof. Let us consider the case $E > 0$. The proof of the other case is similar.

Note that

$$\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2 = \frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2 - \frac{1}{n}[z_{k-n}^2 - z_k^2] = \left(1 - \frac{z_{k-n}^2 - z_k^2}{\sum_{i=1}^{n} z_{k-i}^2}\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right).$$

Let us find a lower bound for the fraction. Using $\sum_{i=1}^{n} z_{k-i}^2 \le n z_{k-n}^2$ and (22), we have

$$\frac{z_{k-n}^2 - z_k^2}{\sum_{i=1}^{n} z_{k-i}^2} \ge \frac{1 - 9/16}{n} = \frac{7}{16n}.$$

Thus, we have proved:

$$\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2 \le \left(1 - \frac{7}{16n}\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right).$$

Using this inequality and the definition (19) of $\{z_k\}_{k\ge n}$, we obtain

$$\begin{aligned} z_{k+1} &= A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right) + E\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right)^{(1+\gamma)/2} \\ &\le \left(1 - \frac{7}{16n}\right) A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right) + \left(1 - \frac{7}{16n}\right)^{(1+\gamma)/2} E\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right)^{(1+\gamma)/2} \\ &\le \left(1 - \frac{7}{16n}\right)^{(1+\gamma)/2} z_k. \end{aligned}$$

□

The next lemma shows that the convergence constant in (23) improves after every n iterations.

Lemma 12 (improving the constant)**.** *Suppose that, starting from number $k_0 \ge n$, the sequence $\{z_k\}_{k\ge 0}$ converges linearly to zero with constant q_0:*

$$z_{k+1} \le q_0 z_k, \qquad k \ge k_0. \tag{24}$$

Then, starting from number $k_0 + n$, the constant q_0 can be replaced with a smaller constant:

$$\begin{aligned} &\textit{(when } E = 0\textit{):} \qquad z_{k+1} \le q_0^2 z_k, \qquad k \ge k_0 + n. \\ &\textit{(when } E > 0\textit{):} \qquad z_{k+1} \le q_0^{1+\gamma} z_k, \qquad k \ge k_0 + n. \end{aligned}$$

Proof. Again, we consider only the case $E > 0$. The other case is similar.

Let $k \ge k_0 + n$. Using the definition (19) of $\{z_k\}_{k\ge n}$ and bound (24), we have the following chain of inequalities:

$$\begin{aligned} z_{k+1} &= A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right) + E\left(\frac{1}{n}\sum_{i=1}^{n} z_{k+1-i}^2\right)^{(1+\gamma)/2} \\ &\le q_0^2 A\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right) + q_0^{1+\gamma} E\left(\frac{1}{n}\sum_{i=1}^{n} z_{k-i}^2\right)^{(1+\gamma)/2} \le q_0^{1+\gamma} z_k. \end{aligned}$$

□

Let us summarize the results we have established.

Lemma 13. *Let $\{z_k\}_{k\ge 0}$ be a recurrent sequence defined in* (19)*. Suppose the initial elements $z_0, \dots, z_{n-1}$ of this sequence are chosen equal to the same number small enough:*

$$z_0 = \cdots = z_{n-1} \le R.$$

Then the sequence $\{z_k\}_{k\ge 0}$ converges monotonically to zero at a Q-superlinear rate:

$$z_{k+1} \le q_k z_k, \qquad k \ge n.$$

The expressions for R and q_k are as follows:

$$\textit{(when } E = 0\textit{):} \qquad R := \frac{1}{2A}, \qquad\qquad q_k := \left(1 - \frac{3}{4n}\right)^{2^{\lfloor k/n \rfloor - 1}}.$$

$$\textit{(when } E > 0\textit{):} \qquad R := \min\left\{\frac{1}{2A}, \left(\frac{1}{4E}\right)^{1/\gamma}\right\}, \qquad q_k := \left(1 - \frac{7}{16n}\right)^{(1+\gamma)^{\lfloor k/n \rfloor}/2}.$$

Proof. Consider the case $E > 0$.

Denote $q := (1 - 7/16n)^{1/2}$. According to (23) and Lemma 12 we can write the following sequence of inequalities:

$$\begin{aligned} z_{k+1} &\le q^{1+\gamma} z_k, && k \ge n, \\ z_{k+1} &\le q^{(1+\gamma)^2} z_k, && k \ge 2n, \\ z_{k+1} &\le q^{(1+\gamma)^3} z_k, && k \ge 3n, \\ &\dots \end{aligned}$$

Combining all these inequalities together, we get

$$z_{k+1} \le q^{(1+\gamma)^{\lfloor k/n \rfloor}} z_k, \qquad k \ge n.$$

□

4.4 Proof of the theorem

Proof. Consider the sequence $\{z_k\}_{k\ge 0}$ defined in (19) with

$$A := \frac{M_f}{\mu_f}, \qquad E := \sqrt{\frac{8(2 + L_f)^{5+2\gamma}}{\mu_f^3}}.$$

Let us set the initial elements $z_0, \dots, z_{n-1}$ of this sequence to the same value:

$$z_0 = \dots = z_{n-1} := \max_{0 \le i \le n-1} \|x_i - x^*\| \le R.$$

According to Lemma 13, the sequence $\{z_k\}_{k\ge 0}$ converges monotonically and Q-superlinearly to zero with constants q_k. In particular, it means that $\{z_k\}_{k\ge 0}$ always stays bounded:

$$z_k \le R \le \frac{\mu_f}{2M_f}, \qquad k \ge 0. \tag{25}$$

Due to the initial condition (15), we can apply Lemma 9 for $k = n - 1$. Since, by construction, the values $\|x_i - x^*\|$ are bounded above by z_i for $i = 0, \dots, n-1$, we have $\|x_n - x^*\| \le z_n$. In view of (25), it means that the new iterate x_n does not leave the R-vicinity of x^*. Therefore, we can apply Lemma 9 again but for $k = n$. Using the same reasoning, we conclude that $\|x_{n+1} - x^*\| \le z_{n+1} \le R$, and so on. Thus, Lemma 9 holds for all $k \ge n - 1$ and the sequence $\{z_k\}_{k\ge 0}$ majorizes $\{\|x_k - x^*\|\}_{k\ge 0}$. □

5 Global rate of convergence

We consider minimizing

$$\phi(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + h(x).$$

In this section we give a proof of the global convergence of NIM in the particular case when $h(x)$ is an ℓ_2-regularizer, $h(x) = \frac{\mu}{2}\|x\|^2$, and there is no inexactness in finding the minimum of the model.

Our proof is based on the work by Gurbuzbalaban et al. (2015). To analyze NIM, we view it as a perturbed scaled gradient method. In the following we use the notation

$$A_k := \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(v_k^i) + \mu I.$$

Also we can rewrite the step of NIM as follows:

$$x_{k+1} = x_k + \alpha p_k, \tag{26}$$

where p_k is the search direction of NIM:

$$\begin{aligned} p_k := \bar{x}_k - x_k &= A_k^{-1}(u_k - g_k - A_k x_k) \\ &= A_k^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(v_k^i)(v_k^i - x_k) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(v_k^i) - \mu x_k\right). \end{aligned}$$

Denote the search direction of the scaled gradient method as

$$p_k^{\mathrm{SG}} := -A_k^{-1}\nabla\phi(x_k).$$

Then (26) can be rewritten as:

$$x_{k+1} = x_k + \alpha p_k^{\mathrm{SG}} + \alpha e_k,$$

where e_k is the error in the approximation of p_k^{SG} by p_k:

$$e_k := p_k - p_k^{\mathrm{SG}}.$$

Lemma 14. *For a twice continuously differentiable strongly convex function ϕ with constant μ and Lipschitz-continuous gradient with constant L we have the following bounds:*

$$\frac{1}{L}\left\|\nabla\phi(x)\right\|_2 \le \left\|x - x^*\right\|_2 \le \frac{1}{\mu}\left\|\nabla\phi(x)\right\|_2$$

and

$$\frac{1}{2L}\left\|\nabla\phi(x)\right\|_2^2 \le \phi(x) - \phi(x^*) \le \frac{1}{2\mu}\left\|\nabla\phi(x)\right\|_2^2.$$

where x^* is the optimum of ϕ.

In what follows we assume that $n \ge 2$.

5.1 Bounding the norm of the error

Lemma 15.

$$\|e_k\|_2 \le \frac{2L}{\mu} \max_{j=k-n+1,\ldots,k-1} \|x_j - x_k\|_2$$

and

$$\|e_k\|_2 \le \frac{4L}{\mu^2} \max_{j=k-n+1,\ldots,k} \|\nabla\phi(x_j)\|_2$$

Proof.

$$e_k = A_k^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(v_k^i)(v_k^i - x_k) - \frac{1}{n}\sum_{i=1}^{n}\left[\nabla f_i(v_k^i) - \nabla f_i(x_k)\right]\right).$$

Taking the norms, we get

$$\|e_k\|_2 \le \left\|A_k^{-1}\right\|_2 \left(\frac{1}{n}\sum_{i=1}^{n} \left\|\nabla^2 f_i(v_k^i)\right\|_2 \left\|v_k^i - x_k\right\|_2 + \frac{1}{n}\sum_{i=1}^{n} \left\|\nabla f_i(v_k^i) - \nabla f_i(x_k)\right\|_2\right)$$

By the Lipschitz-continuity of ∇f_i, we have that $\left\|\nabla^2 f_i(v_k^i)\right\|_2 \leq L$ and $\left\|\nabla f_i(v_k^i) - \nabla f_i(x_k)\right\|_2 \leq L\left\|v_k^i - x_k\right\|_2$. Also $\left\|A_k^{-1}\right\| \leq (1/\mu)$. So

$$\|e_k\|_2 \leq \frac{2L}{\mu n}\sum_{i=1}^{n}\left\|v_k^i - x_k\right\|_2.$$

Since the order of component selection is cyclic,

$$\sum_{i=1}^{n}\left\|v_k^i - x_k\right\|_2 = \sum_{j=k-n+1}^{k}\|x_j - x_k\|_2.$$

Then

$$\begin{aligned}\|e_k\|_2 &\leq \frac{2L}{\mu n}\sum_{j=k-n+1}^{k}\|x_j - x_k\|_2\\ &\leq \frac{2L}{\mu}\max_{j=k-n+1,\ldots,k}\|x_j - x_k\|_2\\ &= \frac{2L}{\mu}\max_{j=k-n+1,\ldots,k-1}\|x_j - x_k\|_2.\end{aligned}$$

The second inequality follows from the triangle inequality

$$\|x_j - x_k\|_2 \leq \|x_j - x^*\| + \|x_k - x^*\|.$$

□

First, we bound the norm of the error by a term proportional to the step length and the norms of the gradients at the previous points.

Lemma 16.

$$\|e_k\|_2 \leq \frac{2L(4L+\mu)(n-1)\alpha}{\mu^3}\max_{t=k-2n+2,\ldots,k-1}\|\nabla\phi(x_t)\|_2.$$

Proof. Now for any $j = k-n+1,\ldots,k-1$:

$$\|x_j - x_k\|_2 = \left\|\sum_{s=j}^{k-1}[x_s - x_{s+1}]\right\|_2 \leq \sum_{s=j}^{k-1}\|x_s - x_{s+1}\|_2 \leq \sum_{s=k-n+1}^{k-1}\|x_s - x_{s+1}\|_2.$$

For the difference of successive points we have:

$$\|x_s - x_{s+1}\|_2 \leq \alpha\left(\left\|p_s^{\mathrm{SG}}\right\|_2 + \|e_s\|_2\right)$$

For the first term,

$$\left\|p_s^{\mathrm{SG}}\right\|_2 \leq \left\|A_s^{-1}\right\|_2\|\nabla\phi(x_s)\|_2 \leq \frac{1}{\mu}\|\nabla\phi(x_s)\|_2.$$

For the second term we use the previous lemma:

$$\|e_s\|_2 \leq \frac{4L}{\mu^2}\max_{t=s-n+1,\ldots,s-1}\|\nabla\phi(x_t)\|_2.$$

Therefore,

$$\|x_s - x_{s+1}\|_2 \leq \frac{\alpha(4L+\mu)}{\mu^2}\max_{t=s-n+1,\ldots,s}\|\nabla\phi(x_t)\|_2$$

So

$$\begin{aligned}\|x_j - x_k\|_2 &\leq \frac{\alpha(4L+\mu)}{\mu^2}\sum_{s=k-n+1}^{k-1}\max_{t=s-n+1,\ldots,s}\|\nabla\phi(x_t)\|_2\\ &\leq \frac{\alpha(n-1)(4L+\mu)}{\mu^2}\max_{t=k-2n+2,\ldots,k-1}\|\nabla\phi(x_t)\|_2.\end{aligned}$$

Finally,

$$\|e_k\|_2 \leq \frac{2L(4L+\mu)(n-1)\alpha}{\mu^3} \max_{t=k-2n+2,\ldots,k-1} \|\nabla\phi(x_t)\|_2 .$$

□

5.2 Proof of the theorem about global convergence

Note that ϕ has Lipschitz-continuous gradient with constant $L+\mu$.

Proof. By Lipschitz-continuity of the gradient we have

$$\begin{aligned}\phi(x_{k+1}) - \phi(x_k) &\leq \nabla\phi(x_k)^\top (x_{k+1}-x_k) + \frac{L+\mu}{2}\|x_{k+1}-x_k\|_2^2 \\ &= -\alpha\nabla\phi(x_k)^\top A_k^{-1}\nabla\phi(x_k) + \alpha\nabla\phi(x_k)^\top e_k \\ &\quad + \frac{(L+\mu)\alpha^2}{2}\left\|p_k^{\mathrm{SG}}\right\|_2^2 + \frac{(L+\mu)\alpha^2}{2}\|e_k\|_2^2 + (L+\mu)\alpha^2 (p_k^{\mathrm{SG}})^\top e_k\end{aligned}$$

Now we bound each term above in terms of the norms of the previous gradients.

For the first term, using $\lambda_{\max}(A_k) \leq L+\mu$, we have

$$-\alpha\nabla\phi(x_k)^\top A_k^{-1}\nabla\phi(x_k) \leq -\frac{\alpha}{L+\mu}\|\nabla\phi(x_k)\|_2^2 .$$

For the second term we use the Cauchy-Schwarz inequality and bound for $\|e_k\|_2$:

$$\begin{aligned}\alpha\nabla\phi(x_k)^\top e_k &\leq \alpha\|\nabla\phi(x_k)\|_2\|e_k\|_2 \\ &\leq \frac{2L(4L+\mu)(n-1)\alpha^2}{\mu^3} \max_{t=k-2n+2,\ldots,k}\|\nabla\phi(x_t)\|_2^2 .\end{aligned}$$

For the third term we use the bound $\left\|p_k^{\mathrm{SG}}\right\|_2 \leq \left\|A_k^{-1}\right\|_2 \|\nabla\phi(x_k)\|_2$:

$$\frac{(L+\mu)\alpha^2}{2}\left\|p_k^{\mathrm{SG}}\right\|_2^2 \leq \frac{(L+\mu)\alpha^2}{2\mu^2}\|\nabla\phi(x_k)\|_2^2 .$$

For the fourth term we use the lemma:

$$\frac{(L+\mu)\alpha^2}{2}\|e_k\|_2^2 \leq \frac{4L^2(L+\mu)(4L+\mu)^2(n-1)^2\alpha^4}{2\mu^6} \max_{t=k-2n+2,\ldots,k-1}\|\nabla\phi(x_t)\|_2^2 .$$

For the firth term we use the Cauchy-Schwarz inequality and the lemma:

$$\begin{aligned}(L+\mu)\alpha^2(p_k^{\mathrm{SG}})^\top e_k &\leq (L+\mu)\alpha^2\left\|p_k^{\mathrm{SG}}\right\|_2\|e_k\|_2 \\ &\leq \frac{2L(L+\mu)(4L+\mu)(n-1)\alpha^3}{\mu^4} \max_{t=k-2n+2,\ldots,k}\|\nabla\phi(x_t)\|_2^2 .\end{aligned}$$

Combining all these bounds, we get

$$\begin{aligned}\phi(x_{k+1}) - \phi(x_k) &\leq \left(-\frac{\alpha}{L+\mu} + \frac{(L+\mu)\alpha^2}{2\mu^2}\right)\|\nabla\phi(x_k)\|_2^2 \\ &\quad + \frac{2L(4L+\mu)(n-1)}{\mu^3}\left(\alpha^2 + \frac{(L+\mu)\alpha^3}{\mu} + \frac{2L(L+\mu)(4L+\mu)(n-1)\alpha^4}{2\mu^3}\right) \max_{t=k-2n+2,\ldots,k}\|\nabla\phi(x_t)\|_2^2 .\end{aligned}$$

Now bound the gradients in terms of function values:

$$\|\nabla\phi(x_k)\|_2^2 \geq 2\mu\left[\phi(x_k)-\phi(x^*)\right] \qquad \text{and} \qquad \max_{t=k-2n+2,\ldots,k}\|\nabla\phi(x_t)\|_2 \leq 2(L+\mu)\max_{t=k-2n+2}\left[\phi(x_k)-\phi(x^*)\right]$$

For $\alpha \leq 2(\mu/(L+\mu))^2$ the coefficient before $\|\nabla\phi(x_k)\|_2$ is non-positive.

Denote $V_k := \phi(x_k) - \phi(x^*)$. Then we have proved the following bound:

$$V_{k+1} \leq p(\alpha)V_k + q(\alpha) \max_{t=k-2n+2,\ldots,k} V_t$$

where

$$p(\alpha) := 1 - \frac{\mu\alpha}{L+\mu} + \frac{(L+\mu)\alpha^2}{2\mu},$$
$$q(\alpha) := \frac{2L(L+\mu)(4L+\mu)(n-1)}{\mu^3}\left(\alpha^2 + \frac{(L+\mu)\alpha^3}{\mu} + \frac{2L(L+\mu)(4L+\mu)(n-1)\alpha^4}{2\mu^3}\right).$$

Using lemma 3.2 of Gurbuzbalaban et al. (2015) we have that for α small enough $p(\alpha)+q(\alpha) < 1$ and that V_k converges linearly with constant $c := (p+q)^{1/(1+2(n-1))}$.

□

6 Order of component selection

There are two standard strategies in NIM for choosing the component i_k to update: 1) *cyclic* when $i_k = (k \bmod n)+1$, and 2) *randomized* when at every iteration $i_k \in \{1,\ldots,n\}$ is chosen uniformly at random. In all our experiments we observed that NIM always converges faster under the cyclic order. Here we analyse this situation.

Let's consider a particular function $\phi(x)$ for optimization using NIM:

$$\phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{2}\|x\|^2 + \frac{1}{3}\|x\|^3, \tag{27}$$
$$f_1(x) = \frac{1}{2}\|x\|^2 + \frac{n}{3}\|x\|^3,\ f_i(x) = \frac{1}{2}\|x\|^2, i > 1.$$

Lemma 17. *Using NIM with unit step size and exact model minimization for function* (27) *leads to the following iterate:*

$$x_{k+1} = \frac{\|v_k^1\|}{1+2\|v_k\|}v_k^1.$$

Proof. Using unit step size means that the next iterate x_{k+1} in NIM coincides with exact minimum of the model $m_k(x)$. This minimum can be found as follows:

$$x_{k+1} = H_k^{-1}(u_k - g_k). \tag{28}$$

Here all the values H_k, u_k, g_k for the function (27) can be calculated directly:

$$H_k = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i) = I + \|v_k^1\|I + \frac{(v_k^1)(v_k^1)^T}{\|v_k^1\|},$$
$$g_k = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(v_k^i) = \frac{1}{n}\sum_{i=1}^{n} v_k^i + \|v_k^1\|v_k^1,$$
$$u_k = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)v_k^i = \frac{1}{n}\sum_{i=1}^{n} v_k^i + 2\|v_k^1\|v_k^1.$$

Substituting these expressions into (28), we obtain:

$$H_k x_{k+1} = \left(I + \|v_k^1\|I + \frac{(v_k^1)(v_k^1)^T}{\|v_k^1\|}\right)x_{k+1} = (1+\|v_k^1\|)x_{k+1} + \frac{(v_k^1)^T x_{k+1}}{\|v_k^1\|}v_k^1 = u_k - g_k = \|v_k^1\|v_k^1. \tag{29}$$

This result means that x_{k+1} can be represented as αv_k^1 for some scalar α. Substituting αv_k^1 instead of x_{k+1} into (29) and finding α, we obtain:

$$x_{k+1} = \frac{\|v_k^1\|}{1 + 2\|v_k\|} v_k^1.$$

□

Now we are ready to prove the main theorem:

Theorem 3. *Using NIM with randomized order for the function* (27) *with center initialization* $\|v_k^1\| < 1$ *leads to the following lower bound:*

$$\mathbf{E}[\|x_{k+1}\|] \geq \frac{1}{3}\left(1 - \frac{1}{n}\right)^k \mathbf{E}[\|v_0^1\|^2].$$

Proof. From the last lemma and using condition $\|v_k^1\| < 1$ we have:

$$\|x_{k+1}\| = \frac{\|v_k^1\|^2}{1 + 2\|v_k\|} \geq \frac{\|v_k^1\|^2}{3}. \tag{30}$$

Let's denote $\xi_k := \mathbf{E}[\|x_{k+1}\|^2]$ and $\delta_k := \mathbf{E}[\|v_k^1\|^2]$, where expectation is taken w.r.t. all components selections at all iterations. In NIM we have the following update rule for the next center:

$$v_{k+1}^i = x_{k+1} I[i = i_k] + v_k^i I[i \neq i_k],$$

where $I[\cdot]$ is indicator function. Now we can obtain recalculation formula for δ_k:

$$\delta_{k+1} = \mathbf{E}[\|v_{k+1}\|^2] = \frac{1}{n}\mathbf{E}[\|x_{k+1}\|^2] + \left(1 - \frac{1}{n}\right)\mathbf{E}[\|v_k^1\|^2] = (1-q)\xi_{k+1} + q\delta_k,$$

where $q = 1 - 1/n$. Also $\delta_0 = \|v_0^1\|^2$. Using this recalculation formula several times, we obtain:

$$\begin{aligned}\delta_k = (1-q)\xi_k + q\delta_{k-1} = (1-q)\xi_k + q((1-q)\xi_{k-1} + q\delta_{k-2}) = \\ = \underbrace{(1-q)\xi_k}_{\geq 0} + \underbrace{q(1-q)\xi_{k-1}}_{\geq 0} + \underbrace{q^2(1-q)\xi_{k-2}}_{\geq 0} + \cdots + \underbrace{q^{k-1}(1-q)\xi_1}_{\geq 0} + q^k\delta_0 \geq q^k\delta_0.\end{aligned}$$

Using (30) we come to the theorem statement:

$$\mathbf{E}[\|x_{k+1}\|] \geq \frac{1}{3}\delta_k \geq \frac{1}{3}q^k\delta_0.$$

□

This theorem proves that NIM with random order can at best have a linear convergence rate. Meanwhile using NIM with cyclic order for the function (27) gives the following upper bound:

$$\|x_{k+1}\| \leq \|v_k^1\|^2,$$

that is equivalent to Q-quadratic convergence w.r.t epochs and R-superlinear convergence w.r.t. iterations.

References

Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P. On the convergence rate of incremental aggregated gradient algorithms. *arXiv preprint arXiv:1506.02081*, 2015.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming, V.140(1)*, pp. 125–161, 2013.