Proceedings of the International Conference on
Nonlinear Analysis and Convex Analysis &
International Conference on Optimization:
Techniques and Applications -II-
(Hakodate, Japan, 2019), 99–108

GLOBAL CONVERGENCE OF A PROXIMAL MEMORYLESS SYMMETRIC RANK ONE METHOD FOR MINIMIZING COMPOSITE FUNCTIONS

SHUMMIN NAKAYAMA AND YASUSHI NARUSHIMA

Abstract. In this paper, we treat a minimization problem of which the objective function is a composite function of a differentiable function (possibly nonconvex), and a convex function (possibly nonsmooth). For solving such a problem, proximal gradient methods are widely used. To accelerate proximal gradient methods, some researchers have proposed proximal gradient methods based on proximal mappings scaled by quasi-Newton matrices. Although it is usually difficult to compute the scaled proximal mapping directly, it can be easily obtained when quasi-Newton matrices are given by the memoryless symmetric rank-one (SR1) update formula. Thus, in this paper, we focus on a proximal quasi-Newton method based on the memoryless SR1 formula. To establish the global convergence of the method, we propose a proximal quasi-Newton method based on the memoryless SR1 formula with a modified spectral-scaling secant condition. We show the global convergence of the proposed method. Finally, we report some preliminary numerical results.

1. Introduction

In this paper, we consider minimization of the following composite function:

$$\min_{x\in\mathbb{R}^n} \quad f(x) := g(x) + h(x), \tag{1.1}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function, and $h : \mathbb{R}^n \to \mathbb{R}$ is a continuous convex function. Here, we note that g is possibly nonconvex and h is possibly nonsmooth. To solve (1.1), the proximal gradient methods are widely used.

2010 *Mathematics Subject Classification.* 90C30, 90C53, 90C06.

Key words and phrases. Nonsmooth optimization, proximal Newton method, memoryless quasi-Newton method, symmetric rank one formula, global convergence.

The usual proximal gradient method is an iterative method of the form:

$$x_{k+1} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left(g(x_k) + \nabla g(x_k)^T (x - x_k) + \frac{1}{2t_k}\|x - x_k\|^2 + h(x) \right) \tag{1.2}$$
$$= \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left(h(x) + \frac{1}{2t_k} \|x - (x_k - t_k \nabla g(x_k))\|^2 \right),$$

where $x_k \in \mathbb{R}^n$ is k-th approximation to a solution, $\nabla g(x_k)$ is the gradient of g at x_k, $\|\cdot\|$ denotes the ℓ_2 norm and $t_k > 0$ is a parameter. By defining the proximal mapping as

$$\mathrm{Prox}_h(\bar{x}) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left(h(x) + \frac{1}{2} \|x - \bar{x}\|^2 \right), \tag{1.3}$$

(1.2) can be represented by

$$x_{k+1} = \mathrm{Prox}_{t_k h} \left(x_k - t_k \nabla g(x_k) \right).$$

Note that the proximal mapping (1.3) can be easily obtained in some special cases (see [1, 11], for example).

To accelerate proximal gradient methods, some researchers proposed proximal quasi-Newton methods which use proximal mappings scaled by quasi-Newton matrices. Lee et al. [7] gave a proximal Newton type method of the form:

$$x_{k+1} = x_k + \alpha_k d_k \tag{1.4}$$

and

$$d_k = \mathrm{Prox}_h^{B_k} \left(x_k - H_k \nabla g(x_k) \right) - x_k, \tag{1.5}$$

where $\alpha_k > 0$ is the step size and $\mathrm{Prox}_h^{B_k}(\bar{x})$ is a scaled proximal mapping given by

$$\mathrm{Prox}_h^{B_k}(\bar{x}) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left(g(\bar{x}) + \nabla g(\bar{x})^T (x - \bar{x}) + \frac{1}{2}\|x - \bar{x}\|_{B_k}^2 + h(x) \right) \tag{1.6}$$
$$= \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left(h(x) + \frac{1}{2}\|x - \bar{x}\|_{B_k}^2 \right).$$

Here, B_k is a symmetric positive definite matrix, $\|x\|_{B_k} = \sqrt{x^T B_k x_k}$ and $H_k = B_k^{-1}$.

Following [7], we now consider a prototype algorithm of the proximal Newton type method.

Algorithm 1 (A prototype algorithm of proximal Newton type methods)**.**

Step 0: Given an initial point $x_0 \in \mathbb{R}^n$, and parameters $\delta \in (0, 1)$, $\beta \in (0, 1)$ and $k = 0$.

Step 1: Give B_k and H_k. Compute d_k by (1.5).
Step 2: If the stopping condition holds, then stop.
Step 3: Determine the step size α_k, which is the first number of the sequence $\alpha \in \{1, \beta, \beta^2, \ldots\}$ satisfying the Armijo condition:

$$f(x_k + \alpha d_k) \leq f(x_k) + \delta\alpha(\nabla g(x_k)^T d_k + h(x_k + d_k) - h(x_k)).$$

Step 4: Update x_{k+1} by (1.4).
Step 5: Set $k = k + 1$, and go to Step 1.

In general, to compute (1.5), we need to solve (1.6) inexactly by using some numerical method. Accordingly, some researchers have proposed inexact proximal Newton type methods, which solve (1.6) inexactly [7, 10]. As another approach, Becker and Fadlli [3] gave a calculation procedure of (1.6) for the case $B_k = D + uu^T$, where D is a positive diagonal matrix and $u \in \mathbb{R}^n$.

Theorem 1.1. *Let $V = D \pm uu^T$ be symmetric positive definite, where D is diagonal with positive diagonal elements, and $u \in \mathbb{R}^n$. Then,*

$$\mathrm{Prox}_h^V(\bar{x}) = D^{-1/2}\mathrm{Prox}_{\bar{h}}(D^{1/2}\bar{x} \mp \alpha D^{-1/2}u),$$

where $\bar{h}(x) := h(D^{-1/2}x)$ and α is the unique root of

$$p(\alpha) = \left\langle u, \bar{x} - D^{-1/2}\mathrm{Prox}_{\bar{h}}(D^{1/2}\bar{x} \mp \alpha D^{-1/2}u) \right\rangle + \alpha.$$

By using the above theorem, $\mathrm{Prox}_h^V(\bar{x})$ can be easily compute when we can easily obtain the usual proximal mapping. Becker and Fadili [3] gave a proximal quasi-Newton method based on the memoryless Symmetric Rank-one (SR1) method. However, the memoryless SR1 updating formula does not guarantee positive definiteness of approximate matrices. Therefore, in this paper, we propose another proximal quasi-Newton method based on a modified SR1 method, which guarantees positive definiteness of approximate matrices.

This paper is organized as follows. In Section 2, we propose an proximal Newton type method based on the memoryless SR1 method. In addition, we give the global convergence property of the method. In Section 3, we report some preliminary numerical results. Finally, we give concluding remarks.

2. Proximal memoryless symmetric rank one method

In this section, we propose a proximal Newton type method based on the SR1 formula, and give its global convergence property.

We first consider a concrete choice of B_k in (1.5). We focus on the memoryless SR1 method proposed by Nakayama et al. [9], which is the quasi-Newton

method with the following SR1 formula:

$$B_k = I + \frac{(\gamma_k y_{k-1} - s_{k-1})(\gamma_k y_{k-1} - s_{k-1})^T}{(\gamma_k y_{k-1} - s_{k-1})^T s_{k-1}}, \tag{2.1}$$

where $s_{k-1} = x_k - x_{k-1}$, $y_{k-1} = \nabla g(x_k) - \nabla g(x_{k-1})$, $I \in \mathbb{R}^{n\times n}$ is the identity matrix and $\gamma_k > 0$ is a scaling parameter. This formula is based on the spectral scaling secant condition by Cheng and Li [5]:

$$B_k s_{k-1} = \gamma_k y_{k-1}. \tag{2.2}$$

Under the assumption $s_{k-1}^T y_{k-1} > 0$, Nakayama et al. [9] showed B_k in (2.1) is symmetric positive definite if and only if γ_k satisfies

$$\gamma_k \notin \left[\frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}, \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \right].$$

Thus, using by Theorem 1.1, the proximal mapping with (2.1) can be easily computed. On the other hand, to establish the global convergence of the method, we need uniformly positive definiteness and boundedness for B_k. For this purpose, we modify the SR1 formula (2.1). Following Nakayama et al. [10], we consider the modified spectral scaling secant condition:

$$B_k s_{k-1} = \gamma_k z_{k-1}, \tag{2.3}$$

which is combined the spectral scaling secant condition (2.2) and the modified secant condition by Li and Fukushima [8]. Here,

$$z_{k-1} = y_{k-1} + \nu_k s_{k-1} \tag{2.4}$$

and $\nu_k \geq 0$ is a bounded parameter. In this paper, we choose ν_k such that

$$s_{k-1}^T z_{k-1} = s_{k-1}^T (y_{k-1} + \nu_k s_{k-1}) \geq \bar{\nu} \|s_{k-1}\|^2 \tag{2.5}$$

holds for some positive constant $\bar{\nu}$. For example, if we choose

$$\nu_k = \begin{cases} 0, & \text{if } s_{k-1}^T y_{k-1} \geq \bar{\nu} \|s_{k-1}\|^2 \\ \max\left\{0, -\frac{s_{k-1}^T y_{k-1}}{s_{k-1}^T s_{k-1}}\right\} + \bar{\nu}, & \text{otherwise,} \end{cases}$$

then (2.5) holds. Furthermore, we choose γ_k satisfying the condition

$$\underline{\gamma} \leq \gamma_k \leq \overline{\gamma}, \tag{2.6}$$

where $\underline{\gamma}$ and $\overline{\gamma}$ are positive constants such that $0 < \underline{\gamma} \leq \overline{\gamma}$ holds. Based on the modified spectral scaling secant condition (2.3), the SR1 formula (2.1) is represented by

$$B_k = I + \frac{(\gamma_k z_{k-1} - s_{k-1})(\gamma_k z_{k-1} - s_{k-1})^T}{(\gamma_k z_{k-1} - s_{k-1})^T s_{k-1}}. \tag{2.7}$$

Also, the inverse of (2.7) is given by

$$H_k = I + \frac{(s_{k-1} - \gamma_k z_{k-1})(s_{k-1} - \gamma_k z_{k-1})^T}{\gamma_k (s_{k-1} - \gamma_k z_{k-1})^T z_{k-1}}. \tag{2.8}$$

In order to show the uniformly positive definiteness, we make the following standard assumption.

Assumption 1. The function g is a continuously differentiable function and its gradient ∇g is Lipschitz continuous, namely, there exists a positive constant L such that

$$\|\nabla g(u) - \nabla g(v)\| \leq L\|u - v\| \text{ for any } u, v \in \mathbb{R}^n. \tag{2.9}$$

Under the above assumption, we note that it follows from (2.4) and (2.9) that

$$\|z_{k-1}\| = \|y_{k-1} + \nu_k s_{k-1}\| = \|\nabla g(x_k) - \nabla g(x_{k-1}) + \nu_k s_{k-1}\| \leq (L + \nu_k)\|s_{k-1}\|.$$

Since ν_k is bounded, there exists a positive constant $\bar{L}$ such that

$$\|z_{k-1}\| \leq \bar{L}\|s_{k-1}\|. \tag{2.10}$$

Then, we have the following theorem.

Theorem 2.1. *Suppose that* Assumption 1*,* (2.5) *and* (2.6) *are satisfied. The matrix B_k is given by* (2.7) *and the parameter γ_k satisfies*

$$\gamma_k \notin \left(\underline{\rho} \frac{s_{k-1}^T z_{k-1}}{z_{k-1}^T z_{k-1}}, \overline{\rho} \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T z_{k-1}} \right), \tag{2.11}$$

where $0 < \underline{\rho} < 1$ and $\overline{\rho} > 1$ are constants. Then, the following statements hold:

(i) *If we choose γ_k satisfying $\gamma_k \leq \underline{\rho} \dfrac{s_{k-1}^T z_{k-1}}{z_{k-1}^T z_{k-1}}$, then there exists a positive constant $m \leq 1$ such that*

$$m\|v\|^2 \leq v^T B_k v \leq \|v\|^2, \quad \forall v \in \mathbb{R}^n. \tag{2.12}$$

(ii) *If we choose γ_k satisfying $\gamma_k \geq \overline{\rho} \dfrac{s_{k-1}^T s_{k-1}}{s_{k-1}^T z_{k-1}}$, then there exists a positive constant $M \geq 1$ such that*

$$\|v\|^2 \leq v^T B_k v \leq M\|v\|^2, \quad \forall v \in \mathbb{R}^n. \tag{2.13}$$

Proof. First, we consider case (i). The eigenvalues of formula (2.8) are 1 with multiplicity $n - 1$ and

$$1 + \frac{\|s_{k-1} - \gamma_k z_{k-1}\|^2}{\gamma_k (s_{k-1} - \gamma_k z_{k-1})^T z_{k-1}}. \tag{2.14}$$

Since $\gamma_k(s_{k-1}-\gamma_k z_{k-1})^T z_{k-1} > 0$ holds when $\gamma_k \leq \underline{\rho}\frac{s_{k-1}^T z_{k-1}}{z_{k-1}^T z_{k-1}}$, (2.14) is greater than 1. Thus, the smallest eigenvalue of H_k is 1, and the largest eigenvalue of H_k is (2.14). It follows from (2.5), (2.6), (2.10) and (2.14) that

$$\begin{aligned}
1 + \frac{\|s_{k-1} - \gamma_k z_{k-1}\|^2}{\gamma_k(s_{k-1} - \gamma_k z_{k-1})^T z_{k-1}} &\leq 1 + \frac{\|s_{k-1} - \gamma_k z_k\|^2}{\gamma_k(1-\underline{\rho}) s_{k-1}^T z_{k-1}} \\
&\leq 1 + \frac{(1+\overline{\gamma}\bar{L})\|s_{k-1}\|^2}{\underline{\gamma}(1-\underline{\rho})\bar{\nu}\|s_{k-1}\|^2} \\
&= 1 + \frac{1+\overline{\gamma}\bar{L}}{\bar{\nu}\underline{\gamma}(1-\underline{\rho})}.
\end{aligned}$$

Since $B_k = H_k^{-1}$, we have (2.12) with $m = \left(1 + \frac{1+\overline{\gamma}\bar{L}}{\bar{\nu}\underline{\gamma}(1-\underline{\rho})}\right)^{-1}$.

Next, we consider case (ii). The eigenvalues of formula (2.7) are 1 with multiplicity $n-1$ and

$$1 + \frac{\|\gamma_k z_{k-1} - s_{k-1}\|^2}{(\gamma_k z_{k-1} - s_{k-1})^T s_{k-1}}. \tag{2.15}$$

Since $(\gamma_k z_{k-1} - s_{k-1})^T s_{k-1} > 0$ hols when $\gamma_k \geq \overline{\rho}\frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T z_{k-1}}$, (2.15) is greater than 1. Thus, the smallest eigenvalue of B_k is 1, and the largest eigenvalue of B_k is (2.15). It follows from (2.5), (2.6), (2.10) and (2.15) that

$$\begin{aligned}
1 + \frac{\|\gamma_k z_{k-1} - s_{k-1}\|^2}{(\gamma_k z_{k-1} - s_{k-1})^T s_{k-1}} &\leq 1 + \frac{\|\gamma_k z_k - s_{k-1}\|^2}{(\overline{\rho}-1)\|s_{k-1}\|^2} \\
&\leq 1 + \frac{(\overline{\gamma}\bar{L}+1)\|s_{k-1}\|^2}{(\overline{\rho}-1)\|s_{k-1}\|^2} \\
&= 1 + \frac{\overline{\gamma}\bar{L}+1}{\overline{\rho}-1}.
\end{aligned}$$

Therefore, we have (2.13) with $M = 1 + \frac{\overline{\gamma}\bar{L}+1}{\overline{\rho}-1}$. Hence, the the proof is completed. □

We next show that the sequence generated by the proposed method converges to a stationary point of (1.1), which means the point x^* satisfies the first optimality condition

$$0 \in \nabla g(x^*) + \partial h(x^*),$$

where $\partial h(x)$ is the subdifferential of h at x. Using Theorem 2.1 and the first optimality condition of (1.6), namely,

$$0 \in \nabla g(x_k) + B_k d_k + \partial h(\mathrm{Prox}_h^{B_k}(x_k - H_k \nabla g(x_k))), \tag{2.16}$$

we have the following theorem corresponding to Proposition 2.5 in [7]. If $d_k = 0$, then it follows from (2.16) that x_k is a stationary point. Also, we can prove the converse in a similar way to [7, 10].

Theorem 2.2. *Suppose that* Assumption 1 *is satisfied. Let the sequence* $\{x_k\}$ *be generated by* Algorithm 1 *with* (2.7) *and* (2.8)*, where* (2.5), (2.6) *and* (2.11) *are satisfied. Then,* x_k *is a stationary point if and only if* $d_k = 0$.

Moreover, we have the following global convergence theorem. We can prove the following theorem in a similar manner to the proof of Theorem 3.1 in [7].

Theorem 2.3. *Suppose that* Assumption 1 *is satisfied. Let the sequence* $\{x_k\}$ *be generated by* Algorithm 1 *with* (2.7) *and* (2.8)*, where* (2.5), (2.6) *and* (2.11) *are satisfied. If the objective function is bounded blew, then*

$$\lim_{k\to\infty} \|d_k\| = 0. \tag{2.17}$$

Furthermore, if generated sequence $\{x_k\}$ *is bounded, then any accumulation point of* $\{x_k\}$ *is a stationary point of* (1.1).

3. Numerical experiments

In this section, we investigate the performance of the proposed method (namely, Algorithm 1 with (2.7)). We compare the proposed method with other typical proximal gradient methods. We test the ℓ_1-regularized logistic regression problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} \log\left(1+\exp(-b_i x^T a_i)\right) + C\|x\|_1,$$

where $a_i \in \mathbb{R}^n$, $b_i \in \{-1, 1\}$, $i = 1, \cdots, m$ and C is a regularization parameter. We use binary classification datasets `gisette_scale`, `a9a` and `leukemia` as (a_i, b_i), $i = 1, \cdots, m$ from [6]. We denote `gisette_scale` as simply `gisette`. Table 1 gives details of these datasets. We set an initial point by $x_0 = (0, \cdots, 0)^T$ and the regularization parameter by $C = 10^{-3}$. We use the parameter ν_k in (3.1) as

$$\nu_k = \begin{cases} 0, & \text{if } s_{k-1}^T y_{k-1} \geq \bar{\nu}\|s_{k-1}\|^2, \\ \bar{\nu}\left(1 - \frac{s_{k-1}^T y_{k-1}}{\|s_{k-1}\|^2}\right), & \text{otherwise,} \end{cases} \tag{3.1}$$

where we set $\bar{\nu} = 10^{-3}$. Since $g(x) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1+\exp(-b_i x^T a_i)\right)$ is convex, it follows from $s_{k-1}^T y_{k-1} \geq 0$ that

$$s_{k-1}^T y_{k-1} + \bar{\nu}\left(1 - \frac{s_{k-1}^T y_{k-1}}{\|s_{k-1}\|^2}\right) s_k = (1-\bar{\nu}) s_{k-1}^T y_{k-1} + \bar{\nu}\|s_{k-1}\|^2 \geq \bar{\nu}\|s_{k-1}\|^2.$$

TABLE 1. Dataset information

Data name	gisette	a9a	leukemia
Number of data m	1000	32561	38
Dimension n	5000	123	7129

Thus we note that (2.5) holds. Following Nakayama et al. [9], we chose

$$\gamma_k = \rho_k \frac{s_{k-1}^T z_{k-1}}{z_{k-1}^T z_{k-1}} \quad (0 < \rho_{\min} \leq \rho_k \leq \rho_{\max} < 1), \tag{3.2}$$

and we set $\rho_k = 0.1, 0.2, ..., 0.9$. Note that the choice (3.1) guarantees condition (2.5). Furthermore, since it follows from (2.5) and (2.10) that

$$\frac{\bar{\nu}}{\bar{L}^2} \leq \frac{s_{k-1}^T z_{k-1}}{z_{k-1}^T z_{k-1}} \leq \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T z_{k-1}} \leq \frac{1}{\bar{\nu}},$$

parameter (3.2) satisfies (2.6). Thus, from Theorem 2.3 the proposed method converges globally. For the parameters of the line search (Step 3 in Algorithm 1), we set $\delta = 0.001$ and $\beta = 0.5$. Each algorithm terminates when $\|d_k\|_\infty \leq 10^{-6}$, where $\|\cdot\|_\infty$ is the infinity norm. All the numerical experiments are performed in Matlab 2017b on a PC with 2.3 GHz Intel Core i5, 8GB RAM running macOS High Sierra.

We compare the proposed method with TFOCS [2][1], PNOPT [7][2] and the inexact proximal memoryless BFGS method [10]. TFOCS is a software based on the proximal gradient method with acceleration techniques. PNOPT is a software of the proximal quasi-Newton method which uses the limited memory BFGS method as a choice of B_k in (1.5). In PNOPT and the inexact proximal memoryless BFGS method, the scaled proximal mapping is solved inexactly. For those softwares, we use the default parameter settings.

Table 2 shows the numerical results of TFOCS, PNOPT, the inexact proximal memoryless BFGS method (mless-BFGS) and the proposed method (mless-SR1). Iter and Time mean the number of iterations and the CUP time (second), respectively. In each column, we write the best result by using bold-face.

We first consider how choice of the parameter ρ_k affects the efficiency of our methods (namely, mless-SR1). We see that the efficiency of mless-SR1 greatly depends on the value of ρ_k. For `gisette` and `a9a`, the method with large values of ρ_k performed better, but there is the reverse tendency for `leukemia`. Next, we compare mless-SR1 with mless-BFGS. Then, mless-BFGS is superior

[1] `http://cvxr.com/tfocs/download/`

[2] `https://web.stanford.edu/group/SOL/software/pnopt/`

TABLE 2. Numerical results

Data name	gisette		a9a		leukemia	
	Iter	Time	Iter	Time	Iter	Time
TFOCS	7025	1384.38	1123	21.28	1056	**4.27**
PNOPT (L-BFGS)	**264**	**137.62**	**41**	3.55	**500**	96.92
mless-BFGS	2144	1898.72	164	9.74	1361	120.23
mless-SR1($\rho_k = 0.1$)	7743	1110.93	860	11.40	2148	11.84
mless-SR1($\rho_k = 0.2$)	8258	1147.18	383	4.81	3494	16.80
mless-SR1($\rho_k = 0.3$)	8563	1148.85	267	3.14	6678	32.87
mless-SR1($\rho_k = 0.4$)	9124	1213.54	227	2.62	11207	56.22
mless-SR1($\rho_k = 0.5$)	7003	932.56	182	2.16	12863	61.39
mless-SR1($\rho_k = 0.6$)	5307	696.41	189	2.19	15236	72.83
mless-SR1($\rho_k = 0.7$)	6676	871.12	164	1.91	16825	80.53
mless-SR1($\rho_k = 0.8$)	4869	630.99	170	1.99	17925	85.72
mless-SR1($\rho_k = 0.9$)	4271	548.00	151	**1.79**	17899	85.58

to mless-SR1 in the viewpoint of Iter. This result seems to be natural because the BFGS method is usually superior to the SR1 method in the case of unconstrained optimization. However, mless-SR1 (with better choice of ρ_k) is superior to mless-BFGS in the viewpoint of Time. This cause may be because mless-BFGS solves the subproblem to compute the scaled proximal mapping in each iteration. From these results, we see that the proposed method remedies other proximal memoryless quasi-Newton methods which need to solve the subproblem. Finally, we compare mless-SR1 with TFOCS and PNOPT in the perspective on Time. Then, we cannot conclude that mless-SR1 is always superior to the other methods, and the numerical performance depends on the choice of problems.

4. Concluding renarks

In this paper, we have proposed a proximal quasi-Newton method based on the memoryless SR1 formula. To establish the global convergence of the method, the proposed method uses the memoryless SR1 formula based on a modified spectral-scaling secant condition. We have shown the global convergence of the proposed method. In numerical experiments, we report some preliminary numerical results.

More recently, Becker et al. [4] gave the explicit form of the scaled proximal mapping (1.6) with two rank updates like as the BFGS formula. Therefore, we

can consider proximal quasi-Newton methods based on the memoryless BFGS formula instead of the memoryless SR1 formula. This is our further study.

Acknowledgements. This research was supported in part by JSPS KAKENHI (grant number 18K11179, 20K11698 and 20K14986) and the Research Institute for Mathematical Sciences in Kyoto University.

References

[1] A. Beck, *First-order Method in Optimization*, MOS-SIAM Series on Optimization. SIAM, 2017.
[2] S. Becker, E. J. Candés and M. C. Grant, *Templates for convex cone problems with applications to sparse signal recovery.* Mathematical Programming Computation **3** (2011), 165–218.
[3] S. Becker and M. J. Fadili, *A quasi-Newton proximal splitting method*, in: Advances in Neural Information Processing Systems 25, 2012, pp. 2618–2626.
[4] S. Becker, J. Fadili and P. Ochs, *On quasi-Newton forward-backward splitting: proximal calculus and convergence*, SIAM Journal on Optimization **29** (2019), 2445–2481.
[5] W. Y. Cheng and D. H. Li, *Spectral scaling BFGS method*, Journal of Optimization Theory and Applications **146** (2010), 305–319.
[6] C. C. Chang and C. J. Lin, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology **2** (2011), 1–27.
[7] J. D. Lee, Y. Sun and M. A. Saunders, *Proximal Newton-type methods for minimizing composite functions*, SIAM Journal on Optimization **24** (2014), 1420–1443.
[8] D. H. Li and M. Fukushima, *A modified BFGS method and its global convergence in nonconvex minimization.* Journal of Computational and Applied Mathematics **129** (2001), 15–35.
[9] S. Nakayama, Y. Narushima and H. Yabe, *A memoryless symmetric rank-one method with sufficient descent property for unconstrained optimization*, Journal of the Operations Research Society of Japan **61** (2018), 53–70.
[10] S. Nakayama, Y. Narushima and H. Yabe, *Inexact proximal memoryless quasi-Newton methods based on the Broyden family for minimizing composite functions*, Computational Optimization and Applications, to appear.
[11] S. Sra, S. Nowozin and S. J. Wright (eds.), *Optimization for Machine Learning*, The MIT Press, 2012.

S. Nakayama
Chuo University, Japan
E-mail address: shummin@kc.chuo-u.ac.jp

Y. Narushima
Keio University, Japan
E-mail address: narushima@ae.keio.ac.jp