A globally convergent regularized Newton method for ℓ_q-norm composite optimization problems

Xiaoqi Yang

Department of Applied Mathematics
The Hong Kong Polytechnic University

Joint work with Yuqia Wu (PolyU) and Shaohua Pan (SCUT)

Workshop on Optimization, Equilibrium and Complementarity

16-19 August, 2023, PolyU

Table of Contents

1 Introduction

2 A Hybrid of PG and Regularized Newton Method

3 Convergence Analysis

4 Numerical Experiments

5 References

Table of Contents

1 Introduction

2 A Hybrid of PG and Regularized Newton Method

3 Convergence Analysis

4 Numerical Experiments

5 References

Consider the ℓ_q-norm regularized composite optimization problem:

$$\min_{x\in\mathbb{R}^n} F(x) = f(Ax) + \lambda\|x\|_q^q, \tag{1}$$

where

- $A \in \mathbb{R}^{m\times n}$, $q \in (0,1)$ and $\lambda > 0$.
- f is twice continuously differentiable with $\inf_{z\in\mathbb{R}^m} f(z) > -\infty$.
- $\|x\|_q := \left(\sum_{i=1}^n |x_i|^q\right)^{1/q}$ denotes the ℓ_q quasi-norm of x.

Due to the nonconvex, nomsmooth and nonLipschitz property of the ℓ_q norm, problem (1) is a class of difficult nonconvex and nonsmooth optimizaton problems.

Problem (1) first appears in statistics as the bridge penalty regression [Frank93], and later appears in optimization as a special case of nonsmooth and nonconvex penalty problems studied by [Luo96] and and [Yang01, Huang03] for nonlinear optimization problems.

Literature Review

There are a lot of first-order optimization algorithms to solve problem (1):

- Hybrid orthogonal matching pursuit-smoothing gradient method (OMP-SG) [Chen10]
- Reweighted ℓ_1 minimization method [Lai13, Lu14, Chen14]
- Proximal gradient method [Wright09, Xu12, Zeng16, Hu17, Hu21]

Next, we first focus on proximal gradient method.

Proximal Gradient Method (PG)

For a proper lower semicontinuous (lsc) function $h\colon \mathbb{R}^n \to (-\infty, \infty]$, its proximal mapping associated to parameter $\mu > 0$ is defined by

$$\operatorname{prox}_{\mu h}(x) := \operatorname*{arg\,min}_{z\in\mathbb{R}^n} \left\{ \frac{1}{2\mu}\|z - x\|^2 + h(z) \right\} \quad \text{for } x \in \mathbb{R}^n.$$

When ∇f is Lipschitz continuous with Lipschitz constant $L_{\nabla f}$, the proximal gradient method for solving (1) is shown as follows. The essential part of PG lies in how to solve (2).

Algorithm 1 (Proximal gradient method)

```
Initialization: Choose initial point $x^0$ and $\gamma > \|A\|_2^2 L_{\nabla f}$. Let $k := 0$.
While the termination condition is not satisfied, do
    $x^{k+1} \in \operatorname{prox}_{\gamma^{-1}(\lambda\|\cdot\|_q^q)}(x^k - \gamma^{-1}\nabla f(x^k)).$    (2)
    $k \leftarrow k + 1$
end while
```

- When $q = 1/2, 2/3$, the proximal mapping of $\|x\|_q^q$ has a closed-form solution [Xu12, Cao13, Zeng16, Hu17]. This means that PG is able to solve (1) for $q = 1/2, 2/3$ with cheap computation cost.
- PG method for solving (1) has a global convergence guarantee if F is a KL function [Attouch10].
- When assuming that the limit point is a local minimizer, a linear convergence rate of the iterates is obtained in [Xu12, Zeng16, Hu17].
- In [Wright09], a class of PGs with nonmonotone line search strategy is proposed.

Features of the proximal gradient method:

- Weak condition for global convergence: KL property of the objective function is sufficient.
- Cheap computation cost: computing a proximal mapping per iterate.
- However, it at most achieves a linear convergence rate.

Features of the Newton-type method:

- The global convergence analysis of Newton-type methods with line search is limited to the subsequential convergence of the iterates, see, e.g., [Nocedal06].
- Expensive to compute the inverse of the (regularized) Hessian.
- Under some regularity conditions (e.g., strongly convex, local error bound condition, local Lipschitz Hessian), local superlinear convergence rate is achieved.

For unconstrained smooth optimization, the weakest condition for a Newton-type method to have a local superlinear convergence rate is a local error bound condition at local minima [Li04, Ueda10].

It is natural to ask whether it is possible to design a globally convergent Newton-type method for (1) with a local superlinear convergence rate.

Proximal Newton-type method

$$\min_{x \in \mathbb{R}^n} \Phi(x) := \phi(x) + h(x),$$

where ϕ is convex and twice differentiable and h is convex and not necessarily differentiable.

- The references [Lee14], [Kanzow21], [Mordukhovich22]: Choose $H_k \succ 0$ (an approximation to $\nabla^2 \phi(x_k)$), solve the following subproblem for a search direction

  $$\Delta x_k \in \operatorname{argmin}_d \nabla \phi(x_k)^\top d + \frac{1}{2} d^\top H_k d + h(x_k + d),$$

  and let

  $$x_{k+1} = x_k + t_k \Delta x_k.$$

Global convergence with superlinear rate is achieved!

Hybrid of PG and quasi-Newton methods

$$\min_{x\in\mathbb{R}^n} \Phi(x) := \phi(x) + h(x), \tag{3}$$

where ϕ is twice differentiable and h is not necessarily differentiable. In [Themelis18], a thorough study of stationarity of (3), criticality and optimality via the FBE $\Phi_\gamma(x)$ was given (see also [Poliqun96], [Beck16] and [Pang17]), and let $\bar{x}^k$ be a PG iterate and **Select** a quasi-Newton direction Δx^k and let the back-tracking $x^{k+1} = \bar{x}^k + t_k \Delta x^k$ satisfy

$$\Phi_{\gamma^{-1}}(x^{k+1}) \leq \Phi_{\gamma^{-1}}(x^k) - \sigma\gamma^2 \|x^k - \bar{x}^k\|^2.$$

Global convergence with superlinear rate is also achieved!

See also [Stella17] and [Ahookhosh21] and [Bareilles22] for a hybrid of PG and Riemannian update on an identified manifold.

The forward-backward envelope (FBE) Φ_γ of Φ:

$$\Phi_\gamma(x) = \inf_{z\in\mathbb{R}^n} \left\{ \phi(x) + \langle \nabla\phi(x), z - x\rangle + \frac{1}{2\gamma}\|z - x\|^2 + h(z) \right\}$$

Table of Contents

1 Introduction

2 A Hybrid of PG and Regularized Newton Method

3 Convergence Analysis

4 Numerical Experiments

5 References

Some Notations

For any given $\emptyset \neq S \subseteq \{1, \dots, n\}$, we define $|S|$ as the cardinality of S and

- $\psi(x) := f(Ax)$ and $g(x) := \|x\|_q^q$ for $x \in \mathbb{R}^n$.

Let $S = \text{supp}(x) := \{i | x_i \neq 0\}$ and

- $\psi_S(u) := f(A_S u)$ and $g_S(u) := \sum_{i \in S} |u_i|^q$ for $u \in \mathbb{R}^{|S|}$.
- $F_S(u) := \psi_S(u) + \lambda g_S(u)$ for $u \in \mathbb{R}^{|S|}$.

Critical point and L-type stationary point

Definition 2.1

- A vector $x \in \mathbb{R}^n$ is called a critical point of F if $0 \in \partial F(x)$, where $\partial F(x)$ denotes the limiting subdifferential of F at x.
- A vector $x \in \mathbb{R}^n$ is called an L-type stationary point of problem (1) if there exists a constant $\mu > 0$ such that $x \in \mathrm{prox}_{\mu^{-1}(\lambda g)}(x - \mu^{-1}\nabla\psi(x))$.

Since g is prox-regular and prox-bounded, one can show that a point x is an L-type stationary point of problem (1) if and only if $0 \in \partial F(x)$.

Proposition 2.2

For model (1), the set of L-type stationary points coincides with that of critical points.

Motivation of HpgSRN

Based on the fact that $\mathrm{dist}(0, \partial F(x)) = \|\nabla F_{\mathrm{supp}(x)}(x_{\mathrm{supp}(x)})\|$, seeking a point $\overline{x}$ satisfying $0 \in \partial F(\overline{x})$ is equivalent to finding $\overline{x}$ such that $\|\nabla F_{\mathrm{supp}(\overline{x})}(\overline{x}_{\mathrm{supp}(\overline{x})})\| = 0$. Given that $\overline{S} := \mathrm{supp}(\overline{x})$ is known but $\overline{x}$ is unknown, we can turn to find $\overline{u} \in \mathbb{R}^{|\overline{S}|}$ with $\overline{u}_i \neq 0$ for all i such that $\|\nabla F_{\overline{S}}(\overline{u})\| = 0$. Then, $\overline{x} = (\overline{u}; 0)$ is our desired point since

$$\|\nabla F_{\overline{S}}(\overline{u})\| = 0 \Longrightarrow 0 \in \partial F(\overline{x}) \text{ with } \overline{x} = (\overline{u}; 0).$$

Based on this line, we find a critical point of F by the following steps:

(a) Use a PG method to seek a good estimate in some neighborhood of a potential critical point.

(b) Apply a regularized Newton method in the subspace associated to the support of the iterate generated by the PG method.

PGls

The detailed iterate of our proposed algorithm needs the algorithm flow of the PG method with a monotone line search (PGls), i.e., a monotone version of SpaRSA [Wright09].

Let $x \in \mathbb{R}^n$ be the current iterate and $\mu > 0$ be an initial step-size. The PGls returns a new iterate x^+ and the used step-size μ_+ such that $F(x^+)$ has a certain decrease.

Algorithm Flow of PGls: $[x^+, \mu_+] = \mathcal{G}(x, \mu; \widetilde{\tau}, \widetilde{\alpha}, \lambda)$

```
Input: $x \in \mathbb{R}^n$ and parameters $\mu > 0, \widetilde{\tau} > 1$ and $\widetilde{\alpha} > 0$.
Let $l = 0, \mu_l = \mu$ and $x^l \in \mathrm{prox}_{\mu_l^{-1}(\lambda g)}\big(x - \mu_l^{-1}\nabla\psi(x)\big)$.
while $F(x^l) > F(x) - (\widetilde{\alpha}/2)\|x^l - x\|^2$
    • Let $\mu_{l+1} = \widetilde{\tau}\mu_l$ and $l \leftarrow l + 1$;
    • Seek $x^l \in \mathrm{prox}_{\mu_l^{-1}(\lambda g)}\big(x - \mu_l^{-1}\nabla\psi(x)\big)$;
end (while)
Let $x^+ = x^l$ and $\mu_+ = \mu_l$.
```

A Hybrid of PG and Regularized Newton Method (HpgSRN)

Initialization: Choose $\widetilde{\tau} > 1, \widetilde{\alpha} > 0$, $\mu_{\max} > \mu_{\min} > 0$. Choose an initial $x^0 \in \mathbb{R}^n$ and a tolerance $\epsilon \geq 0$. Let $k = 0$.

Step 1: proximal gradient step

(S1) Choose an initial step-size $\mu_k \in [\mu_{\min}, \mu_{\max}]$. Set $[\overline{x}^k, \overline{\mu}_k] = \mathcal{G}(x^k, \mu_k; \widetilde{\tau}, \widetilde{\alpha}, \lambda)$.

(S2) **If** $\overline{\mu}_k \|x^k - \overline{x}^k\|_\infty \leq \epsilon$, output x^k; **otherwise** go to (S3).

(S3) Let $\overline{\omega}_k = \overline{\mu}_k + \lambda q(q-1)|\overline{x}^k|_{\min}^{q-2}$. **If**

$$\mathrm{sign}(x^k) = \mathrm{sign}(\overline{x}^k) \;\text{ and }\; \overline{\mu}_k + \lambda q(q-1)|x^k|_{\min}^{q-2} \geq \frac{1}{2}\overline{\omega}_k, \tag{4}$$

then go to **Step 2**; **otherwise** let $x^{k+1} = \overline{x}^k$ and $k \leftarrow k+1$. Go to **Step 1**.

Step 2: subspace regularized Newton step

(S4) Let $S_k = \mathrm{supp}(x^k)$ and $u^k = x^k_{S_k}$. Seek a subspace Newton direction Δu^k by solving $G^k \Delta u = -\nabla F_{S_k}(u^k)$, where $G^k = \nabla^2 F_{S_k}(u^k) + (b_1\zeta_k + b_2\|\nabla F_{S_k}(u^k)\|^\sigma)I$.Let $d^k_{S_k} = \Delta u^k$ and $d^k_{S_k^c} = 0$.

(S5) Let m_k be the smallest nonnegative integer m such that

$$F_{S_k}(u^k + \beta^m d^k_{S_k}) \leq F_{S_k}(u^k) + \varrho\beta^m \langle \nabla F_{S_k}(u^k), d^k_{S_k} \rangle. \tag{5}$$

(S6) Let $\alpha_k = \beta^{m_k}$ and $x^{k+1} = x^k + \alpha_k d^k$ and $k \leftarrow k+1$. Go to **Step 1**.

Table of Contents

1. Introduction
2. A Hybrid of PG and Regularized Newton Method
3. Convergence Analysis
4. Numerical Experiments
5. References

Technical lemmas

For any given $\gamma > 0, s \in \mathbb{R}$, define a real-valued function

$$h_{\gamma,s}(t) := \frac{\gamma}{2}(t-s)^2 + \lambda|t|^q \ \text{ for } \ t \in \mathbb{R}. \tag{6}$$

It is easy to see that $t = 0$ is always a local minimizer of $h_{\gamma,s}$ and that the absolute value of another possible local minimizer is greater than $\overline{\nu}$, where $\overline{\nu} := \left(\frac{\lambda q(1-q)}{\gamma}\right)^{\frac{1}{2-q}}$.

Lemma 3.1

For any given $0 < \upsilon < M < \infty$, there exists a constant $\varpi > 0$ such that for any $\gamma > 0$ and $s \in \mathbb{R}$ with $|\overline{t}(\gamma, s)| \in [\upsilon, M]$,

$$h''_{\gamma,s}(\overline{t}(\gamma,s)) = \gamma + \lambda q(q-1)|\overline{t}(\gamma,s)|^{q-2} \geq \varpi.$$

HpgSRN is Different from PG

From the iterate steps of HpgSRN, the sequence $\{x^k\}_{k\in\mathbb{N}}$ consists of two parts, i.e., $\{x^k\}_{k\in\mathbb{N}} = \{x^k\}_{k\in\mathcal{K}_1} \cup \{x^k\}_{k\in\mathcal{K}_2}$, where

$$\mathcal{K}_1 := \left\{k \in \mathbb{N} \mid x^{k+1} \text{ is generated by Step 1}\right\} \text{ and } \mathcal{K}_2 := \mathbb{N}\backslash\mathcal{K}_1.$$

In other words, if condition (4) is satisfied in k-th iterate, then $k \in \mathcal{K}_2$.

Corollary 3.1

There exists $\overline{k} \in \mathbb{N}$ such that for any $k_1, k_2 \in \mathbb{N}$ with $k_2 - k_1 > \overline{k}$, $[k_1, k_2] \cap \mathcal{K}_2 \neq \emptyset$.

Corollary 3.1 states that $\mathcal{K}_2$ contains infinite indices, so HpgSRN is different from PG method. In fact, under an additional assumption, we will improve this result so that after a finite number of steps, the iterates of HpgSRN always enter into Step 2.

Technical lemmas

Assumption 1

$\nabla^2 f$ is locally Lipschitz continuous on $\mathbb{R}^m$.

Lemma 3.2

Let $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ be the sequences yielded by HpgSRN. Then, under Assumption 1, the following assertions hold.

(i) There exists $\widehat{\gamma} > 0$ such that for all $k \in \mathbb{N}$, $F(x^{k+1}) \leq F(x^k) - \frac{\widehat{\gamma}}{2}\|x^k - \overline{x}^k\|^2$.

(ii) $\lim_{k\to\infty} \|x^k - \overline{x}^k\| = 0$.

(iii) There exists $\widetilde{c} > 0$ such that $\mathrm{dist}(0, \partial F(x^k)) \leq \widetilde{c}\|x^k - \overline{x}^k\|$ for all $k \in \mathcal{K}_2$.

(iv) Each accumulation point of $\{x^k\}_{k\in\mathbb{N}}$ is an L-type stationary point of (1).

Among others, (i) states that $\{F(x^k)\}_{k\in\mathbb{N}}$ is sufficiently decreasing, while (iii) reveals the subdifferential gap of F at x^k for all $k \in \mathcal{K}_2$. Part (iv) gives the subsequential convergence result of the iterate sequence.

Technical lemmas

Lemma 3.3

Let $\{x^k\}_{k\in\mathbb{N}}$ and $\{\overline{x}^k\}_{k\in\mathbb{N}}$ be the sequences given by HpgSRN. Then, under Assumption 1, the following assertions hold.

(i) There exists an index set $S_* \subseteq [n]$ such that for all sufficiently large k,

$$\mathrm{supp}(x^k) = \mathrm{supp}(\overline{x}^k) = S_*;$$

and furthermore, every cluster point x^* of $\{x^k\}_{k\in\mathbb{N}}$ satisfies $\mathrm{supp}(x^*) = S_*$.

(ii) There exists $\widehat{k} \in \mathbb{N}$ such that for all $k \geq \widehat{k}$, $k \in \mathcal{K}_2$.

The second part of this lemma means that under Assumption 1, after a finite number of iterates, HpgSRN reduces to a regularized Newton method for minimizing the function F_{S_*}, where S_* is the one in part (i).

Kurdyka-Łojasiewicz (KL) property

Definition

A proper extended real-valued function $h\colon \mathbb{R}^n \to (-\infty, \infty]$ is said to have the Kurdyka-Łojasiewicz (KL) property at a point $\overline{x} \in \mathrm{dom}\partial h$ if there exist $\eta \in (0, \infty]$, a neighborhood $\mathcal{U}$ of $\overline{x}$, and a continuous concave function $\varphi\colon [0, \eta) \to \mathbb{R}_+$ satisfying

$$\varphi(0) = 0, \varphi \text{ is continuously differentiable on } (0, \eta) \text{ and } \varphi'(s) > 0, \ \forall s \in (0, \eta) \tag{7}$$

such that for all $x \in \mathcal{U} \cap \{x \in \mathbb{R}^n \mid h(\overline{x}) < h(x) < h(\overline{x}) + \eta\}$,

$$\varphi'(h(x) - h(\overline{x}))\mathrm{dist}(0, \partial h(x)) \geq 1.$$

If φ can be chosen as $\varphi(s) = c\sqrt{s}$ for some constant $c > 0$, then h is said to have the KL property of exponent $1/2$ at $\overline{x}$. If h has the KL property of exponent $1/2$ at each point of $\mathrm{dom}\,\partial h$, then h is called a KL function of exponent $1/2$.

Convergence rate of objective function value sequence

Proposition 3.4

Suppose that Assumption 1 holds, and that F is a KL function of exponent $1/2$. Then $\{F(x^k)\}_{k\in\mathbb{N}}$ converges to some value F^* in a Q-linear rate.

We only achieve the linear convergence rate of the sequence of $\{F(x^k)\}_{k\in\mathbb{N}}$ here. Later, under an additional assumption, we will show the linear convergence rate of the iterate sequence $\{x^k\}_{k\in\mathbb{N}}$.

Convergence rate of iterate sequence

Assumption 2

It holds that $\liminf\limits_{\mathcal{K}_2 \ni k \to \infty} \frac{-\langle \nabla F_{S_k}(u^k), d^k_{S_k} \rangle}{\|\nabla F_{S_k}(u^k)\| \|d^k_{S_k}\|} > 0$, where $u^k = x^k_{S_k}$.

Theorem 3.5

Suppose Assumptions 1 and 2 hold. The following assertions hold.

(i) If F is a KL function, then $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty$, and consequently, $\{x^k\}_{k\in\mathbb{N}}$ converges to an L-type stationary point of (1), say x^*.

(ii) If F is a KL function of exponent $1/2$ at x^*, then $\{x^k\}_{k\in\mathbb{N}}$ converges R-linearly to x^*.

Discussion of Assumption 2

- Assumption 2 essentially requires that the angle between $\nabla F_{S_k}(u^k)$ and $d^k_{S_k}$ is sufficiently away from $\pi/2$ and close to π. It is very common in the subsequential convergence analysis of line search Newton-type methods (see, e.g., [Nocedal06]), which guarantees that $\lim_{k\to\infty} \|\nabla F_{S_k}(u^k)\| = 0$.

- The authors of [Themelis18] proposed ZeroFPR, a hybrid of PG and quasi-Newton method for minimizing so-called forward-backward envelop of a nonsmooth composite problem. They achieve a global convergence under the condition

$$\exists \text{ a constant } \widehat{c} > 0 \text{ such that } \|d^k\| \leq \widehat{c}\|x^k - \overline{x}^k\| \text{ for all } k. \tag{8}$$

In fact, Assumption 2 is strictly weaker than (8) in the setting of our algorithm.

Lemma 3.6

Suppose that Assumption 1 holds. If d^k yielded by Step 2 of HpgSRN satisfies condition (8) for all $k \in \mathcal{K}_2$, then Assumption 2 holds.

Discussion of KL Property with Exponent 1/2

The KL property of exponent 1/2 plays a crucial role in achieving the linear convergence rate of a class of first-order method. The following proposition establishes the equivalence between the KL property of exponent 1/2 of F and that of F_S.

Proposition 3.7

For any given $\overline{x} \in \mathbb{R}^n \backslash \{0\}$, F has the KL property of exponent 1/2 at $\overline{x}$ if and only if $F_{\overline{S}}$ with $\overline{S} = \mathrm{supp}(\overline{x})$ has the KL property of exponent 1/2 at $\overline{u} = \overline{x}_{\overline{S}}$.

- By Proposition 3.7, to check the KL property with exponent $1/2$ of F at x^*, it suffices to verify that of F_{S_*} at $x^*_{S_*}$. Due to the sufficient smoothness of F_{S_*} at $x^*_{S_*}$, the verification of the latter is easier than that of the former.
- By [Zeng16, Lemma 3], the nonsingularity of $\nabla^2 F_{S_*}(x^*_{S_*})$ implies the KL property of exponent $1/2$ for F_{S_*} at $x^*_{S_*}$.
- Then by Theorem 3.5 (ii), if $\{x^k\}_{k\in\mathbb{N}}$ converges to x^* and $\nabla^2 F_{S_*}(x^*_{S_*})$ is nonsingular, then $\{x^k\}_{k\in\mathbb{N}}$ converges to x^* in a linear convergence rate.

Superlinear Convergence of HpgSRN

By Theorem 3.5, if Assumptions 1 and 2 hold and F is a KL function, the sequence $\{x^k\}_{k\in\mathbb{N}}$ is convergent. In the sequel, we denote its limit by x^*. By Lemma 3.3, $\mathrm{supp}(x^*) = S_*$. Write

$$u^* := x^*_{S_*} \text{ and } \mathcal{U}^* := \left\{u \in \mathbb{R}^{|S_*|} \mid \nabla F_{S_*}(u) = 0, \nabla^2 F_{S_*}(u) \succeq 0\right\}.$$

Note that $\mathcal{U}^*$ is not necessarily the set of local minima of F_{S_*}.

To achieve the superlinear convergence rate of $\{x^k\}_{k\in\mathbb{N}}$, we need to bound ζ_k involved in the matrix G_k by $\mathrm{dist}(u^k, \mathcal{U}^*)$ as in the following lemma.

Lemma 3.8

Suppose that Assumptions 1 and 2 hold, and that F is a KL function. If $\nabla^2 F_{S_*}(u^*) \succeq 0$, then there exists $c_H > 0$ such that for all sufficiently large k, $\zeta_k \le c_H \mathrm{dist}(u^k, \mathcal{U}^*)$.

It is worth noting that [Ueda10, Lemma 5.2] achieved the result of Lemma 3.7 by a stronger condition $\nabla^2 F_{S_*}(u^*) \succ 0$.

Theorem 3.9

Suppose that Assumptions 1 and 2 hold, and that F is a KL function. If $\nabla^2 F_{S_*}(u^*) \succeq 0$ and there exist $\delta > 0$ and $\kappa_1 > 0$ such that for all $u \in \mathbb{B}(u^*, \delta)$,

$$\kappa_1 \mathrm{dist}(u, \mathcal{U}^*) \leq \|\nabla F_{S_*}(u)\|, \tag{9}$$

then the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to x^* in a Q-superlinear rate of order $1+\sigma$.

- The proof of the superlinear convergence of E-RNM in [Ueda10] requires the local optimality of x^*. After checking its proof, we found that the local optimality of x^* was only used to achieve the result of Lemma 3.8. Thus the local optimality of x^* can be removed.
- The current weakest condition for a second-order method to have a local superlinear convergence rate is "local error bound condition":

  $$k\mathrm{dist}(u, X^*) \leq \|\nabla f(x)\|, \ \forall x \in B(x^*, \delta),$$

  where X^* is the set of local optimal solutions. See [Ueda10], [Li04], [19].
- The local error bound condition (9) is a little stronger than the metric subregularity of ∇F_{S_*} at u^* for the origin because $\mathcal{U}^*$ may be a strict subset of $\nabla F_{S_*}^{-1}(0)$, but it does not require the isolatedness of u^* and its local optimality.

Table of Contents

1. Introduction
2. A Hybrid of PG and Regularized Newton Method
3. Convergence Analysis
4. Numerical Experiments
5. References

We apply HpgSRN to solving the ℓ_q-norm regularized linear and logistic regression problems on real data, which respectively take the form of (1) with $f = f_1$ or f_2, where

- $f_1(z) := \frac{1}{2}\|z - b\|^2$
- $f_2(z) := \sum_{i=1}^{m} \log\big(1 + \exp(-b_i z_i)\big)$

for $z \in \mathbb{R}^m$. Here, $b \in \mathbb{R}^m$ is a given vector. Clearly, such f satisfies Assumption 1 and the associated F is a KL function. All numerical tests are conducted on a desktop running in MATLAB R2020b and 64-bit Windows System with an Intel(R) Core(TM) i7-10700 CPU 2.90GHz and 32.0 GB RAM.

Description of the Implementation of HpgSRN

- The initial step-size μ_k is chosen by the Barzilai-Borwein (BB) rule,

$$\mu_k = \max\left\{10^{-20}, \min\left\{10^{20}, \frac{\langle x^k - x^{k-1}, \nabla\psi(x^k) - \nabla\psi(x^{k-1})\rangle}{\|x^k - x^{k-1}\|^2}\right\}\right\}.$$

- For each $k \in \mathcal{K}_2$, we call the MATLAB function eigs to compute the approximate smallest eigenvalue of $\nabla^2 F_{S_k}(u^k)$, which requires about $O(|S_k|^2)$ flops by [Stewart02]. Since $|S_k|$ is usually much smaller than n, this computation cost is not expensive.

- In addition, we choose

  $$\widetilde{\tau} = 10,\ \widetilde{\alpha} = 10^{-8},\ \sigma = 0.5,\ b_1 = 1{+}10^{-3},\ b_2 = 10^{-3}, \varrho = 10^{-4}, \beta = 2.$$

- We solve the linear system in (S4) via a direct method if $|S_k| < 500$, otherwise a conjugate gradient method.
- Our preliminary tests indicate that (1) with $q = 1/2$ usually has better performance than (1) with other $q \in (0, 1)$ in terms of the CPU time and the sparsity. This coincides with the conclusion in [Hu17].

Description of Comparison

- We compare the performance of HpgSRN with that of ZeroFPR [Themelis18] and that of PGls to check the effect of the additional subspace regularized Newton step on HpgSRN. The parameters of PGls are chosen to be the same as those involved in Step 1 of HpgSRN except $\widetilde{\tau} = 2$.

- For the three algorithms, we adopt the stopping criterion

$$\gamma \| x^k - \mathrm{prox}_{\gamma^{-1}\lambda g}(x^k - \gamma^{-1}\nabla\psi(x^k))\|_\infty < 10^{-3} \text{ or } k \geq 50000,$$

  where $\gamma = L/0.95$ and L is an estimation of the Lipschitz constant of $\nabla\psi(\cdot)$. It is well-known that the Lipschitz constants of $A^\top \nabla f_1(A\cdot)$ and $A^\top \nabla f_2(A\cdot)$ are $\|A\|^2$ and $0.25\|A\|^2$, respectively.

- As in ZeroFPR, we choose $x^0 = 0$ as the starting point. Although $x^0 = 0$ is a local minimizer of F and hence an L-type stationary point. It is not a good one in terms of objective value; see the difference between $F(0)$ and the objective function value for each example in tables given later.

Description of Data Set

The data set used to test HpgSRN on ℓ_q regularized least square model:

- we conduct this experiments with (A, b) from LIBSVM datasets (see `https://www.csie.ntu.edu.tw`). As suggested in [Huang10], for **housing** and **space_ga**, we expand their original features with polynomial basis functions.

Explanation of Table 1:

- The numerical results including the number of iterations (Iter#), the CPU times in seconds (Time), the objective function values (Fval) and the cardinality of the outputs (Nnz). In particular, to check the effect of the regularized Newton steps in HpgSRN, we record its number of iterations in the form $M(N)$, where M means the total number of iterates and N means the number of regularized Newton steps.
- The second column of Table 1 lists the values of $\|A\|^2$ and $F(0)$, which reflect the condition number of the Hessian matrix of the loss function ψ and the quality of the starting point x^0 respectively.
- For each dataset, we solve (1) associated to f_1 and $\lambda = \lambda_c \|A^\top b\|_\infty$ for two different λ_c's with the three solvers.

Table 1: Numerical comparisons on ℓ_q-norm regularized linear regressions with LIBSVM datasets

Data (m, n)	$\|\|A\|\|^2$ $F(0)$	λ_c	Index	HpgSRN	ZeroFPR	PGls
space_ga9 (3107, 5505)	4.01e3 5.77e3	10^{-3}	Iter#	17(5)	43	180
			Time	0.45	0.98	0.93
			Fval	36.47	37.24	37.15
			Nnz	7	7	6
		10^{-4}	Iter#	230(64)	476	3058
			Time	2.26	9.03	16.48
			Fval	20.93	20.31	21.57
			Nnz	15	19	15
housing7 (506, 77520)	3.28e5 1.50e5	10^{-3}	Iter#	639(157)	4164	25133
			Time	14.45	2.13e2	4.08e2
			Fval	2.25e3	2.57e3	2.56e3
			Nnz	27	49	57
		10^{-4}	Iter#	1765(485)	18807	50000
			Time	49.26	9.81e2	8.59e2
			Fval	8.89e2	9.27e2	9.17e2
			Nnz	82	123	135
E2006.test (3308, 72812)	4.79e4 2.46e4	10^{-4}	Iter#	3(0)	3	3
			Time	0.03	0.25	0.03
			Fval	2.45e2	2.45e2	2.45e2
			Nnz	1	1	1
		10^{-5}	Iter#	3(0)	4	4
			Time	0.05	0.25	0.04
			Fval	2.40e2	2.40e2	2.40e2
			Nnz	1	1	1

cont'd

E2006.train (16087, 150348)	1.91e5 1.03e5	10^{-4}	Iter#	3(0)	3	3
			Time	0.09	1.06	0.09
			Fval	1.22e3	1.22e3	1.22e3
			Nnz	1	1	1
		10^{-5}	Iter#	4(0)	4	4
			Time	0.11	1.05	0.11
			Fval	1.20e3	1.20e3	1.20e3
			Nnz	1	1	1
log1p.E2006.test (3308, 1771946)	1.46e7 2.46e4	10^{-4}	Iter#	372(88)	827	1416
			Time	33.54	2.87e2	1.16e2
			Fval	2.35e2	2.43e2	2.37e2
			Nnz	5	4	6
		10^{-5}	Iter#	755(166)	6708	22305
			Time	1.01e2	2.28e3	2.30e3
			Fval	1.54e2	1.53e2	1.49e2
			Nnz	385	460	389
log1p.E2006.train (16087, 4265669)	5.86e7 1.03e5	10^{-4}	Iter#	286(58)	855	1621
			Time	77.95	8.57e2	3.85e2
			Fval	1.16e3	1.16e3	1.16e3
			Nnz	7	5	4
		10^{-5}	Iter#	944(195)	5610	33112
			Time	3.14e2	5.26e3	8.83e3
			Fval	1.02e3	1.02e3	1.01e3
			Nnz	141	184	155

Description of Data Set

For the ℓ_q-norm regularized logistic regressions, we also use (A, b) from LIBSVM datasets. For each data, we solve (1) associated to f_2 and $\lambda = \lambda_c \max_{1\leq j\leq n} \|A_j\|_1$ for two different λ_c's with the three solvers.

Numerical comparisons on ℓ_q-norm regularized logistic regressions with LIBSVM datasets

Data (m, n)	$\|A\|^2$ $F(0)$	λ_c	Index	HpgSRN	ZeroFPR	PGls
colon-cancer (62, 2000)	1.94e4 42.98	10^{-2}	Iter#	48(6)	730	94
			Time	0.04	0.74	0.06
			Fval	7.97	10.58	7.77
			Nnz	10	9	9
		10^{-3}	Iter#	94(9)	1853	175
			Time	0.07	2.07	0.11
			Fval	1.03	1.07	1.07
			Nnz	11	12	12
rcv1 (20242, 47236)	4.48e2 1.40e4	10^{-2}	Iter#	65(10)	448	1193
			Time	1.00	6.35	11.24
			Fval	4.23e3	4.35e3	4.24e3
			Nnz	165	167	164
		10^{-3}	Iter#	365(96)	2081	5536
			Time	7.78	29.27	88.65
			Fval	1.28e3	1.53e3	1.27e3
			Nnz	704	741	717
news20 (19996, 1355191)	1.73e3 1.39e4	10^{-2}	Iter#	44(6)	170	981
			Time	2.65	36.61	53.14
			Fval	9.73e3	1.04e4	9.53e3
			Nnz	51	42	50
		10^{-3}	Iter#	410(99)	1528	18538
			Time	41.45	3.44e2	1.43e3
			Fval	4.31e3	4.71e3	4.25e3
			Nnz	385	371	401

Conclusion of numerical experiments

To sum up, HpgSRN outperforms the other two algorithms in the following aspects:

- HpgSRN requires the least CPU time for all the test examples compared to ZeroFPR and PGIs, and for those large scale examples, HpgSRN is at least ten times faster than ZeroFPR and PGIs.
- The outputs of the objective function value and the sparsity yielded by HpgSRN have a comparable even better quality. This indicates that the introduction of second-order steps improves greatly the performance of the first-order method. We also observe that for most of examples, the iterates generated by the regularized Newton step account for about 10%–35% of the total iterates.

Table of Contents

1 Introduction

2 A Hybrid of PG and Regularized Newton Method

3 Convergence Analysis

4 Numerical Experiments

5 References

[Hu17] Y. Hu, C. Li, K. Meng, J. Qin, and X. Yang, Group sparse optimization via $\ell_{p,q}$ regularization, The Journal of Machine Learning Research, 18 (2017), pp. 960–1011

[Themelis18] A. Themelis, L. Stella, and P. Patrinos, Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms, SIAM Journal on Optimization, 28 (2018), pp. 2274–2303.

[Ueda10] K. Ueda and N. Yamashita, Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization, Applied Mathematics and Optimization, 62 (2010), pp. 27–46.

[Wright09] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, Sparse reconstruction by separable approximation, IEEE Transactions on Signal Processing, 57 (2009), pp. 2479–2493.

[Xu12] Z. Xu, X. Chang, F. Xu, and H. Zhang. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. IEEE Transactions on neural networks and learning systems, 23 (2012), pp. 1013–1027.

Thank You for Your Attentions!