An Inexact Proximal Semismooth Newton Method with Superlinear Convergence to Degenerate Solutions Under the Hölderian Error Bound

LEE Ching-pei
Academia Sinica

Joint work with Stephen J. Wright.
Talk at ICCOPT 2022 (Lehigh, Jul. 27, 2022)

Outline

Problem Setting

Consider the following generalized equation problem:

$$\text{find } x \in \mathcal{H} \quad \text{such that} \quad 0 \in (A + B)(x), \tag{GE}$$

- $\mathcal{H}$ is a Hilbert space: inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$
- $A, B : \mathcal{H} \rightrightarrows 2^{\mathcal{H}}$
- B maximal monotone
- Assume that the solution set Ω to (GE) is non-empty.
- A single-valued and continuous everywhere, and locally Lipschitz continuous in a neighborhood U of Ω:

$$\exists L \geq 0 : \quad \|A(x) - A(y)\| \leq L\|x - y\|, \quad \forall x, y \in U.$$

Regularized Optimization

A more approachable special case we will separately discuss in detail is regularized optimization:

$$\min_{x \in \mathcal{H}} \quad F(x) := f(x) + \Psi(x), \tag{Reg-Opt}$$

- f continuously differentiable with ∇f Lipschitz continuous
- $\Psi : \mathcal{H} \to [-\infty, \infty]$ convex, proper, and closed
- Namely, $A = \nabla f, B = \partial \Psi$ in (GE)

Newton's Method

- For (GE), if $A \in \mathcal{C}^1$, ∇A is Lipschitz continuous, $B \equiv 0$, $0 \in A(x^*)$, $\nabla A(x^*)$ is nonsingular, the iterates $\{x_t\}$ converge to x^*, Newton's method

$$p_t := \nabla A(x_t)^{-1} A(x_t), \quad x_{t+1} = x_t + p_t$$

ensures quadratic convergence to x^* when we are close enough to it:

$$\|x_{t+1} - x^*\| = O\left(\|x_t - x^*\|^2\right)$$

and quadratic convergence of the residual to 0:

$$\|A(x_{t+1})\| = O\left(\|A(x_t)\|^2\right)$$

- For large-scale problems such that inverting the Jacobian is too costly, we use iterative methods to obtain truncated Newton directions: $p_t \approx (\nabla A(x_t))^{-1} A(x_t)$ instead
- With suitable stopping conditions for the iterative method, the same convergence rates can be retained

Proximal Newton for (Reg-Opt)

- For smooth optimization, namely (Reg-Opt) with $\Psi \equiv 0$, the (truncated) Newton direction can be seen as
$$p_t \approx \operatorname*{argmin}_{p \in \mathcal{H}} \langle \nabla f(x_t),\, p \rangle + \frac{1}{2} \langle p,\, \nabla^2 f(x_t) p \rangle.$$
- Thus a natural way to extend Newton's method to the regularized setting in which $\Psi \neq 0$ is to solve
$$p_t \approx \operatorname*{argmin}_{p \in \mathcal{H}} \langle \nabla f(x_t),\, p \rangle + \frac{1}{2} \langle p,\, \nabla^2 f(x_t) p \rangle + \Psi(x_t + p).$$
- Lee et al. (2014)[1]: If $f \in \mathcal{C}^2$ is strongly convex and $\nabla^2 f$ is Lipschitz continuous, we still get quadratic convergence for $\|x_t - x^*\|$
- We can generalize this approach back to (GE)
$$x_{t+1} \approx \left(\nabla A(x_t) + B\right)^{-1} \left(\nabla A(x_t) - A\right)(x_t),$$
and similar local convergence results can be easily obtained

[1] Jason D. Lee, Yuekai Sun, and Michael A. Saunders. Proximal Newton-type methods for minimizing composite functions.
SIAM Journal on Optimization, 24(3):1420–1443, 2014

Questions to Ask

- What if ∇A is singular and Ω is not a singleton?
- Is differentiability of A really needed? Does semismooth Newton still work in this setting (nondifferentiable A and a possibly singular generalized Jacobian ∂A, further together with the additional set-valued B term)?
- If we only have either nondifferentiability of A or singularity of ∇A, the problem should not be too difficult, but the combination of these two degeneracy conditions makes the situation more complicated

This Talk I

- Not requiring A to be differentiable, not requiring its (Clarke) generalized Jacobian to be nonsingular, allowing Ω to be an unbounded set (so not assuming iterates converging to a point)
- Using a damping strategy, we get superlinear convergence of inexact forward-backward-semismooth-Newton for (GE) for both $\operatorname{dist}(x_t, \Omega)$ and the residual

$$r(x) := \left\| x - (\mathsf{Id} + B)^{-1}(\mathsf{Id} - A)x \right\|, \qquad \text{(Forward-Backward Residual)}$$

(which is 0 iff $0 \in (A+B)x$ and is continuous), where Id is the identity operator, under a Hölderian error bound condition: there are $\kappa, q > 0$ such that

$$\operatorname{dist}(x, \Omega) = \operatorname{dist}\left(x, (A+B)^{-1}(0)\right) \leq \kappa r(x)^q, \qquad \text{(HEB)}$$

for all x in a neighborhood U of Ω (Q-superlinear convergence can be guaranteed for $q > (-1+\sqrt{33})/8$)

This Talk II

- For (Reg-Opt), we futher obtain finite termination for $q > 1$
- For (Reg-Opt), we also propose a globally convergent (only need convexity of f and Ψ) algorithm that possesses all the local properties above (need (HEB)), and ensures that $F(x_t)$ is strictly decreasing and also converges Q-superlinearly to $F^* := \min_{x \in \mathcal{H}} F(x)$, without needing f to be twice-differentiable

Outline

Motivations

- To deal with nondifferentiable A:
  - Define $\partial A(x)$ as the (Clarke) generalized Jacobian of A at x, which is well-defined in U as A is locally Lipschitz continuous
  - Take any element from $\partial A(x)$, abuse the notation to still call it $\nabla A(x)$
- To deal with possible singularity of ∇A:
  - Add a damping term that vanishes as we approach Ω: we do not know Ω a priori, so need to rely on (Forward-Backward Residual)
- As the solution is inexact, also allow approximation of $\nabla A(x)$

Algorithm

- Given the current iterate x_t, we update the iterate by

$$x_{t+1} \approx (H_t + B)^{-1}(H_t - A)(x_t),\ H_t := (\mu_t \mathrm{Id} + J_t),\ \mu_t := cr(x_t)^{\rho}, \qquad \text{(Proximal semismooth Newton)}$$

  for some given $c > 0, \rho \geq 0$

- J_t is a positive semidefinite linear operator with $\exists \nabla A(x_t) \in \partial A(x_t)$ such that $\|J_t - \nabla A(x_t)\| = O\left(r(x_t)^{\theta}\right)$, for $\theta \in [\rho, 1]$

  (only need A maximal monotone around points with $r(x) = 0$)

- The resolvent $(\mathrm{Id} + B)^{-1}$ of B is single-valued and well-defined, and hence so is $(H_t + B)^{-1}$ as J_t is positive semidefinite.

- For the approximate solution, consider the following criterion with $\nu \geq 0$:

$$r_t(x_{t+1}) := \left\| x_{t+1} - (\mathrm{Id} + B)^{-1}\left((H_t - A)(x_t) - (H_t - \mathrm{Id})(x_{t+1})\right)\right\|$$
$$\leq \nu r(x_t)^{1+\rho} \qquad \text{(Stop)}$$

  $r_t(x_{t+1})$: forward-backward residual of (Proximal semismooth Newton), so it is 0 when x_{t+1} is an exact solution of that subproblem

Outline

Semismoothness

- It is known that when $B \equiv 0$, semismooth Newton that uses arbitrary elements of the generalized Jacobian can achieve superlinear convergence if $\nabla A(x)$ is nonsingular at the point of convergence

Definition 1 (Semismooth)

A is semismooth of order p at x if it is directionally differentiable at x, and for any $\nabla A(x + \Delta x) \in \partial(A(x + \Delta x))$ with $\Delta x \to 0$,

$$A(x + \Delta x) - A(x) - (\nabla A(x + \Delta x))\,\Delta x = O\left(\|\Delta x\|^{1+p}\right).$$

When $p = 1$, it is called strongly semismooth.

- A generalization of A differentiable with ∇A Hölder continuous of order p

Local Convergence

Theorem 2

Consider solving (GE) *using* (Proximal semismooth Newton) *and* (Stop), *with* A *single-valued and continuous,* B *maximal monotone, and* $\Omega \neq 0$. *Assume* (HEB) *holds for some* $q > 0$ *in a neighborhood* V *of* Ω *and* A *is locally Lipschitz continuous and semismooth of order* p *for some* $p \in (0, 1]$ *within the same neighborhood. If the following inequalities are satisfied:*

$$\begin{cases} (1+\rho)q & > 1, \\ (1+p)q & > 1, \\ \left(1+p-\frac{\rho}{q}\right)(1+p)q & > 1, \end{cases} \tag{Cond-Q}$$

we obtain Q*-superlinear convergence in* V *with the form*

$$r\left(x_{t+1}\right) = O\left(r(x_t)^s\right), \operatorname{dist}\left(x_{t+1}, \Omega\right) = O\left(\operatorname{dist}\left(x_t, \Omega\right)^s\right),$$
$$s := \min\left\{(1+\rho)q, (1+p)q, \left(1+p-\frac{\rho}{q}\right)(1+p)q\right\}.$$

Corollary 3

Consider the setting of Theorem 2. If instead of (Cond-Q), *the following inequalities hold:*

$$\begin{cases} (1+p)q & > 1, \\ \left(1+p-\frac{\rho}{q}\right)(1+p)q & > 1, \\ \rho+q & > 1, \\ \rho & > 0, \end{cases} \tag{Cond-R}$$

then we obtain Q-superlinear convergence within V for $\{r(x_t)\}$ of the form

$$r(x_{t+1}) = O\left(r(x_t)^{1+s_2}\right),$$

where

$$s_2 := \min\left\{(1+p)q, \left(1+p-\frac{\rho}{q}\right)(1+p)q, \rho+q, 1+\rho\right\} - 1 > 0,$$

and R-superlinear convergence within V for $\{\operatorname{dist}(x_t, \Omega)\}$:
$\lim_{t\to\infty} \operatorname{dist}(x_t, \Omega)^{\frac{1}{t}} = 0.$

Theorem 4

If the conditions in either Theorem 2 or Corollary 3 hold true and that the initial point x_0 is close enough to Ω, then $x_t \to x^*$ strongly for some $x^* \in \Omega$.[2]

[2] Note that we are considering Hilbert spaces.

Implications

(Cond-Q): $(1+\rho)q > 1, \quad (1+p)q > 1, \quad \left(1+p-\frac{\rho}{q}\right)(1+p)q > 1$

- $q > 1$: might get faster-than-quadratic rates; also ok to have $p = 0, \rho = 0$: constant damping but still with superlinear convergence
- Quadratic convergence: happens if $p = \rho = q = 1$, namely the case of the ordinary error bound plus strongly semismooth A.
- If $p = 1$ and $q \le 1$, the condition becomes

$$q > \frac{1}{2}, \quad 2q - \frac{1}{2} > \rho, \quad q > \frac{-1+\sqrt{33}}{8}.$$

  In particular, we can set $\rho = 2(-1+\sqrt{33})/8 - 1/2 = \left(-3+\sqrt{33}\right)/4$ to allow the widest range of q.
- If $q = 1$, then $p \ge \rho > 0$ implies superlinear convergence.

(Cond-R): $(1+p)q > 1, \left(1+p-\frac{\rho}{q}\right)(1+p)q > 1, \rho + q > 1, \rho > 0$

- When $p = 1$: superlinear convergence only requires $q > 1/2$ and $\rho = 1/2$
- Weaker than (Cond-Q) when $q \le 1$, but stronger when $q > 1$

Outline

Equivalence Between (HEB) and Sharpness

For (Reg-Opt), when f and Ψ are both convex, actually

$$\operatorname{dist}(x, \Omega) \leq r(x)^q$$

is equivalent to

$$F(x) - F^* \geq \kappa_2 \operatorname{dist}(x, \Omega)^{1+\frac{1}{q}} \qquad \text{(Sharpness)}$$

for some $\kappa_2 > 0$. This relates the objective distance and the iterate distance to the solutions

- $q > 1$ only possible when $\Psi \neq 0$
- For $q = \infty$, (Sharpness) is the so-called weak-sharp minima, and it is known that for problems with such a property, a solution in Ω can be obtained in a finite number of iterations for a wide range of algorithms
- We get a finite termination result for a broader range ($q > 1$) of problems and a very broad class of algorithms, and semismoothness is not needed in this result

Finite Termination

Theorem 5

Consider (Reg-Opt) *with* F *satisfying* (HEB) *for some* $q > 1$ *and some* $\kappa > 0$. *If* f *and* Ψ *are both convex, then any algorithm guaranteeing*

$$\liminf_{t \to \infty} r(x_t) = 0 \tag{1}$$

ensures that there is some $t_0 < \infty$ *such that* $r(x_{t_0}) = 0$ *and hence* $x_{t_0} \in \Omega$.

- Any convergent algorithm (such that a limit point of its iterates is a solution) attains finite termination in this case

Outline

Issues of Existing Approaches

- For globalization for (Reg-Opt), the approximate solution to the subproblem is called $\hat{x}_{t+1}$, which is just a candidate for the next iterate
- Some line search for ensuring global convergence
- Use a modified version of (Stop) to get descent directions:

$$q_t(\hat{x}_{t+1}) \leq q_t(x_t), \quad r_t(\hat{x}_{t+1}) \leq \nu r(x_t)^{1+\rho}, \qquad \text{(Stop')}$$

  where $r_t(\cdot)$ is the residual of the subproblem, and

$$q_t(x) := \langle \nabla f(x_t),\, x - x_t \rangle + \frac{1}{2} \langle J_t\,(x - x_t)\,,\, x - x_t \rangle + \Psi\,(x)$$

- Yue et al. (2019):[3] backtracking from $\alpha_t = 1$ until $F(x_t + \alpha_t(\hat{x}_{t+1} - x_t)) \leq F(x_t) - \gamma\alpha_t \|\hat{x}_{t+1} - x_t\|^2$
  - Requires Lipschitzian $\nabla^2 f$ (to use Taylor expansion for sufficient decrease)
  - Does not guarantee unit step size acceptance when $q = \rho = 1$
  - Only considered (Reg-Opt) and got superlinear convergence only for $q = 1$

[3] Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property. *Mathematical Programming*, 174(1-2):327–358, 2019

- Mordukhovich et al. (2022):[4] accept unit step size when $r(x)$ decreases at a prespecified linear rate; otherwise backtracking above
  - Again need $\nabla^2 f$ Lipschitz continuous, and only consider (Reg-Opt)
  - Guarantee unit step size eventually accepted if local superlinear convergence is present, but objective value might not be strictly decreasing
  - Prove Q-superlinear convergence for $r(x_t)$ and R-superlinear convergence for $\operatorname{dist}(x_t, \Omega)$ with conditions the same as (Cond-R)
  - Earlier preprint proved Q-superlinear convergence for both $r(x_t)$ and $\operatorname{dist}(x_t, \Omega)$ for $q > (-1 + \sqrt{5})/2 \approx 0.618$ with suitable ρ (ours is $q > (-1 + \sqrt{33})/8 \approx 0.593$)
  - Convergence faster than quadratic for $q > 1$ (we have seen that actually it is finite termination)

---

[4] Boris S. Mordukhovich, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A globally convergent proximal Newton-type method in nonsmooth convex optimization. *Mathematical Programming*, 2022.
Online first. The first version appears at arXiv:2011.08166v1 with some slightly different results

Our Algorithm

Algorithm 1: A Proximal-Newton Method Guaranteeing Strict Decrease and Superlinear Convergence for the Objective Value

input : $x_0 \in \mathcal{H}$, $\beta, \gamma \in (0,1)$, $\nu \in [0,1)$, $c > 0$, $\delta \geq 0$

Compute an upper bound L for the Lipschitz constant of ∇f

$\rho \leftarrow 1$

for $t = 0, 1, \ldots$ **do**

  Find an approximate solution $\hat{x}_{t+1}$ of the subproblem satisfying (Stop')

  $\alpha_t \leftarrow 1, \quad p_t \leftarrow \hat{x}_{t+1} - x_t$

  while *True* **do**

    $y_{t+1}(\alpha_t) \leftarrow x_t + \alpha_t p_t$

    $\bar{x}_{t+1}(\alpha_t) \leftarrow \mathrm{prox}_{\Psi/L}\left(y_{t+1}(\alpha_t) - \nabla f(y_{t+1}(\alpha_t))/L\right)$

    if $F(\bar{x}_{t+1}(\alpha_t)) \leq F(x_t) - \gamma \alpha_t^2 \|p_t\|^{2+\delta}$ **then**

      $x_{t+1} \leftarrow \bar{x}_{t+1}(\alpha_t)$

      Break

    else $\alpha_t \leftarrow \beta \alpha_t$

Global Convergence

Lemma 6

If f is Lipschitz-continuously differentiable (convexity not needed except for subproblem construction, but can work around this by using PSD approximations for the generalized Hessian) and Ψ is convex, closed, and proper, then we have for Algorithm 1 that

$$\lim_{t \to \infty} r(x_t) = 0.$$

Unit Step Size Acceptance and Superlinear Convergence

Theorem 7

Consider Algorithm 1 with the setting of Theorem 2 satisfied and f convex. If (Cond-Q) holds and δ is large enough such that $\|p_t\|^{2+\delta} = o(\mathrm{dist}(x_t, \Omega)^{(q+1)/q})$, then there is $t_0 \geq 0$ such that $\alpha_t = 1$ is accepted for all $t \geq t_0$, and

$$\begin{aligned} \mathrm{dist}(x_{t+1}, \Omega) &= O\left(\mathrm{dist}(x_t, \Omega)^{1+s}\right), \\ r(x_{t+1}) &= O\left(r(x_t)^{1+s}\right), \\ F(x_{t+1}) - F^* &= O\left(\left(F(x_{t+1}) - F^*\right)^{1+s}\right), \quad \forall t \geq t_0. \end{aligned}$$

- For the allowed range of Q-superlinear convergence, $\delta = 2$ is large enough
- The key for working around Lipschitzian Hessian is to use (Sharpness) and the objective bound from proximal gradient to get objective change guarantees

Simplification for Smooth Problems

- If $\Psi \equiv 0$: the bound described above directly comes from convexity of f
- Don't need another (proximal) gradient step and can keep all guarantees

Algorithm 2: A Simple semismooth Newton Method

input : $x_0 \in \mathcal{H}$, $\beta, \gamma \in (0,1)$, $\nu \in [0,1)$, $c > 0, \rho \in (0,1]$, $\delta \geq 0$
for $t = 0, 1, \ldots$ **do**
  Find $\hat{x}_{t+1}$ satisfying (Stop')
  $p_t \leftarrow \hat{x}_{t+1} - x_t$, $\alpha_t \leftarrow 1$
  while $F(x_t + \alpha_t p_t) > F(x_t) - \gamma \alpha_t^2 \|p_t\|^{2+\delta}$ **do** $\alpha_t \leftarrow \beta \alpha_t$
  $x_{t+1} \leftarrow x_t + \alpha_t p_t$

- All guarantees follow that for Algorithm 1 if $\Psi \equiv 0$
- If $\rho = 0.5$, f has Lipschitzian Hessian, and c is chosen according to the Lipschitz constant of $\nabla^2 f$: coincides with Mishchenko (2021); Doikov and Nesterov (2021), who mainly focus on global complexity without considering inexactness or considering when f is not twice-differentiable

Special case: Levenberg–Marquardt Method

- Our approximation of allowing $\|J_t - \nabla A(x_t)\| = O\left(r\left(x_t\right)^{\theta}\right)$ with $\theta \in [\rho, 1]$ covers the case of the Levenberg–Marquardt method when the nonlinear equation has at least one solution
- In this case the residual converges to 0 at a fast rate, so the error of the estimation of the Hessian using the Jacobian will satisfy this condition

Thanks for Listening

Questions?

References I

Nikita Doikov and Yurii Nesterov. Gradient regularization of newton method with bregman distances. Technical report, 2021. arXiv:2112.02952.

Jason D. Lee, Yuekai Sun, and Michael A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

Konstantin Mishchenko. Regularized Newton method with global $O(1/k^2)$ convergence. Technical report, 2021. arXiv:2112.02089.

Boris S. Mordukhovich, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A globally convergent proximal Newton-type method in nonsmooth convex optimization. *Mathematical Programming*, 2022. Online first. The first version appears at arXiv:2011.08166v1 with some slightly different results.

References II

Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property. *Mathematical Programming*, 174(1-2):327–358, 2019.