Newtonian Methods in Nonsmooth Optimization via the Lens of Variational Analysis

VO THANH PHAT
Assistant Professor - University of North Dakota, USA
(Based on the joint work with
Pham Duy Khanh, HCMC University of Education, Vietnam)
Boris Mordukhovich, Wayne State University, USA

26th Midwest Optimization Meeting
Workshop on Large Scale Optimization and Applications
Hosted by the University of Waterloo
November 8-9, 2024

OUTLINE

- Structured Nonconvex Optimization Problems
- Classical Newton's Method and Tools of Variational Analysis
- Coderivative-Based Newton Method for Structured Nonconvex Optimization Problems.
- Applications
- Future Investigation

Structured Nonconvex Optimization Problems

Structured Nonconvex Optimization Problems

Our target optimization problem is

$$\text{minimize} \quad \varphi(x) := f(x) + g(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^2$-smooth, and $g : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$ is a lower semicontinuous, and prox-bounded[1]

- Note that both f and g are generally nonconvex and nonsmooth. This makes (1) appropriate for applications to signal and image processing, machine learning, statistics, control, system identification, etc.
- If, in particular, g is the indicator function of a closed set, then (1) becomes a constrained optimization problems with numerous applications.

[1] *A function is prox-bounded if it is bounded from below by some quadratic function.*

Classical Newton's Method and Tools of Variational Analysis

Classical Newton's Method

Consider the gradient system

$$\nabla \varphi(x) = 0, \tag{2}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^2$-smooth function.

The classical version of Newton's method: $x^0 \in \mathbb{R}^n$ is given, then

$$x^{k+1} = x^k + d^k \quad \text{with } -\nabla \varphi(x^k) = \nabla^2 \varphi(x^k) d^k, \quad k = 0, 1, \ldots \tag{3}$$

This algorithm is locally well-defined and superlinearly converges to a solution $\bar{x}$ of (2) if $\nabla^2 \varphi(\bar{x})$ is nonsingular.[2]

[2] *Izmailov, A. F., & Solodov, M. V. (2014). Newton-type methods for optimization and variational problems (Vol. 1). New York: Springer.*

Requirement of generalized derivatives for generalized Newton method:

- Comprehensive calculus rules: sum rules, chain rules, product rules, etc.
- Explicit calculations in a number of settings important for applications.
- Can characterize the convexity, generalized convexity, local optimality, etc.

To develop Newton's method, we use the generalized derivatives including limiting first- and second-order subdifferentials introduced by Mordukhovich, which are presented in the books [3][4].

---

[3] *Mordukhovich, B. S. (2018). Variational analysis and applications (Vol. 30). Cham: Springer.*

[4] *Rockafellar, R. T., & Wets, R. J. B. (2009). Variational analysis (Vol. 317). Springer Science & Business Media.*

Generalized Differentiation

The normal cone to $\Omega \subset \mathbb{R}^n$ at $\bar{x} \in \Omega$ from

$$N_\Omega(\bar{x}) := \big\{ v \;\big|\; \exists\, x_k \to \bar{x},\; \alpha_k \ge 0,\; w_k \in \Pi_\Omega(x_k),\; \alpha_k(x_k - w_k) \to v \big\}$$

where Π_Ω stands for the Euclidean projection. The coderivative of $F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ at $(\bar{x}, \bar{y}) \in \operatorname{gph} F$

$$D^*F(\bar{x},\bar{y})(v) := \big\{ u \in \mathbb{R}^n \;\big|\; (u, -v) \in N_{\operatorname{gph} F}(\bar{x},\bar{y}) \big\}, \quad v \in \mathbb{R}^m.$$

When $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is $\mathcal{C}^1$-smooth, then

$$D^*F(\bar{x})(v) = \big\{ \nabla F(\bar{x})^* v \big\}, \quad v \in \mathbb{R}^n,$$

via the adjoint/transpose Jacobian matrix. The (first-order) subdifferential of $\varphi\colon \mathbb{R}^n \to (-\infty, \infty]$ at $\bar{x} \in \operatorname{dom} \varphi$

$$\partial\varphi(\bar{x}) := \big\{ v \in \mathbb{R}^n \;\big|\; (v, -1) \in N_{\operatorname{epi} \varphi}(\bar{x}, \varphi(\bar{x})) \big\}.$$

Generalized Differentiation - Second-order Subdifferential

Second-order subdifferential/generalized Hessian[5] of φ at $\bar{x}$ relative to $\bar{v} \in \partial\varphi(\bar{x})$ is

$$\partial^2\varphi(\bar{x},\bar{v})(u) := \big(D^*\partial\varphi\big)(\bar{x},\bar{v})(u), \quad u \in \mathbb{R}^n$$

If $\varphi \in \mathcal{C}^2$-smooth around $\bar{x}$, then

$$\partial^2\varphi(\bar{x},\bar{v})(u) = \big\{\nabla^2\varphi(\bar{x})u\big\}, \quad u \in \mathbb{R}^n$$

It is realized that the generalized Hessian $\partial^2\varphi$ enjoys well-developed second-order calculus and can be viewed as an appropriate replacement of the Hessian $\nabla^2\varphi$ for nonsmooth problems. $\partial^2\varphi$ is fully computed in terms of the given data for broad classes of problems in optimization and variational analysis.

[5] *Mordukhovich, B.S.: Sensitivity analysis in nonsmooth optimization. In: Field, D.A., Komkov, V.(eds) Theoretical Aspects of Industrial Design, 32–46. SIAM Proc. Appl. Math. 58. Philadelphia, PA (1992)*

Ideas

Goal: Approximate an optimal solution to the following optimization problem

$$\text{minimize} \quad \varphi(x) \quad \text{subject to} \quad x \in \mathbb{R}^n \tag{4}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ is not necessarily smooth.

Ideas:

- $\nabla\varphi(x^k) \longrightarrow [v^k \in \partial\varphi(x^k)]$.
- $\nabla^2\varphi(x^k) \longrightarrow \partial^2\varphi(x^k, v^k)$, for some $v^k \in \partial\varphi(x^k)$.

However, $\partial\varphi(x^k)$ may be empty. So, we will choose the pair $(\widehat{x}^k, \widehat{v}^k) \in \operatorname{gph}\partial\varphi$ such that $\partial\varphi(\widehat{x}^k)$ is nonempty in which $\widehat{x}^k$ is not far from x^k in some senses.

General Framework of Coderivative-Based Newton Method

Suppose that $\bar{x}$ is a stationary point, i.e., $0 \in \partial\varphi(\bar{x})$. Our generalized coderivative-based Newton method can be formulated as: $x^0 \in \mathbb{R}^n$ is given, then

$$x^{k+1} = \widehat{x}^k + d^k \quad \text{with } -\widehat{v}^k \in \partial^2\varphi(\widehat{x}^k, \widehat{v}^k)(d^k),\ \widehat{v}^k \in \partial\varphi(\widehat{x}^k),$$

where $(\widehat{x}^k, \widehat{v}^k)$ satisfying the following inequality

$$\|(\widehat{x}^k, \widehat{v}^k) - (x^k, 0)\| \le \eta \|x^k - \bar{x}\|. \tag{5}$$

The step in (5) is called the **approximate step** in our algorithm.

In fact, we can choose $(\widehat{x}^k, \widehat{v}^k)$ as an "approximate projection" of $(x^k, 0)$ on $\mathrm{gph}\partial\varphi$ in the sense that

$$\|(\widehat{x}^k, \widehat{v}^k) - (x^k, 0)\| \le \eta \mathrm{dist}((x^k, 0), \mathrm{gph}\partial\varphi).$$

Convergence Analysis of Coderivative-Based Newtonian Methods

Question: Do we guarantee the convergence of the iterative sequence generated by our aforementioned method[6]?

[6] $x^0 \in \mathbb{R}^n$ is given, then

$$x^{k+1} = \widehat{x}^k + d^k \quad \text{with } -\widehat{v}^k \in \partial^2\varphi(\widehat{x}^k, \widehat{v}^k)(d^k),\ \widehat{v}^k \in \partial\varphi(\widehat{x}^k), \tag{6}$$

where $k \in \mathbb{N}$.

Prox-regularity

Definition

$\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is prox-regular[a] at $\bar{x} \in \operatorname{dom}\varphi$ for $\bar{v} \in \partial\varphi(\bar{x})$ if φ is lower semicontinuous around $\bar{x}$ and there are $\varepsilon > 0$ and $\rho \geq 0$ such that for all $x \in \mathbb{B}_\varepsilon(\bar{x})$ with $\varphi(x) \leq \varphi(\bar{x}) + \varepsilon$ we have

$$\varphi(x) \geq \varphi(u) + \langle v, x-u\rangle - \frac{\rho}{2}\|x-u\|^2$$

for all $(u,v) \in (\operatorname{gph}\partial\varphi) \cap \mathbb{B}_\varepsilon(\bar{x},\bar{v})$.

[a] *Poliquin, R., & Rockafellar, R. (1996). Prox-regular functions in variational analysis. Transactions of the American Mathematical Society, 348(5), 1805-1838.*

φ is subdifferentially continuous at $\bar{x}$ for $\bar{v}$ if the convergence $(x_k, v_k) \to (\bar{x},\bar{v})$ with $v_k \in \partial\varphi(x_k)$ yields $\varphi(x_k) \to \varphi(\bar{x})$. If both properties hold, φ is continuously prox-regular. This is the major class in second-order variational analysis.

Semismoothness*

Definition

[a] A mapping $F\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is semismooth* at $(\bar{x}, \bar{y}) \in \operatorname{gph} F$ if whenever $(u, v) \in \mathbb{R}^n \times \mathbb{R}^m$ we have the equality

$$\langle u^*, u\rangle = \langle v^*, v\rangle \quad \text{for all} \quad (v^*, u^*) \in \operatorname{gph} D^*F\big((\bar{x}, \bar{y}); (u, v)\big).$$

[a] *Gfrerer, H., & Outrata, J. V. (2021). On a semismooth* Newton method for solving generalized equations. SIAM Journal on Optimization, 31(1), 489-517.*

Example:

(i) A continuously differentiable mapping is semismooth*.

(ii) A set-valued mapping with the graph represented as a union of finitely many closed and convex sets is semismooth*.

Convergence Analysis

- **Local Convergence**: Guarantee the local convergence to a local optimal solution $\bar{x}$ if φ is continuously prox-regular at $\bar{x}$ for 0, $\partial\varphi$ is semismooth* at $\bar{x}$, and $\bar{x}$ satisfies the second-order sufficient optimality condition in the sense that

$$0 \in \partial\varphi(\bar{x}) \quad \text{and} \quad \partial^2\varphi(\bar{x}, 0) \text{ is positive definite.}$$

- **Convergence Rate:** superlinear in the sense that

$$\lim_{k\to\infty} \|x^{k+1} - \bar{x}\| / \|x^k - \bar{x}\| = 0.$$

More detail in our work[7].

[7] *Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: Coderivative-based Newton methods in structured nonconvex and nonsmooth optimization, arXiv:2403.04262.*

Implementation of the Approximate Step

- When φ is a $\mathcal{C}^{1,1}$-smooth function, we can choose

$$\widehat{x}^k := x^k \text{ and } \widehat{v}^k := \nabla\varphi(x^k).$$

- When $\mathrm{Prox}_{\lambda\varphi}$[8] can be computed explicitly, we can choose

$$\widehat{x}^k := \mathrm{Prox}_{\lambda\varphi}(x^k) \quad \text{and} \quad \widehat{v}^k := \frac{1}{\lambda}\left(x^k - \mathrm{Prox}_{\lambda\varphi}(x^k)\right).$$

The natural question is how to implement the approximate step when $\varphi = f + g$, where f is $\mathcal{C}^2$-smooth, g is prox-bounded function? We will discuss in more detail in the next section.

[8] $\mathrm{Prox}_{\lambda\varphi}(x) = \mathrm{argmin}_{y\in\mathbb{R}^n}\{\varphi(y) + 1/(2\lambda)\|y - x\|^2\}$

Coderivative-Based Newton Method for Structured Nonconvex Optimization Problems

Coderivative-Based Newton Method for Structured Nonconvex Optimization Problems

Consider the problem

$$\text{minimize} \quad \varphi(x) := f(x) + g(x), \tag{7}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^2$-smooth, and $g : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$ is a lower semicontinuous, and prox-bounded function.

To use our coderivative-based Newton method to find a stationary point $\bar{x}$ to (7), i.e., $0 \in \partial\varphi(\bar{x})$, we need to clarify two following questions:

- How do we implement the **approximate step**?
- How do we guarantee the **global convergence** of the iterative sequence?

Approximate Step in Structured Nonconvex Optimization Problems

When $\varphi = f + g$, where f is $\mathcal{C}^2$-smooth, g is prox-bounded, we can choose

$$\widehat{x}^k \in \mathrm{Prox}_{\lambda g}(x^k - \lambda \nabla f(x^k))$$

and

$$\widehat{v}^k := \nabla f(\widehat{x}^k) - \nabla f(x^k) + \frac{1}{\lambda}(x^k - \widehat{x}^k).$$

In this case, we have

- $\widehat{v}^k \in \partial\varphi(\widehat{x}^k)$.
- There is $\eta > 0$ such that

$$\|(\widehat{x}^k, \widehat{v}^k) - (x^k, 0)\| \le \eta \|x^k - \bar{x}\|.$$

$\Longrightarrow$ We can apply our method to guarantee the **locally superlinear convergence** of $\{x^k\}$.

Globalization

To obtain the global convergence of our method for the nonconvex optimization problem, a natural approach is to consider the sequence $\{x^k\}$ as follows:

$$x^{k+1} := \widehat{x}^k + \tau_k d^k \tag{8}$$

with an appropriate stepsize selection $\tau_k \in (0, 1]$, with an expectation that the descent property holds

$$\varphi(x^{k+1}) = \varphi(\widehat{x}^k + \tau_k d^k) < \varphi(x^k), \quad k = 0, 1, \ldots.$$

However this is impossible to guarantee due to the discontinuity of the cost function φ.

Globalization

Fortunately, we can approximate the cost function φ by a differentiable function called forward-backward envelope[9,10,11], and we can guarantee the descent property of this function, i.e.,

$$\varphi_\lambda(x^{k+1}) = \varphi_\lambda(\widehat{x}^k + \tau_k d^k) < \varphi_\lambda(x^k), \quad k = 0, 1, \ldots,$$

where φ_λ is defined by

$$\varphi_\lambda(x) := \inf_{y \in \mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), y - x \rangle + g(y) + \frac{1}{2\lambda}\|y - x\|^2 \right\}, \tag{9}$$

[9] *Patrinos, P., & Bemporad, A. (2013, December). Proximal Newton methods for convex composite optimization. In 52nd IEEE Conference on Decision and Control (pp. 2358-2363). IEEE.*

[10] *Stella, L., Themelis, A., & Patrinos, P. (2017). Forward–backward quasi-Newton methods for nonsmooth optimization problems. Computational Optimization and Applications, 67(3), 443-487.*

[11] *Themelis, A., Stella, L., & Patrinos, P. (2018). Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. SIAM Journal on Optimization, 28(3), 2274-2303.*

Globalized Coderivative-Based Newton Method

$x^0 \in \mathbb{R}^n$ is given, then

$$x^{k+1} = \widehat{x}^k + \tau_k d^k \quad \text{with } -\widehat{v}^k \in \partial^2 \varphi(\widehat{x}^k, \widehat{v}^k)(d^k),$$

where

$$\widehat{x}^k \in \mathrm{Prox}_{\lambda g}(x^k - \lambda \nabla f(x^k)) \text{ and } \widehat{v}^k := \nabla f(\widehat{x}^k) - \nabla f(x^k) + \frac{1}{\lambda}(x^k - \widehat{x}^k)$$

and $\tau_k \in (0, 1]$ satisfying

$$\varphi_\lambda(\widehat{x}^k + \tau_k d^k) \le \varphi_\lambda(x^k) - \sigma \left\| \widehat{v}^k \right\|^2.$$

Convergence analysis of our method

- **Well-Posedness**: The sequence $\{x^k\}$ is well-defined. Both sequences $\{\widehat{v}^k\}$ and $\{\widehat{x}^k - x^k\}$ converge to 0 as $k \to \infty$. Finally, any accumulation point of $\{x^k\}$ is a stationary point.
- **Global Convergence:** Guarantee the global convergence of $\{x^k\}$ if g is continuously prox-regular at $\bar{x}$ for $-\nabla f(\bar{x})$, $\nabla^2 f$ is strictly differentiable at $\bar{x}$, ∂g is semismooth* at $\bar{x}$, and $\bar{x}$ satisfies the second-order sufficient optimality condition in the sense that

$$0 \in \partial\varphi(\bar{x}) \quad \text{and} \quad \partial^2\varphi(\bar{x}, 0) \text{ is positive definite.}$$

- **Convergence Rate:** superlinear in the sense that

$$\lim_{k\to\infty} \|x^{k+1} - \bar{x}\|/\|x^k - \bar{x}\| = 0.$$

Applications

Applications

Given an $m \times n$ matrix A and a vector $b \in \mathbb{R}^m$, the ℓ_0-ℓ_2 regularized least square regression problem whose importance has been well recognized in applications to practical models of statistics, machine learning, etc[12]. This problem is formulated as:

$$\min \ \ \varphi(x) := \frac{1}{2}\|Ax - b\|^2 + \mu_0\|x\|_0 + \mu_2\|x\|_2^2 \text{ subject to } x \in \mathbb{R}^n$$

where μ_0 and μ_2 are positive parameters, and where $\|x\|_0$ is the ℓ_0 norm of x counting the number of nonzero elements of x.

Our numerical experiment in [13] shows that our method behaves better than proximal gradient method for solving the above problem.

[12] *Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Operations Research, 68(5), 1517-1537.*

[13] *Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: Coderivative-based Newton methods in structured nonconvex and nonsmooth optimization, arXiv:2403.04262.*

Future Investigation

Future Investigation

- Establish the generalized Newton method for solving difference programming.
- Establish the generalized Newton method for solving multiobjective optimization problems.
- Establish the generalized Newton method for solving bilevel optimization problems.
- Establish the coderivative-based stochastic Newton method for solving nonsmooth and nonconvex optimization problems with high dimension.
- Applications to other important classes of models in data science, machine learning, statistic, and related disciplines.

Our works

- Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: *A generalized Newton method for subgradient systems.* **Math. Oper. Res.** 48, 1811–1845 (2022).
- Khanh, P.D., Mordukhovich, B.S., Phat, V.T., Tran, D. B.: *Generalized damped Newton algorithms in nonsmooth optimization via second-order subdifferentials.* **J. Global Optim.** 86, 93–122 (2023).
- Khanh, P.D., Mordukhovich, B.S., Phat, V.T., Tran, D. B.: *Globally convergent coderivative-based generalized Newton methods in nonsmooth optimization.* **Math. Program.** 205, 373–429 (2024).
- Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: *Coderivative-based Newton methods in structured nonconvex and nonsmooth optimization*, arXiv:2403.04262.

THANK YOU FOR YOUR ATTENTION