On some recent developments on Kurdyka-Łojasiewicz (KL) inequality

Guoyin Li
UNSW Sydney, Australia

26th Midwest Optimization Meeting (MOM26)

Base on joint work with B.S. Mordukhvoich, T.K. Pong, P. Yu and J. Zhu

Outline

1. Introduction on KL inequality and Motivations

Outline

1. Introduction on KL inequality and Motivations
2. Part I: An extended analysis framework
   - An abstract convergence framework
   - Interplay between generalized metric subregularity and KL property via strict saddle point condition
   - Applications to high-order regularization methods with momentum steps

Outline

1. Introduction on KL inequality and Motivations
2. Part I: An extended analysis framework
   - An abstract convergence framework
   - Interplay between generalized metric subregularity and KL property via strict saddle point condition
   - Applications to high-order regularization methods with momentum steps
3. Part II: Estimating the KL exponents

Outline

Motivation

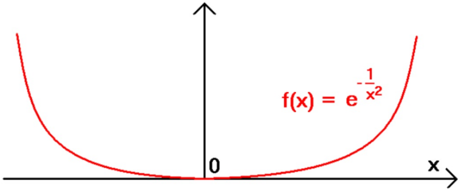
Our motivation starts with the KL property.

KL inequality

- (Łojasiewicz's gradient inequality, 1963) Let f be an analytic function on $\mathbb{R}^n$ with $\nabla f(\overline{x}) = 0$. Then, exists a rational number $\theta \in (0, 1]$ and $c, \delta > 0$ such that

$$\|\nabla f(x)\| \geq c|f(x) - f(\overline{x})|^{\theta} \text{ for all } x \text{ with } \|x - \overline{x}\| \leq \delta.$$

- This can fail for C^{∞} function, in general.



- Extended by Kurdyka to C^1 definable function. Further extended by Bolte, Daniilidis, Lewis to nonsmooth cases

KL Property and Convergence Analysis

Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be a proper lower l.s.c. function, and let $\vartheta : [0, \eta) \to \mathbb{R}_+$ be a continuous concave function with $\vartheta(0) = 0$, ϑ is continuously differentiable on $(0, \eta)$ and $\vartheta'(s) > 0$ for all $s \in (0, \eta)$.

Definition (KL property (Bolte, Daniilidis, Lewis, 07))

We say that f has the *Kurdyka-Łojasiewicz* (*KL*) *property* at $\overline{x}$ with respect to the desingularization function ϑ if there exists $\varepsilon > 0$ such that

$$\vartheta'(f(x) - f(\overline{x}))d(0, \partial f(x)) \geq 1$$

for all $x \in B_{\mathbb{R}^m}(\overline{x}, \varepsilon) \cap [f(\overline{x}) < f < f(\overline{x}) + \eta]$, where $d(\cdot, S)$ stands for the *distance function* associated with the set S.

- KL property is satisfied by a wide range of functions such as the semi-algebraic functions (e.g. Max/Min of finitely many polynomials).
- ∂f is the limiting subdifferential (cf. Mordukhovich).
- If $\vartheta(t) = c\, t^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$, reduces to the form of Łojasiewicz inequality.

If the desingularization function ϑ takes the form of $\vartheta(t) = c\,t^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$, then we say f satisfies the *KL property* at $\bar{x}$ with the *KL exponent* θ.

Prototypical result on convergence rate: Let $\{x_k\}$ be a bounded sequence generated by a descent algorithm with a potential function f. Let f be a KL function with exponent $\theta \in [0, 1)$. Then the following results hold (Attouch, Bolte, '09):

(i) If $\theta = 0$, then $\{x_k\}$ converges finitely.

(ii) If $\theta \in (0, \frac{1}{2}]$, then $\{x_k\}$ converges locally linearly.

(iii) If $\theta \in (\frac{1}{2}, 1)$, then $\{x_k\}$ converges locally sublinearly.

- These techniques has been widely used. E.g., in proximal type algorithms Attouch, Bolte, & Svaiter '13, Bolte, Sabach & Teboulle '14, Lewis & Drusvyatskiy '18, Boţ, Csetnek & Nguyen '19 and in Alternating direction method of multipliers (ADMM) and Douglas-Rachford algorithm L., Pong '15, '16.

An innocent looking example

Consider applying the standard proximal point method for $f(t) = |t|^{\frac{3}{2}}$.

- Iteration: $t_{k+1} = \mathrm{argmin}_{t \in \mathbb{R}}\{f(t) + \frac{\lambda}{2}(t - t_k)^2\}, \quad t_0 = 1$, where λ is a fixed positive parameter.
- Equivalent to

$$t_k = \frac{3}{2\lambda}(t_{k+1})^{\frac{1}{2}} + t_{k+1}.$$

- Simplifying this, and noting that $t_k \to 0$,

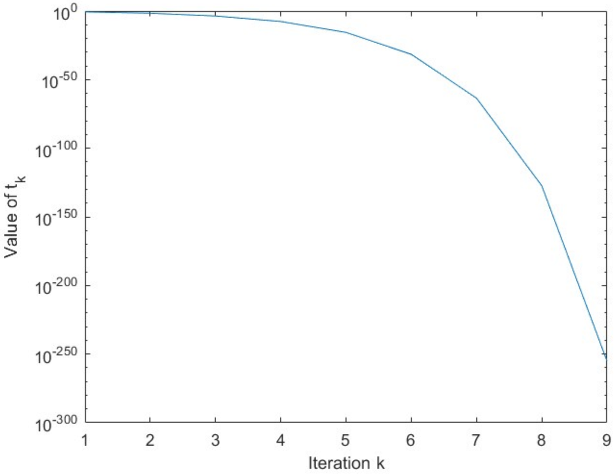
$$t_{k+1} = \left[\frac{t_k}{\frac{3}{4\lambda} + \sqrt{t_k + \frac{9}{16\lambda^2}}}\right]^2 = O(t_k^2),$$

An innocent looking example

Consider applying the standard proximal point method for $f(t) = |t|^{\frac{3}{2}}$.

- Iteration: $t_{k+1} = \mathrm{argmin}_{t\in\mathbb{R}}\{f(t) + \frac{\lambda}{2}(t - t_k)^2\}, \quad t_0 = 1$, where λ is a fixed positive parameter.
- Equivalent to
$$t_k = \frac{3}{2\lambda}(t_{k+1})^{\frac{1}{2}} + t_{k+1}.$$
- Simplifying this, and noting that $t_k \to 0$,
$$t_{k+1} = \left[\frac{t_k}{\frac{3}{4\lambda} + \sqrt{t_k + \frac{9}{16\lambda^2}}}\right]^2 = O(t_k^2),$$
- Quadratic convergence rate.

Illustration of the rate



Example Cont.

Consider applying the standard proximal point method for $f(t) = |t|^{\frac{3}{2}}$.

- Quadratic convergence rate.

Example Cont.

Consider applying the standard proximal point method for $f(t) = |t|^{\frac{3}{2}}$.

- Quadratic convergence rate.
- But, KL analysis only tells us the iterates converge in a linear rate.

Example Cont.

Consider applying the standard proximal point method for $f(t) = |t|^{\frac{3}{2}}$.

- Quadratic convergence rate.
- But, KL analysis only tells us the iterates converge in a linear rate.
- Question:

  Can we discuss superlinear/quadratic convergence within a suitable analysis framework (extending the KL framework)?

Newton type method

- Superlinear/quadratic convergence of Newton type methods have been studied by many researchers. A lot of exciting developments and progresses
  - Newton's method and Quasi Newton method
  - Nonsmooth Newton method
  - Regularized Newton method and many more.

Newton type method

- Superlinear/quadratic convergence of Newton type methods have been studied by many researchers. A lot of exciting developments and progresses
  - Newton's method and Quasi Newton method
  - Nonsmooth Newton method
  - Regularized Newton method and many more.
- A recent variant: Cubic regularization method (Nesterov & Polyak, 06)

Cubic regularization method

- Basic update: For a C^2-function f,

$$x_{k+1} \in \arg\min_{y \in \mathbb{R}^m} f_\sigma(y),$$

where

$$f_\sigma(y) = f(x_k) + \nabla f(x_k)^T(y - x) + \frac{1}{2}(y - x_k)^T \nabla^2 f(x_k)(y - x_k) + \frac{\sigma}{6}\|y - x_k\|^3,$$

- Subproblem can be solved via various techniques (convex optimization techniques, eigenvalue problem etc); Global Complexity.
- Quadratic convergence to a second-order stationary point was recently established under an error bound condition (Yue, Zhou, & So, 2019)

Error bound condition

- Error bound condition: there exist $\kappa, \rho > 0$ such that

$$d(x, \Theta) \leq \kappa \|\nabla f(x)\| \text{ for all } x \in \mathcal{N}(\Theta, \rho).$$

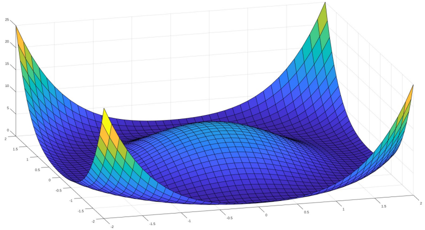
  where Θ is the collection of *second-order stationary points* of f.

$$\Theta := \{x \in \mathbb{R}^m \mid \nabla f(x) = 0, \ \nabla^2 f(x) \succeq 0\}.$$

  and $\mathcal{N}(\Theta, \rho) := \{x \in \mathbb{R}^m \mid d(x, \Theta) \leq \rho\}$.
- Was shown to be satisfied with phase retrieval problem and matrix completion problem with overwhelming probability.

Error bound condition cont.

- Can be satisfied in nonconvex and degeneracy case. E.g. $f(x) = (\|x\|^2 - r)^2$ with $r > 0$.



- $\nabla f(x) = 4(\|x\|^2 - r)x$ and $\nabla^2 f(x) = 8xx^T + 4(\|x\|^2 - r)I_m$;
- $\Gamma = \{x : \nabla f(x) = 0\} = \{x : \|x\| = \sqrt{r}\} \cup \{0\}$ and $\Theta = \{x : \|x\| = \sqrt{r}\}$

Error bound condition: there exist $\kappa, \rho > 0$ such that

$$d(x, \Theta) \leq \kappa \, d\big(0, \nabla f(x)\big) \text{ for all } x \in \mathcal{N}(\Theta, \rho).$$

where Θ is the collection of *second-order stationary points* of f.

- Has a similar form with metric subregularity but with subtle difference.
- Can we provide more simple verifiable sufficient conditions for this error bound condition (or its weaker variants)?

Main Questions:

- A framework for general descent methods (covering cubic regularization methods with momentums steps) so that superlinear/quadratic convergence can be identified?

Main Questions:

- A framework for general descent methods (covering cubic regularization methods with momentums steps) so that superlinear/quadratic convergence can be identified?
  Ans: Yes, and superlinear/quadratic convergence requires a generalized metric subregularity condition

Main Questions:

- A framework for general descent methods (covering cubic regularization methods with momentums steps) so that superlinear/quadratic convergence can be identified?
  Ans: Yes, and superlinear/quadratic convergence requires a generalized metric subregularity condition
- Simple verifiable sufficient conditions for this generalized metric subregularity condition?

Main Questions:

- A framework for general descent methods (covering cubic regularization methods with momentums steps) so that superlinear/quadratic convergence can be identified?
  Ans: Yes, and superlinear/quadratic convergence requires a generalized metric subregularity condition
- Simple verifiable sufficient conditions for this generalized metric subregularity condition?
  Ans: Yes, under the KL + strict saddle point conditions
- The convergence rate can be tied up with the KL exponents. Can we estimate these exponents?
  Ans: Yes, one approach is to exploit the underlying polynomial or conic structure.
- How sharp are the derived convergence rates?
  Ans: There are cases where the rates are indeed attained.

Part I: An extended analysis framework

In this part, we

- discuss an abstract framework for general descent methods so that superlinear convergence can be identified under a generalized metric subregularity condition
- link the generalized metric subregularity condition with KL condition via the strict saddle point conditions
- apply it to high-order regularization methods with momentum steps.

Based on: G. Li, B.S. Mordukhovich and J. Zhu, Generalized metric subregularity with applications to high-order regularized Newton methods, preprint, 2024.

Metric subregularity for subdifferential mapping

- Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be a proper l.s.c. function;
- Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be an admissible function, that is, $\psi(t) \to 0 \Rightarrow t \to 0$
- Given a target set $\Omega \subseteq \Gamma = \{x : 0 \in \partial f(x)\}$ and $\overline{x} \in \Omega$.

Definition

(i) The subdifferential ∂f satisfies the *(pointwise) generalized metric subregularity property* with respect to (ψ, Ω) at $\overline{x}$ if there exist $\kappa, \delta \in (0, \infty)$ such that

$$\psi(d(x, \Omega)) \leq \kappa\, d(0, \partial f(x)) \quad \text{for all} \quad x \in B_{\mathbb{R}^m}(\overline{x}, \delta).$$

(ii) The subdifferential ∂f satisfies the *uniform generalized metric subregularity property* with respect to (ψ, Ω) if there exist $\kappa, \rho \in (0, \infty)$ such that the above inequality holds for all $x \in \mathcal{N}(\Omega, \rho) = \{x \in \mathbb{R}^m \mid d(x, \Omega) \leq \rho\}$.

Comments and Illustrative Examples

Recall that the subdifferential ∂f satisfies the *(pointwise) generalized metric subregularity property* with respect to (ψ, Ω) at $\overline{x}$ if there exist $\kappa, \delta \in (0, \infty)$ such that

$$\psi\big(d(x,\Omega)\big) \leq \kappa\, d\big(0, \partial f(x)\big) \quad \text{for all} \quad x \in B_{\mathbb{R}^m}(\overline{x}, \delta).$$

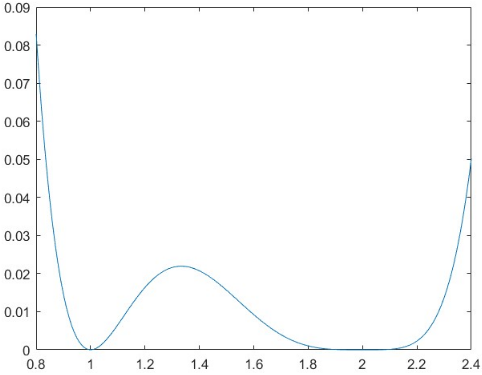
- if $\psi(t) = t$ & $\Omega = \Gamma \rightsquigarrow$ usual metric subreg. (cf. Dontchev, Rockafellar, 2009)
- if $\psi(t) = t^p$ with $p > 1$ & $\Omega = \Gamma \rightsquigarrow$ Hölder metric subreg. (Ahookhosh, Aragón-Artacho, Fleming 2019; Kruger 2015; L., Mordukhovich 2012);
- if $\psi(t) = t^p$ with $p \in (0, 1)$ & $\Omega = \Gamma \rightsquigarrow$ high-order metric subreg. (Mordukhovich, Ouyoung, 2015);
- $\exists$ cases where ψ is not of exponent type (e.g. exponential cone program) Lindstrom, Lourenço, Pong, 2023.

Recall that the subdifferential ∂f satisfies the *uniform generalized metric subregularity property* with respect to (ψ, Ω) if there exist $\kappa, \rho \in (0, \infty)$ such that

$$\psi\big(d(x, \Omega)\big) \le \kappa\, d\big(0, \partial f(x)\big) \quad \text{for all} \quad x \in \mathcal{N}(\Omega, \rho).$$

- If $\psi(t) = t$ & $\Omega = \Theta \rightsquigarrow$ the error bound condition.
- Generally, is strictly stronger than the pointwise version for the same (ψ, Ω). Sometimes, can fail to identify the quadratic convergence rate.

For example, $f(x) := (x-1)^2(x-2)^4$ and $\Omega = \Theta = \{1, 2\}$. Cubic regularization method with initial point $x_0 = 0.5$ leads to quadratic convergence to the point 1. Note that the error bound condition fails while pointwise metric subreg. holds at 1.



Outline

Descent method at large

Consider a couple sequence $\{(x_k, e_k)\} \subseteq \mathbb{R}^m \times \mathbb{R}_+$ generated by some algorithms such that

(i) *Surrogate condition*: there exists $c > 0$ such that

$$\|x_{k+1} - x_k\| \le c\, e_k \quad \text{for all} \quad k \in \mathbb{N} \tag{H0}$$

(ii) *Descent condition*:

$$f(x_{k+1}) + a\,\varphi(e_k) \le f(x_k) \tag{H1}$$

where $a > 0$ and φ is an admissible function.

(iii) *Relative error condition*:

$$\exists\ w_{k+1} \in \partial f(x_{k+1}) \ \text{ such that } \ \|w_{k+1}\| \le b\,\beta(e_k), \tag{H2}$$

where b is a fixed positive constant, and $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ is an admissible function.

The framework is flexible. E.g.,

- For many existing descent algorithms, the construction of the algorithm satisfies

$$f(x_{k+1})+a\|x_{k+1}-x_k\|^2 \leq f(x_k) \text{ and } \|\nabla f(x_{k+1})\| \leq \beta\|x_{k+1}-x_k\|$$

  So, $\varphi(t)=t^2$, $\beta(t)=t$ and $e_k=\|x_{k+1}-x_k\|$;

- For cubic regularization method, $\varphi(t)=t^3$, $\beta(t)=t^2$ and $e_k=\|x_{k+1}-x_k\|$;

- Having e_k helps to deal with momentum steps.

Abstract convergence result – a glimpse

- $\xi : [0,\eta) \to \mathbb{R}_+$ is a nondecreasing continuous function with $\xi(0) = 0$ for some $\eta > 0$.
- $\overline{x} \in \Omega$ is a cluster point of x_k, Ω is some (target) set.
- Denote $\Lambda_{k,k+1} := \xi(f(x_k) - f(\overline{x})) - \xi(f(x_{k+1}) - f(\overline{x}))$.

Key Recurrence Inequality: Consider the case where *the surrogate sequence of successive change grows mildly*, i.e., there exist $\ell_1 \in [0,1), \ell_2, \ell_3 \in [0,\infty)$ such that

$$e_k \leq \underbrace{\ell_1 e_{k-1} + \ell_2 \Lambda_{k,k+1}}_{\text{Appeared in KL Analysis}} + \underbrace{\ell_3 d(x_k,\Omega)}_{\text{New term}} \quad \text{for all large } k.$$

Key Recurrence Inequality: There exist $\ell_1 \in [0,1), \ell_2, \ell_3 \in [0,\infty)$ such that

$$e_k \leq \underbrace{\ell_1 e_{k-1} + \ell_2 \Lambda_{k,k+1}}_{\text{Appeared in KL Analysis}} + \underbrace{\ell_3 d(x_k,\Omega)}_{\text{New term}} \quad \text{for all large } k,$$

where Ω is some (target) set.

- Convergence. Let $s_k = \ell_3 d(x_k, \Omega)$. If s_k asymptotically shrinks *, then x_k converges towards a point in the target set Ω;
- Sublinear/linear convergence can be deduced similar as in KL analysis;

What about superlinear convergence?

*A sequence is called asymptotically shrinking if $s_k \leq \tau(s_{k-1})$ where τ satisfies $\limsup_{t\to 0^+} \sum_{n=0}^{\infty} \frac{\tau^n(t)}{t} < \infty$.

Superlinear convergence

Superlinear convergence

- under (pointwise) generalized metric subregularity with respect to (ψ, Ω), rate explicitly depends on ψ, φ and β; [†]

<u>Comments:</u>

- For the previous example, $f(t) = |t|^{\frac{3}{2}}$, generalized metric subregularity holds at 0 with $\psi(t) = t^{1/2} \rightsquigarrow$ quadratic convergence rate.
- For cubic regularization methods with momentum steps, $\rightsquigarrow$ quadratic convergence rate under (pointwise) metric subregularity w.r.t. $\Omega = \Theta$.

[†] it is possible to derive superlinear convergence rate under the assumption of KL property with growth control of the desingularization function ϑ, rate explicitly depends on ϑ, φ and β. But the derived rate is weaker.

Outline

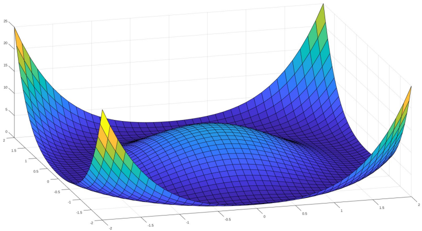
Sufficient conditions

An important question is: For a C^2-function f, how to check the generalized (pointwise) metric subregularity condition, when the target set is the set of second-order stationary points of f?

Here, we provide one possible way in connecting to KL property.

Motivating Example

Consider $f(x) = (\|x\|^2 - r)^2$ with $r > 0$.



- $\nabla f(x) = 4(\|x\|^2 - r)x$ and $\nabla^2 f(x) = 8xx^T + 4(\|x\|^2 - r)I_m$;
- $\Gamma = \{x \mid \nabla f(x) = 0\} = \{x : \|x\| = \sqrt{r}\} \cup \{0\}$ and $\Theta = \{x \mid \|x\| = \sqrt{r}\}$

What do we observe here?

- $\Gamma \neq \Theta$.
- But $d(x, \Gamma) = d(x, \Theta)$ for any x in a small neighborhood of $\overline{x} \in \Theta$.

A useful lemma

Lemma

Given a $\mathcal{C}^2$-smooth function $f : \mathbb{R}^m \to \mathbb{R}$ and $\overline{x} \in \Theta$. Suppose that both the KL property and strict saddle point property holds at $\overline{x}$. Then, there exists $\gamma > 0$ such that

$$d(x, \Theta) = d(x, \Gamma) \text{ for all } x \in B_{\mathbb{R}^m}(\overline{x}, \gamma). \tag{3.0}$$

- Strict saddle point property at $\overline{x} \in \Gamma$: if $\overline{x}$ is either a local minimizer for f, or a strict saddle point for f (i.e., $\lambda_{\min}(\nabla^2 f(\overline{x})) < 0$.
- KL property can be replaced by the more general weak separation property (WSP) at $\overline{x} \in \Gamma$ in the paper (which covers the convex composite cases under regularity)
- Generalized metric subregularity w.r.t. Θ can be deduced under KL + strict saddle point property.

Classes with explicit generalized metric subreguarity

The results can be used to determine explicit generalized metric subreguarity such as

- Over-parameterized compressive sensing models
- Rank-one matrix/tensor approximation
- Generalized phase retrieval problems.

We illustrate the first class below.

Consider the least squares problem with ℓ_1*-regularization*

$$\min_{x \in \mathbb{R}^m} \|Ax - b\|^2 + \nu \|x\|_1,$$

where $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$, $\nu > 0$, and $\|\cdot\|_1$ is the usual ℓ_1-norm.

Example (Over-parameterization model)

A recent interesting way to solve this problem is to transform it into an *equivalent smooth problem (e.g. Poon & Peyré, MP, 2023)*

$$\min_{(u,v) \in \mathbb{R}^m \times \mathbb{R}^m} f_{OP}(u, v) := \|A(u \circ v) - b\|^2 + \frac{\nu}{2}(\|u\|^2 + \|v\|^2),$$

where $u \circ v$ is the Hadamard (entrywise) product between the vector u and v in the sense that $(u \circ v)_i := u_i v_i$, $i = 1, \ldots, m$.

For the problem,

$$\min_{x=(u,v)\in\mathbb{R}^m\times\mathbb{R}^m} f_{OP}(u,v) := \|A(u\circ v)-b\|^2+\frac{\nu}{2}(\|u\|^2+\|v\|^2),$$

f_{OP} satisfies generalized metric subregularity at $\bar{x}\in\Theta$ w.r.t (ψ,Θ), where Θ is the set of 2nd-order stationary pts. [‡]

- Under strict complementarity condition (SCC) at $\bar{x}$, [§] $\psi(t)=t$;
- Otherwise, $\psi(t)=t^3$.

As an illustration of the idea, it can be proved by seeing

- f_{OP} is C^2, and it satisfies a (stronger version of) strict saddle point property (e.g. Poon & Peyré, 2023);
- Identifying the KL exponent for f_{OP} depending on whether strict complementarity condition holds.

[‡] The result can be extended to the case when the least squares loss $\|Ax-b\|^2$ is replaced by $g(Ax)$ where g is a C^2-strongly convex function.

[§] SCC: $0\in 2A^T(A\bar{x}-b)+\mathrm{ri}\left(\nu\,\partial\|\cdot\|_1(\bar{x})\right)$,

Outline

Application to high-order regularization methods

We now discuss the convergence rate analysis for high-order regularization methods

Basic Assumptions:

- f is $\mathcal{C}^2$-smooth and bounded below.
- $\mathcal{L}(f(x_0)) \subseteq \mathcal{F}$ for some compact convex set $\mathcal{F}$.
- ∇f is Lipschitz continuous with modulus $L_1 > 0$ on $\mathcal{F}$, and the Hessian of f is Hölder-continuous on $\mathcal{F}$ with exponent q, [¶] i.e., $L_2 > 0$ and $q \in (0, 1]$ such that

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_2 \|x - y\|^q \text{ for all } x, y \in \mathcal{F}.$$

[¶] The case where the Hessian of f is Hölder-continuous was considered e.g. in Grapigla & Nesterov, 2017.

Algorithm 1 Regularization method with momentum [‖]

1: **Input:** $x_0 = \widehat{x}_0 \in \mathbb{R}^m$, $\overline{\sigma} \in \left(\frac{2L_2}{q+2}, L_2\right]$ and $\zeta \in [0,1)$.
2: **for** $k = 0, 1, \ldots$ **do**
3: **Regularization step:** Choose $\sigma_k \in [\overline{\sigma}, 2L_2]$ and find

$$\widehat{x}_{k+1} \in \arg\min_{y \in \mathbb{R}^m} f_{\sigma_k}(x_k).^{**} \tag{3.0}$$

4: **Momentum step:**

$$\beta_{k+1} = \min\left\{\zeta, \|\nabla f(\widehat{x}_{k+1})\|, \|\widehat{x}_{k+1} - x_k\|\right\},$$

$$\widetilde{x}_{k+1} = \widehat{x}_{k+1} + \beta_{k+1}(\widehat{x}_{k+1} - \widehat{x}_k).$$

5: **Monotone step:** $x_{k+1} = \arg\min_{x \in \{\widehat{x}_{k+1}, \widetilde{x}_{k+1}\}} f(x)$.
6: **end for**

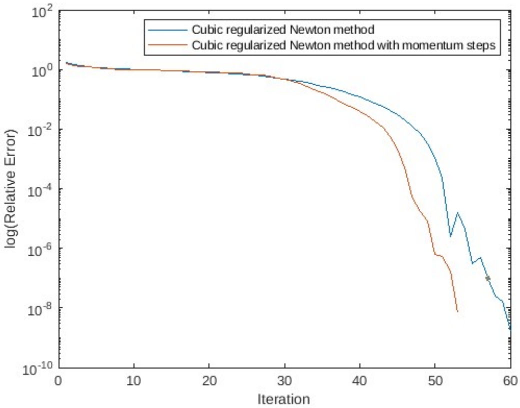
‖ In the case $q = 1$, has been considered in Lan et. al. 22 in convex cases and with complexity guarantees.

** Here, we have

$$f_\sigma(y) = f(x_k) + \nabla f(x_k)^T(y - x) + \frac{1}{2}(y - x_k)^T \nabla^2 f(x_k)(y - x_k) + \frac{\sigma}{(q+1)(q+2)}\|y - x_k\|^{q+2}..$$

Why momentum steps?

Illustrating cubic regularization method vs Algorithm 1 with momentum parameter $\zeta = 0.1$.



Matrix completion problem

Superlinear Convergence results

Apply Algorithm 1 for a C^2-function f whose Hessian is qth-order Hölder continuous. ††

Proposition

Suppose that there exists $\eta > 0$ such that the generalized metric subregularity condition holds with respect to (ψ, Θ), i.e.,

$$\psi(d(x,\Theta)) \le \|\nabla f(x)\| \ \text{ for all } \ x \in B_{\mathbb{R}^m}(\overline{x}, \eta)$$

and $\tau(t)/t \to 0$ with $\tau(t) = \psi^{-1}(Ct^{q+1})$ for some $C > 0$. Then, the sequence $\{x_k\}$ generated converges to $\overline{x} \in \Theta$ at least superlinearly with the rate

$$\limsup_{k\to\infty} \frac{\|x_k - \overline{x}\|}{\tau(\|x_{k-1} - \overline{x}\|)} < \infty.$$

†† Sublinear/linear convergence can also be discussed

Over-parameterized models

Consider the ℓ_1-regularization model and the associated over-parameterized smooth optimization problem

$$\min_{x=(u,v)\in\mathbb{R}^m\times\mathbb{R}^m} f_{OP}(u,v) := \|A(u\circ v)-b\|^2+\frac{\nu}{2}(\|u\|^2+\|v\|^2),$$

Corollary

The iterative sequence $\{x_k\}$ of Algorithm 1 converges to a global minimizer $\overline{x}$ of (OP), and

(i) *Under the strict complementary condition, $\{x_k\}$ converges to $\overline{x}$ in a quadratic rate, i.e., $\limsup_{k\to\infty}\frac{\|x_k-\overline{x}\|}{\|x_{k-1}-\overline{x}\|^2}<\infty$.*

(ii) *If the strict complementary condition fails, then $\{x_k\}$ converges to $\overline{x}$ with a sublinear rate $O(k^{-2})$.*

Part II: Estimating KL exponents

We have seen the KL exponents (if they exist) give us concrete information on the (asymptotic) convergence rates. How to estimate these exponents for general nonsmooth & nonconvex functions in general?

One possible strategy:

- Lift and project approach, then exploit the underlying polynomial structure or conic structure (such as semi-definite representability and C^2-cone structure)

Based on: P. Yu, G. Li and T.K. Pong, Kurdyka-Łojasiewicz exponent via inf-projection, FOCM 2022, arXiv:1902.03635,

Why polynomial or conic structure?

- Problems with polynomial or conic structures are ubiquitous.
- Many useful tools/concepts potentially can be used e.g. facial structure and singular degree for conic optimization (Borwein & Wolkowicz; Drusvyatskiy & L. & Wolkowicz; Sturm; Lourenco; Pataki; Roshchina & Tunçel), semi-algebraic geometry (Bochnak & Coste & Roy) etc.

Lift and project approach via inf-projection

We call the function $f(x) := \inf_{y \in \mathbb{Y}} F(x, y)$ for $x \in \mathbb{X}$ an inf-projection of F.

- The strict epigraph of f, defined as $\{(x, r) \in \mathbb{X} \times \mathbb{R} : \; f(x) < r\}$, is equal to the projection of the strict epigraph of F onto $\mathbb{X} \times \mathbb{R}$.
- Arises naturally in studying sensitivity analysis as value function.
- Used frequently in characterizing complicated functions via optimal value of conic programs.

Lemma (KL exponent via inf-projection Yu, L. Pong, 2022)

Let $F : \mathbb{X} \times \mathbb{Y} \to \mathbb{R} \cup \{\infty\}$ be a proper closed function and define $f(x) := \inf_{y \in \mathbb{Y}} F(x, y)$ and $Y(x) := \mathrm{Argmin}_{y \in \mathbb{Y}} F(x, y)$ for $x \in \mathbb{X}$. Let $\bar{x} \in \mathrm{dom}\, \partial f$. Suppose that

1. *It holds that $\partial F(\bar{x}, \bar{y}) \neq \emptyset$ for all $\bar{y} \in Y(\bar{x})$.*
2. *F is level-bounded in y locally uniformly in x.*
3. *The function F satisfies the KL property with exponent $\alpha \in [0, 1)$ at every point in $\{\bar{x}\} \times Y(\bar{x})$.*

Then f satisfies the KL property at $\bar{x}$ with exponent α.

Note: F is level-bounded in y locally uniformly in x means for any x and $\beta \in \mathbb{R}$, there exists $\rho > 0$ such that

$$\{(u, y) : \|u - x\| \leq \rho, F(u, y) \leq \beta\}$$

is bounded

LMI-representable functions

Definition

We say f is LMI-representable if there exists $d > 0$ and matrices $\{A_{00}, A_0, A_1, \ldots, A_n\} \subset \mathcal{S}^{d_i}$ such that

$$\mathrm{epi}\, f = \left\{ (x,t) \in \mathbb{R}^n \times \mathbb{R} : \; A_{00} + \sum_{j=1}^{n} A_j x_j + A_0 t \succeq 0 \right\}.$$

Examples of LMI representable functions: ℓ_1-norm, ℓ_2-norm, convex quadratic functions and indicator function of second-order cone.

Theorem (Sum of LMI-representable functions)

Let $f = \sum_{i=1}^{m} f_i$, where each $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a proper closed function which is LMI-representable. Suppose that

- *Strict feasibility condition is satisfied for the LMI representation;*
- *Strict complementarity condition holds, $0 \in \mathrm{ri}\, \partial f(\bar{x})$.*

Then f satisfies the KL property at $\bar{x}$ with exponent $\frac{1}{2}$.

Idea of the proof:

- Write $f(x) = \inf_{(s,t)} F(x,s,t)$ with $F(x,s,t) = t + \delta_D(x,s,t)$ where $D = \{(x,s,t) : t \geq \sum_{i=1}^{m} s_i, s_i \geq f_i(x)\}$ is a set described by semi-definite constraints.
- Argue the resulting semi-definite program has singular degree one, then apply error bound result in SDP and inf-projection theorem.

Explicit examples

Each of the following functions satisfies the KL property with exponent $\frac{1}{2}$ at an $\bar{x}$ satisfying $0 \in \mathrm{ri}\,\partial f(\bar{x})$:

1. **Group Lasso with overlapping blocks of variables:**

$$f(x) = \frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^{s} w_i \|x_{J_i}\|,$$

where $b \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times n}$, $\bigcup_{i=1}^{s} J_i = \{1, \ldots, n\}$, x_{J_i} is the subvector of x indexed by J_i, and $w_i \geq 0$, $i = 1, \ldots, s$.

2. **Group fused Lasso (Alaíz etal, 2013):**

$$f(x) = \frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^{s} w_i \|x_{J_i}\| + \sum_{i=2}^{s} \nu_i \|x_{J_i} - x_{J_{i-1}}\|,$$

where $b \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times rs}$, J_i is an equi-partition of $\{1, \ldots, n\}$ in the sense that $\bigcup_{i=1}^{s} J_i = \{1, \ldots, n\}$, $J_i \cap J_j = \emptyset$ and $|J_i| = |J_j| = r$ for $i \neq j$, $w_i, \nu_i \geq 0$, $i = 1, \ldots, s$.

Nuclear norm regularization

Similar strategy can be applied for the model problem

$$f(X) := \sum_{k=1}^{p} f_k(X) + \tau \|X\|_*, \tag{4.0}$$

where $X \in \mathbb{R}^{m\times n}$, $\|X\|_*$ denotes the nuclear norm of X (the sum of all singular values of X) and each $f_k : \mathbb{R}^{m\times n} \to \mathbb{R} \cup \{\infty\}$ is a proper closed LMI-representable function.

We do this by using the SDP representation (Rechet, Fazel & Parrilo, 2010)

$$\|X\|_* = \frac{1}{2} \inf_{U,V} \left\{ \mathrm{tr}(U) + \mathrm{tr}(V) : \begin{bmatrix} U & X \\ X^T & V \end{bmatrix} \succeq 0,\ U \in \mathcal{S}^m, V \in \mathcal{S}^n \right\}$$

Theorem (Nuclear norm regularization, Yu, L. Pong, 2022)

Let $f(X) = \sum_{i=1}^{m} f_i(X) + \tau\|X\|_*$ with each f_i is LMI-representable. Suppose that

- *Strict feasibility condition is satisfied for each of the LMI representation;*
- *Strict complementarity condition holds, $0 \in \mathrm{ri}\, \partial f(\bar{x})$.*

Then f satisfies the KL property at $\bar{X}$ with exponent $\frac{1}{2}$.

Note: In the case $m = 1$ and $f_1(X) = \frac{1}{2}\|\mathcal{A}X - b\|^2$, this can be derived using the error bound result in Zhou & So 2017 under the strict complementarity condition.

Beyond semi-algebraic structure: C^2-cone reduciblity

Definition (Shapiro, 2003)

A closed set $\mathfrak{D} \subseteq \mathbb{X}$ is said to be

- C^2-cone reducible at $\bar{w} \in \mathfrak{D}$ if $\exists$ a closed convex pointed cone $K \subseteq \mathbb{Y}$, $\rho > 0$ and a mapping $\Theta : \mathbb{X} \to \mathbb{Y}$ such that
  - (1) Θ is twice continuously differentiable in $B(\bar{w}, \rho)$;
  - (2) $\Theta(\bar{w}) = 0$ and $D\Theta(\bar{w}) : \mathbb{X} \to \mathbb{Y}$ is onto,
  - (3) $\mathfrak{D} \cap B(\bar{w}, \rho) = \{w : \Theta(w) \in K\} \cap B(\bar{w}, \rho)$.
- C^2-cone reducible if $\mathfrak{D}$ is C^2-cone reducible at $\bar{w}$ for all $\bar{w} \in \mathfrak{D}$.

Examples:

- Polyhedra, second order cone, positive semi-definite cone.
- $\mathfrak{D} = \{w : g_i(w) \leq 0, i = 1, \ldots, m\}$, $g_i \in C^2$, LICQ holds at $\bar{w} \in \mathfrak{D}$ implies that $\mathfrak{D}$ is C^2-cone reducible at $\bar{w}$.

Theorem

Let $\ell : \mathbb{Y} \to \mathbb{R}$ be a function that is strongly convex on any compact convex set and has locally Lipschitz gradient, $\mathcal{A} : \mathbb{X} \to \mathbb{Y}$ be a linear map, and $v \in \mathbb{X}$. Consider the function

$$f(x) := \ell(\mathcal{A}x) + \langle v, x\rangle + \sigma_{\mathfrak{D}}(x)$$

with $\mathfrak{D}$ being a C^2-cone reducible closed convex set. Suppose that

$$\mathcal{A}^{-1}\{\mathcal{A}\bar{x}\} \cap \mathrm{ri} N_{\mathfrak{D}}(-\mathcal{A}^*\nabla\ell(\mathcal{A}\bar{x}) - v) \neq \emptyset.$$

Then f satisfies the KL property at $\bar{x}$ with exponent $\frac{1}{2}$.

Note: The ri condition can be dropped if $N_{\mathfrak{D}}(\cdot)$ is a polyhedral set.

Explicit examples

Let $\ell : \mathbb{R}^m \to \mathbb{R}$ be strongly convex on any compact convex set and have locally Lipschitz gradient, $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ be linear.

Each of the following functions satisfies the KL property with exponent $\frac{1}{2}$ at an $\bar{X}$ satisfying the ri condition

- **(PSD cone constraint)**

$$f(X) = \ell(\mathcal{A}X) + \langle V, X\rangle + \delta_{\mathcal{S}^n_+}(X)$$

- **(Schatten p-norm regularization)**

$$f(X) = \ell(\mathcal{A}X) + \langle V, X\rangle + \tau\|X\|_p \quad \text{for all } X \in \mathcal{S}^n,$$

  where $p \in [1, 2] \cup \{+\infty\}$and $\|X\|_p$ is the Schatten p-norm.
- Problems with **entropy regularization.**

One can also leverage polynomial structure.

- A convex piecewise polynomial function of degree at most $d \geq 2$ on $\mathbb{R}^n$ is a KL function with exponent $1 - \frac{1}{(d-1)^n+1}$ (Bolte et al. 2015)
- (Gwoździewicz 1999 and Kollar 2002) If f is a polynomial with degree d and 0 is a strict local minimizer, then, KL exponent $\tau = 1 - \frac{1}{(d-1)^n+1}$;
- Dropping the strict minimizer assumption in Gwoździewicz's result, we have a new estimate of KL exponent $\tau = 1 - R(n,d)^{-1} = 1 - \frac{1}{d(3d-3)^{n-1}}$ (Kurdyka 2012, and L., Mordukhovich and Pham 2015).

These approaches also allow us to consider other models such as

- (Least squares with rank constraint)

$$f(X) = \frac{1}{2}\|\mathcal{A}X - b\|^2 + \delta_{\mathrm{rank}(\cdot)\leq r}(X)$$

  for $X \in \mathbb{R}^{m\times n}$, $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$.

- (Sparse generalized eigenvalue problem)

$$f(x) = \frac{x^T Ax}{x^T Bx} + \delta_{\|\cdot\|=1}(x) + \lambda\|x\|_0$$

  for $A, B \in S^n$, B is positive definite.

Conclusions and future work

Conclusions

- Discuss two aspects of KL property: usage for superlinear convergence analysis & identifying the KL exponents
- A form of generalized metric subregularity w.r.t to target set places a role in identifying the superlinear convergence.
- Some sufficient conditions are provided for generalized metric subregularity w.r.t 2nd-order stationary pts via KL property + strict saddle point conditions
- One approach in estimating the KL exponents: Lift and project approach, then exploit polynomial or conic structure.

Future work:

- Verifiable sufficient conditions for generalized metric subregularity in nonsmooth setting? [‡‡]
- Can the analysis framework be further extended to cover non-monotone and/or stochastic setting?
- The lift and project approach may depend on the representation of the lifting. Is there an optimal lifting?

[‡‡] $\exists$ nice concepts/results for strict (active) saddle point property for nonsmooth functions (Davis & Drusvyatskiy, 22). Also, it is known that locally Lip. semi-algebraic (more generally tame) function is semismooth (Bolte & Daniilidis & Lewis, 09).

Thanks !