Estimating exponents of Kurdyka-Łojasiewicz inequality and error bounds

Guoyin Li
The University of New South Wales (UNSW)

OWS Seminar
December 7, 2020

Based on joint work with
R. I. Boţ, M. Dao, B.S. Mordukhovich, T.S. Pham, T.K. Pong and P. Yu

Outline

Outline

Outline

Outline

Menu

Entrée

Motivation

Main course

1. Error bounds

2. Kurdyka-Łojasiewicz (KL) inequality

Chef's recommendation

Estimating exponents of error bounds and KL inequality

Sparse Optimization

Sparse optimization:

$$\min_{x\in\mathbb{X}} \quad f(x) := g(x) + h(x).$$

Here, $\mathbb{X}$ is a finite dimensional space and

- g is a loss function which typically measures the data misfitting.
- h is a regularization function which enforces some specific simple or low complexity structure of the solution;
  - h is typically nonsmooth (e.g. $h(x) = \|x\|_p$, $0 < p \leq 1$, cardinality function, rank function, the nuclear norm) and sometimes can be nonconvex.

Examples

- Lasso

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1$$

where $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^p$ and $\lambda \geq 0$.

- Group Lasso (Yuan etal. 2006)

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2 + \sum_{i=1}^{m} \lambda_i \|x_{J_i}\|$$

where $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^p$, $\lambda_i \geq 0$ and $\bigcup_{i=1}^m J_i = \{1,\ldots,n\}$.

- Sparse generalized eigenvalue problem (Tan etal. 2018)

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & \dfrac{x^TAx}{x^TBx} + \lambda\,\|x\|_0 \\ \text{s.t.} & \|x\| = 1, \end{array}$$

where $A, B \in S^n$, B is positive definite, $\lambda \geq 0$ and $\|x\|_0$ is the cardinality of x.

- Least squares with nuclear norm regularization

$$\min_{X\in\mathbb{R}^{m\times n}} \quad \frac{1}{2}\|\mathcal{A}X-b\|^2+\lambda\|X\|_*$$

where $X\in\mathbb{R}^{m\times n}$, $\mathcal{A}:\mathbb{R}^{m\times n}\to\mathbb{R}^p$, $b\in\mathbb{R}^p$ and $\lambda\geq 0$.

- Least squares with rank constraint

$$\begin{array}{ll}\min\limits_{X\in\mathbb{R}^{m\times n}} & \frac{1}{2}\|\mathcal{A}X-b\|^2\\ \text{s.t} & \operatorname{rank}(X)\leq r.\end{array}$$

Sparse optimization is ubiquitous. It has been found applications in a wide range of fields:

- machine learning and statistics;
- signal processing;
- finance;
- structure engineering.

Sparse Optimization

Why it is useful and important?

- A solution for sparse optimization has a desired low complexity structure, so that, it can be efficiently stored, implemented and utilised, and is robust to the data inexactness.

First order methods

First order methods include proximal gradient method and its accelerated version, Douglas-Rachford splitting, Alternating direction methods of multipliers (ADMM) etc.

E.g. Proximal gradient algorithm:

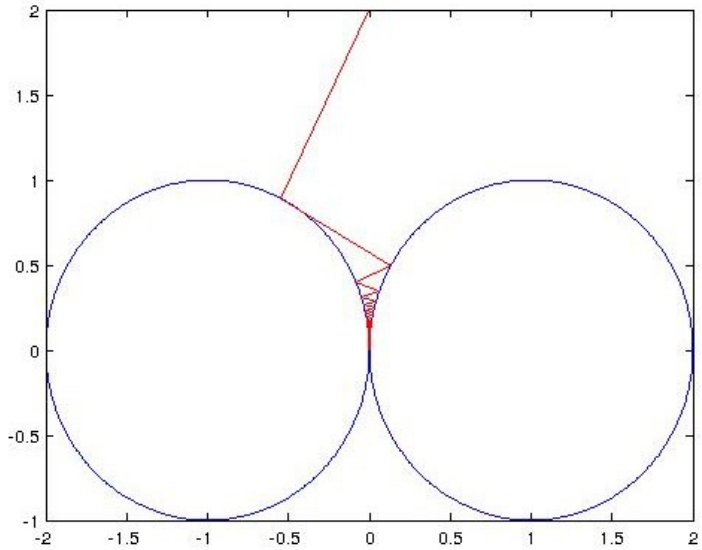
Given $f = g + h$ with g is C^1 with Lipschitz gradient. Initialize x_0. For $k = 1, 2, \ldots,$

$$x_{k+1} \in \text{prox}_{\gamma h}(x_k - \gamma \nabla g(x_k))$$

Here, $\text{prox}_{\gamma h}(x) = \text{Argmin}_{y \in \mathbb{R}^n}\{\frac{1}{2}\|x - y\|^2 + \gamma h(y)\}$.

Note: For $g = \frac{1}{2}d_C^2$ and $h = \delta_D$ where δ is the indicator function with convex sets C, D. Proximal gradient method with $\gamma = 1$ reduces to alternating projection algorithm: $x_{k+1} = P_D(P_C(x_k))$.

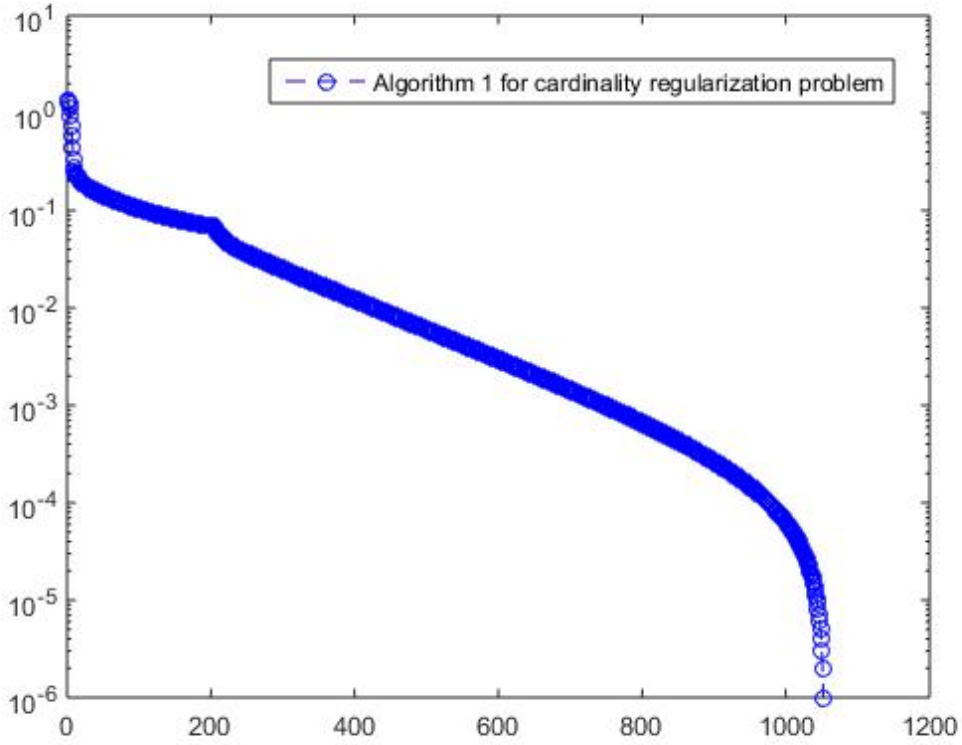
- Alternating projection algorithm for degenerate case:



Direct computation shows that $x_k \to 0$ and $\|x_k\| = O(\frac{1}{\sqrt{k}})$.

- Alternating projection algorithm for non-degenerate case: linear convergence (see e.g. Bauschke & Borwein, 1996; Lewis, Luke & Malick, 2009, Drusvyatskiy, Ioffe & Lewis 2015)

- Inertial proximal gradient method for an equivalent reformulation for sparse generelized eigenvalue problem (Boţ, Dao & L. 2020)



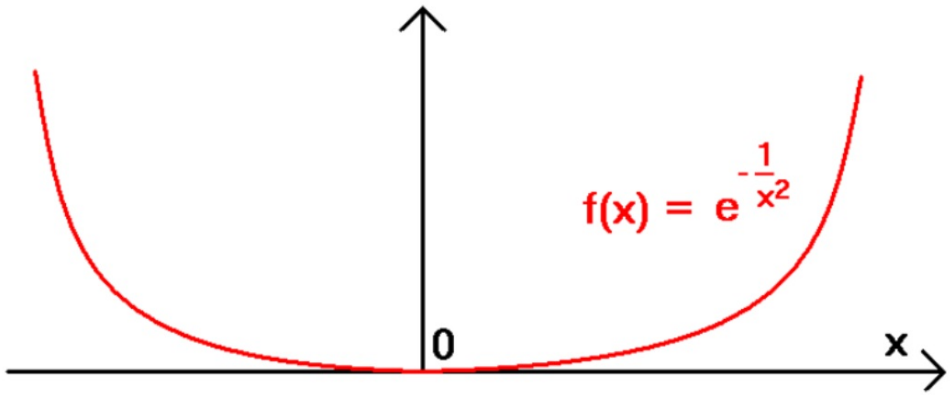
How do we understand the convergence behavior or convergence rate of these numerical algorithms?

Key tool I: 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 $\alpha \in (0, 1]$ and $c, \delta > 0$ such that

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

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



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

Definition

We say that a proper closed function $f : \mathbb{X} \to \mathbb{R} \cup \{\infty\}$ satisfies the Kurdyka-Łojasiewicz (KL) property at $\bar{x} \in \operatorname{dom} \partial f$ if there are $\nu \in (0, \infty]$, a neighborhood V of $\bar{x}$ and a continuous concave function $\varphi : [0, \nu) \to [0, \infty)$ with $\varphi(0) = 0$ such that

(i) φ is continuously differentiable on $(0, \nu)$ with $\varphi' > 0$ on $(0, \nu)$;

(ii) For any $x \in V$ with $f(\bar{x}) < f(x) < f(\bar{x}) + \nu$, it holds that

$$\varphi'(f(x) - f(\bar{x}))\operatorname{dist}(0, \partial f(x)) \geq 1. \tag{3.0}$$

Note: ∂f is the so-called limiting subdifferential. The Łojasiewicz inequality corresponds to the case where $\varphi(s) = cs^{1-\alpha}$.

KL exponent

Definition (KL exponent, Attouch et al. 10)

We say that a proper closed function f has the Kurdyka-Łojasiewicz (KL) property at $\bar{x} \in \operatorname{dom} \partial f$ with exponent α if there exist $c, \nu > 0$ and a neighborhood $\mathcal{N}$ of $\bar{x}$ such that:

- for all $x \in \mathcal{N}$ with $f(\bar{x}) < f(x) < f(\bar{x}) + \nu$, one has

$$\operatorname{dist}(0, \partial f(x)) \geq c\big[f(x) - f(\bar{x})\big]^{\alpha}.$$

A proper closed function f satisfying the KL property with exponent α at all points in $\operatorname{dom} \partial f$ is called a KL function with exponent α.

Prototypical result on convergence rate:

For proximal gradient algorithm and some of its variants: Let $\{x^k\}$ be a bounded sequence generated by the algorithm. Let f be a KL function with exponent $\alpha \in [0, 1)$. Then the following results hold.

(i) If $\alpha = 0$, then $\{x^k\}$ converges finitely.

(ii) If $\alpha \in (0, \frac{1}{2}]$, then $\{x^k\}$ converges locally linearly.

(iii) If $\alpha \in (\frac{1}{2}, 1)$, then $\{x^k\}$ converges locally sublinearly with order $O(k^{-\tau})$ and $\tau = \frac{1-\alpha}{2\alpha-1}$.

Holds also for proximal alternating minimization algorithm (Attouch et al. '10), Douglas-Rachford algorithm (L., Pong '15), etc., if f is replaced by a suitable potential function.

Key tool II: Error bounds

For $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, we consider the following inequality system

$$(S) \quad f(z) \leq 0.$$

- To judge whether x is an approximate solution of (S), we want to know $d(x, [f \leq 0]) := \inf\{\|x - z\| : f(z) \leq 0\}$.
- However, we often measure $[f(x)]_+ := \max\{f(x), 0\}$.
- So, we seek an error bound: there exist $\tau, \alpha > 0$ such that

$$d(x, [f \leq 0]) \leq \tau \left([f(x)]_+ + [f(x)]_+^{\alpha}\right)$$

either locally or globally.

Key tool II: Error bounds

For $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, we consider the following inequality system

$$(S) \quad f(z) \leq 0.$$

- To judge whether x is an approximate solution of (S), we want to know $d(x, [f \leq 0]) := \inf\{\|x - z\| : f(z) \leq 0\}$.
- However, we often measure $[f(x)]_+ := \max\{f(x), 0\}$.
- So, we seek an error bound: there exist $\tau, \alpha > 0$ such that

$$d(x, [f \leq 0]) \leq \tau \left([f(x)]_+ + [f(x)]_+^{\alpha}\right)$$

either locally or globally.

Error bounds and its exponent

Definition

We say f has a
(1) global error bound with exponent α if there exist $\tau > 0$ such that

$$d(x, [f \leq 0]) \leq \tau\left([f(x)]_+ + [f(x)]_+^{\alpha}\right) \text{ for all } x \in \mathbb{R}^n$$

(2) local error bound with exponent α around $\overline{x}$ if there exist $\tau, \epsilon > 0$ such that

$$d(x, [f \leq 0]) \leq \tau\left([f(x)]_+ + [f(x)]_+^{\alpha}\right) \text{ for all } x \in \mathbb{B}(\overline{x}; \epsilon).$$

If $\alpha = 1$, we say f has a Lipschitz type global (resp. local) error bound.

Error bound is useful in

- analyzing the convergence properties of algorithms (e.g. Luo 2000, Fukushima 2005, Attouch etal. 2009, Tseng 2010 and Izmailov & Solodov 2014);
- sensitivity analysis of optimization problem/variational inequality problem (e.g. Jourani 2000, Ye 2002)
- identifying the active constraints (e.g. Facchinei etal. 1998 and Pang 1997)
- studying maximal monotone operator (Dutta & Borwein 2015)

Some Known Results

- Lipschitz type global error bound holds when f is maximum of finitely many affine functions (Hoffman 1951)
- Global error bound can fail even when f is convex and continuous (e.g. $f(x_1, x_2) = x_1 + \sqrt{x_1^2 + x_2^2}$).
- Many further developments (e.g. Ioffe, Klatte, Kummer, Kruger, Lewis, Li, Ng, Outrata, Pang, Robinson, Thera etc...)

Quadratic cases

- Global error bound with exponent $1/2$ holds when f is a convex quadratic function. (Luo and Luo, 1994).
- Local error bound with exponent $1/2$ holds when f is a (nonconvex) quadratic function. (Luo and Sturm, 1998).
- Open questions raised by Luo and Sturm: what happens for the case f can be expressed as finitely many quadratic functions?

Interplay between KL inequality and error bounds

- Let f be a proper closed convex function with $\operatorname{argmin} f \neq \emptyset$. Then, the following are equivalent (Bolte etal 2017)
  - f has KL exponent $\alpha \in (0, 1)$;
  - for all $\overline{x} \in \operatorname{Argmin} f$, local error bound holds for $f - \inf f$ at $\overline{x}$ with exponent $1 - \alpha$.

How to estimate these exponents?

Our strategy:

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

Motivating Example

Consider $f(x) = x^2$. Then, $[f \leq 0] = \{0\}$ and so,

$$d(x, [f \leq 0]) = |x| \leq (x^2)^{\frac{1}{2}} = [f(x)]_+^{\frac{1}{2}}.$$

More generally, consider $f(x) = x^d$ with d is an even number. Then,

$$d(x, [f \leq 0]) = |x| \leq (x^d)^{\frac{1}{d}} = [f(x)]_+^{\frac{1}{d}}.$$

- Can we extend the results from convex quadratic functions to convex polynomials? If yes, how about nonconvex cases involving polynomial structures?

Recent development for polynomial systems

- Global error bound with exponent $\frac{1}{(d-1)^n+1}$ holds when f is a convex polynomial with degree d on $\mathbb{R}^n$ (L. 2010).
- Global error bound with exponent $\frac{1}{(d-1)^n+1}$ holds when f is a convex piecewise polynomial with degree d on $\mathbb{R}^n$. (L. 2013).
- local error bound with exponent $\max\left\{\frac{2}{(2d-1)^n+1}, \frac{1}{\beta(n-1)d^n}\right\}$ if f is maximum of finitely many convex polynomials with degree d on $\mathbb{R}^n$, where $\beta(s)$ is the central binomial coefficient $\binom{s}{[s/2]}$ (Borwein, L. & Yao, 2014).

Recent development for polynomial systems cont.

- 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)
- If f is the maximum of m polynomials of degree at most $d \geq 2$ on $\mathbb{R}^n$, then the KL exponent is $1 - \frac{1}{(d+1)(3d)^{n+m-2}}$ (L., Mordukhovich and Pham, 2015)

Next, we illustrate how to derive the exponent in error bound/KL inequality for the case of convex polynomials.

What is special about polynomials?

- Polynomial optimization problems can be solved via a sequential SDP approximation scheme (in some cases, one single SDP is enough). (Lasserre 2000, Parrilo 2000, De Klerk & Laurent 2010, Nie 2014).

- For a convex polynomial f on $\mathbb{R}^n$ with degree d, we have
  (1) $\inf f > -\infty \Rightarrow \operatorname{argmin} f \neq \emptyset$ (Belousov & Klatte 2000);
  (2) $d(0, \nabla f(x_k)) \to 0 \;\Rightarrow\; f(x_k) \to \inf f$ (L. 2010);
  (3) If $f^\infty(v) = 0$, then $f(x + tv) = f(x)$ for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$ (Teboulle & Auslender, 2003).

  Note: $f^\infty(v) = \sup_{t>0} \frac{f(x+tv)-f(x)}{t}$ for all $x \in \operatorname{dom} f$.

What is special about polynomials?

- Polynomial optimization problems can be solved via a sequential SDP approximation scheme (in some cases, one single SDP is enough). (Lasserre 2000, Parrilo 2000, De Klerk & Laurent 2010, Nie 2014).

- For a convex polynomial f on $\mathbb{R}^n$ with degree d, we have
  (1) $\inf f > -\infty \Rightarrow \operatorname{argmin} f \neq \emptyset$ (Belousov & Klatte 2000);
  (2) $d(0, \nabla f(x_k)) \to 0 \;\Rightarrow\; f(x_k) \to \inf f$ (L. 2010);
  (3) If $f^\infty(v) = 0$, then $f(x + tv) = f(x)$ for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$ (Teboulle & Auslender, 2003).

  Note: $f^\infty(v) = \sup_{t>0} \frac{f(x+tv)-f(x)}{t}$ for all $x \in \operatorname{dom} f$.

Let $\kappa(n,d) = (d-1)^n + 1$.

Theorem (L. 2010)

For a convex polynomial f on $\mathbb{R}^n$ with degree d. Then there exists $\tau > 0$ such that

$$d(x,[f \le 0]) \le \tau\big([f(x)]_+ + [f(x)]_+^{\kappa(n,d)^{-1}}\big) \text{ for all } x \in \mathbb{R}^n. \quad (4.0)$$

convex quadratic $\rightsquigarrow$ $d = 2$ (and so, $\kappa(n,d)^{-1} = 1/2$).
previous example x^d $\rightsquigarrow$ $n = 1$ (and so, $\kappa(n,d)^{-1} = 1/d$).

What is behind the proof?

Growth property for polynomials and its variants

- (Gwoździewicz 1999) In addition, if f is a polynomial with degree d and 0 is a strict local minimizer, then, there exist $\beta, \delta > 0$ s.t. $d(x, f^{-1}(0)) \leq \beta |f(x)|^{\rho}$ for all $\|x\| \leq \delta$, with $\rho = \frac{1}{(d-1)^n+1} = \kappa(n,d)^{-1}$.
- Further development on dropping the strict minimizer assumption with weaker estimate in Gwoździewicz's result (Kurdyka 2012, and L., Mordukhovich and Pham 2015).

Outline of the proof

Induction on the dimension k of $[f \leq 0]$

(1) If $k = 0$, then strict minimizer, so Gwoździewicz's result can be applied.

(2) Suppose the result is true for $k = p$;

(3) For the case $k = p + 1$, find a direction v such that $f^{\infty}(v) = 0$, and so, $f(x + tv) = f(x)$ for all x and for all t. Reduce the case to $k = p$.

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 complicate functions via optimal value of conic programs.

Theorem (KL exponent via inf-projection Yu, L. Pong, 2019)

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

(i) *It holds that $\partial F(\bar{x}, \bar{y}) \neq \emptyset$ for all $\bar{y} \in Y(\bar{x})$.*

(ii) *F is level-bounded in y locally uniformly in x.*

(iii) *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

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

Example 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 Yu, L. Pong, 2019)

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 $0 \in \mathrm{ri}\,\partial f(\bar{x})$:

(i) **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 $w_i \geq 0$, $i = 1, \ldots, s$.

(ii) **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, 2019)

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.

Note: One could also cover entropy regularization. The two above cases can also be derived using the machinery of Cui, Ding and Zhao 2017 via spectral functions.

These approaches also allow us to consider other model 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

- Sparse optimization problems is an important and challenging topic, and first order method used widely in this context.
- The convergence rate of the first order method relies on the KL exponent of a suitable potential function.
- One approach in obtaining the KL exponent is to develop calculus rule (such as inf-projection) of KL function and also exploit the underlying polynomial/conic structure.

What we did not cover?

- Proximal error bound (in the sense of Luo & Tseng).
  See e.g. Drusvyatskiy & Lewis 2018, Zhou & So 2017, L. Pong, 2018.
- The link between KL inequality with metric (sub)regularity.
  See e.g. Bolte, Daniilidis, Olivier & Laurent, 2010.
- KL inequality in infinite dimensional space
  See e.g. Hauer & Mazón 2019.

Some questions:

- The lift and project approach may depend on the representation of the lifting. Is there an optimal lifting?
- What about the modulus of error bound and KL inequality?
- Some of the results with KL exponent $1/2$ relies on the ri condition. Can this be relaxed?
- Further calculus rules?

Want to know more?

(1) H. Attouch, J. Bolte, and B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, Math. Program. 137 (2013), 91-129.

(2) G. Li, T.K. Pong, Calculus of the exponent of Kurdyka-Lojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math.* 18 (2018), no. 5, 1199-1232

(3) G. Li, B.S. Mordukhovich and T.S. Pham, New fractional error bounds for nonconvex polynomial systems with applications to Hölderian stability in optimization, *Math. Program*, 153 (2015), 333-362.

(4) P. Yu, G. Li and T.K. Pong, Deducing Kurdyka-Ł ojasiewicz exponent via inf-projection, arXiv:1902.03635

Thanks !