Error bounds, PL, and Quadratic Growth for Weakly Convex Functions, and Linear Convergences of Proximal Point Methods

Feng-Yi Liao[1], Lijun Ding[2], and **Yang Zheng**[1]

[1]Department of Electrical & Computer Engineering, UC San Diego
[2]Department of Industrial & Systems Engineering at Texas A&M University

2024 INFORMS Optimization Society Conference
March 24, 2024

Outline

Motivation: linear convergence of GD methods

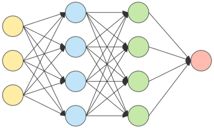
EB, PL, and QG for weakly convex functions

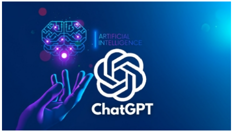
Linear convergences of proximal point methods

Conclusions

Motivation

The success of many machine learning applications





- (Sub)gradient-based methods and their variants are the workhorse algorithms
  - Gradient descent (GD), stochastic GD, coordinate descent, quasi-Newton, etc.
- For **smooth** and **convex** cases, their performances are most well-understood.
  - For example, if $f(x)$ is **strongly convex and** L**-smooth**, then the basic GD algorithm $x_{k+1} = x_k - t_k \nabla f(x_k)$ has linear convergence

$$f(x_{k+1}) - f^* \leq \omega_1 \times (f(x_k) - f^*), \qquad 0 < \omega_1 < 1$$
$$\|x_{k+1} - x^*\| \leq \omega_2 \times \|x_k - x^*\|, \qquad 0 < \omega_2 < 1$$

- But strong convexity is a strong assumption; many machine learning models lack either convexity or smoothness or both.

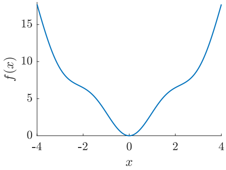
Linear convergence of gradient descent

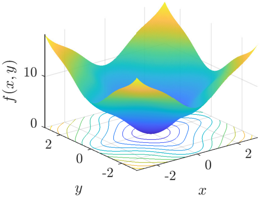
Alternative regularity conditions (weaker than strong convexity)

- One famous condition, introduced by Polyak [1963], is

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \beta \times (f(x) - f^\star), \ \forall x \in \mathbb{R}^n,$$

  where the suboptimality is upper bounded by the gradient norm.
- Holds for strongly convex functions, and also non-convex functions like





- Many other problems like least squares, linear quadratic regulator (LQR) in control, conic optimization (SDPs), etc.
- It is a special case of the Łojasiewicz' inequality [1963] — **Polyak-Łojasiewicz (PL)** inequality (or gradient dominance)

Linear convergence of gradient descent

Simpler proof of linear convergence

- Consider an unconstrained smooth optimization $\min_{x \in \mathbb{R}^n} \; f(x)$, where $f(x)$ satisfies the PL inequality and is L-smooth

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.$$

- Applying one GD step: $x_{k+1} = x_k - t_k \nabla f(x_k)$, leads to

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= f(x_k) + \left(\frac{-2t_k + L t_k^2}{2} \right) \|\nabla f(x_k)\|^2. \end{aligned}$$

- If we choose $0 < t_k < 2/L$, then $L t_k^2 - 2t_k < 0$.
- Applying PL inequality, we have the following linear convergence

$$\begin{aligned} & f(x_{k+1}) \leq f(x_k) + (L t_k^2 - 2t_k) \times \beta \times (f(x_k) - f^*) \\ \Rightarrow \quad & f(x_{k+1}) - f^* \leq \omega_1 \times (f(x_k) - f^*), \qquad 0 < \omega_1 < 1 \end{aligned}$$

Equivalence among regularity conditions

- **Relationship with many other conditions**, including
  - EB: error bounds [Luo and Tseng, 1993].
  - QG: quadratic growth [Anitescu, 2000]
  - ESC: essential strong convexity [Liu et al., 2013].
  - RSI: restricted secant inequality [Zhang & Yin, 2013].
  - a few others
- A nice summary is given in a paper by Karimi et al., 2016

[Karimi et al., 2016, Theorem 2] For the class of L-smooth functions, we have

$$\text{SC} \to \text{RSI} \to \text{EB} \equiv \text{PL} \to \text{QG}$$

If $f(x)$ is further convex, we have RSI $\equiv$ EB $\equiv$ PL $\equiv$ QG.

- This result only focuses on the class of L-smooth functions (the key proof is based on gradient curves)
- Many interesting nonsmooth cases, e.g., $|x|$ or indicator functions of cones

$$f(y) = -b^{\mathsf{T}} y + \delta_{\mathbb{S}^n_+}(c - A^{\mathsf{T}} y)$$

This talk

The class of weakly convex functions

- A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is called ρ-weakly convex if the following function

$$f(x) + \frac{\rho}{2}\|x\|^2$$

  is convex
- A much broader class of functions:
  - any convex (potentially nonsmooth) functions, like $|x|$
  - any L-smooth (potentially nonconvex) functions, like $-x^2 + \sin^2(x)$
  - many cost functions in modern machine learning applications (Drusvyatskiy and Davis, 2020; Atenas et al. 2023)

Message 1: For the class of ρ weakly-convex functions, we have

$$\text{(SC)} \to \text{(RSI)} \to \text{(EB)} \equiv \text{(PL)} \to \text{(QG)}$$

If $f(x)$ is further convex (might be nonsmooth) or the QG coefficient satisfies $\mu_{\mathrm{q}} > \rho/2$, we have (RSI) $\equiv$ (EB) $\equiv$ (PL) $\equiv$ (QG).

- **Message 2**: Exact or inexact PPM will enjoy linear convergence under PL/EB/QG for convex optimization.

Outline

Motivation: linear convergence of GD methods

EB, PL, and QG for weakly convex functions

Linear convergences of proximal point methods

Conclusions

SC, EB, PL, and QG

For a ρ-weakly convex function, its Fréchet subdifferential is well-defined

$$\partial f(x) = \left\{ s \in \mathbb{R}^n \mid \liminf_{y \to x} \frac{f(y) - f(x) - \langle s, y - x \rangle}{\|y - x\|} \geq 0 \right\}.$$

Let $S := \arg\min f(x)$ be the set of optimal solutions. Suppose $S \neq \emptyset$.

1. **Strong Convexity (SC)**: there exists a positive constant $\mu_{\mathrm{s}} > 0$ such that

$$f(x) + \langle g, y - x \rangle + \mu_{\mathrm{s}} \cdot \|y - x\|^2 \leq f(y), \quad \forall g \in \partial f(x). \qquad \text{(SC)}$$

2. **Polyak-Łojasiewic (PL) inequality**: there exists a constant $\mu_{\mathrm{p}} > 0$ such that

$$\mu_{\mathrm{p}} \cdot (f(x) - f^\star) \leq \mathrm{dist}^2(0, \partial f(x)) \qquad \text{(PL)}$$

3. **Error bound (EB)**: there exists a constant $\mu_{\mathrm{e}} > 0$ such that

$$\mathrm{dist}(x, S) \leq \mu_{\mathrm{e}} \cdot \mathrm{dist}(0, \partial f(x)) \qquad \text{(EB)}$$

4. **Quadratic Growth (QG)**: there exists a constant $\mu_{\mathrm{q}} > 0$ such that

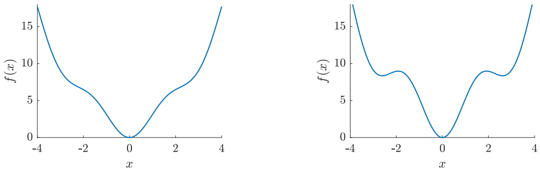
$$\mu_{\mathrm{q}} \cdot \mathrm{dist}^2(x, S) \leq f(x) - f^\star \qquad \text{(QG)}$$

Examples

In principle, these properties are all generalizations of quadratic functions to non-quadratic, nonconvex, and even nonsmooth cases

- which still maintain favourable "quadratic-like" properties.

- **Example 1:** $f(x) = x^2$. Naturally, all properties hold.
- **Example 2:** $f(x) = x^2$ if $|x| \leq 1$; otherwise $f(x) = \frac{1}{2}(x^4 + 1)$; All properties hold, but it is not L-smooth globally.
- **Example 3:** $f_1(x) = x^2 + 2\sin^2(x)$ (left) and $f_2(x) = x^2 + 6\sin^2(x)$ (right)



Both of them satisfy (QG), but the right one does not satisfy (PL) or (EB).

Relationship and equivalency

Theorem

Let f be a proper closed ρ-weakly convex function. We have

$$(\text{SC}) \to (\text{RSI}) \to (\text{EB}) \equiv (\text{PL}) \to (\text{QG}).$$

Furthermore, if 1) $f(x)$ is convex (i.e., $\rho = 0$), or 2) the (QG) coefficient satisfies $\mu_{\mathrm{q}} > \frac{\rho}{2}$, then the following equivalence holds

$$(\text{RSI}) \equiv (\text{EB}) \equiv (\text{PL}) \equiv (\text{QG}).$$

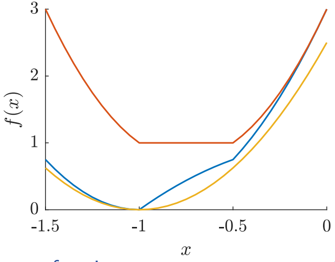
An example of nonconvex function with $\mu_{\mathrm{q}} > \frac{\rho}{2}$, satisfying PL/EB/QG

$$f(x) = \begin{cases} -x^2 + 1 & \text{if } -1 < x < -0.5, \\ 3(x+1)^2 & \text{otherwise.} \end{cases}$$

- It satisfies (QG) with $\mu_{\mathrm{q}} = 3/2$, since

$$f(x) \geq \frac{3}{2}(x+1)^2, \forall x \in \mathbb{R}$$

- $f(x)$ is ρ-weakly convex with $\rho = 2$.



Literature and Proof techniques

An extensive list of literature [an incomplete summary below]

- **Smooth case**: a nice summary appears in [Karimi et al. 2016, Theorem 2], which is a special case of ρ-weakly convex functions.
- **Nonsmooth but convex case**:
  - Equivalence between (EB) and (QG): [Drusvyatskiy and Lewis, 2018, Theorem 3.3] and [Artacho and Geoffroy, 2008, Theorem 3.3]
  - Equivalence between (PL) and (QG): [Bolte et al., 2017, Theorem 5]
  - (PL), (EB), (QG) are equivalent: [Ye et al., 2021, Proposition 2], [Zhu et al. (2023)]
- **Nonsmooth and nonconvex case**: The most closely related work is Drusvyatskiy et al. (2021) on nonsmooth optimization using taylor-like models.
- Our proof from (PL) $\rightarrow$ (EB) relies on a notion of *slop techniques* in Drusvyatskiy et al., 2021.

Proof sketches

Let f be a proper closed ρ-weakly convex function. We have

$$\text{(SC)} \rightarrow \text{(RSI)} \rightarrow \text{(EB)} \equiv \text{(PL)} \rightarrow \text{(QG)}.$$

- The proof: (SC) $\rightarrow$ (RSI) $\rightarrow$ (EB) $\rightarrow$ (PL) $\rightarrow$ (QG) are relatively simple.
- Take (EB) $\rightarrow$ (PL) for example. A function is ρ weakly convex iff

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n, v \in \partial f(x). \tag{1}$$

  - We would like to prove (EB) $\rightarrow$ (PL), i.e.,

$$\operatorname{dist}(x, S) \leq \mu_{\mathrm{e}} \cdot \operatorname{dist}(0, \partial f(x)) \quad \Rightarrow \quad \mu_{\mathrm{p}}(f(x) - f^*) \leq \operatorname{dist}^2(0, \partial f(x))$$

  distance to the solution set $\rightarrow$ suboptimality gap of the cost

  - Fix $x \in \mathbb{R}^n$ and take $y = \Pi_S(x)$; from the quadratic lower bound (1),

$$f^\star \geq f(x) + \langle v, \Pi_S(x) - x \rangle - \frac{\rho}{2}\|\Pi_S(x) - x\|^2, \quad \forall v \in \partial f(x)$$

Proof sketches

Let f be a proper closed ρ-weakly convex function. We have

$$\text{(SC)} \to \text{(RSI)} \to \text{(EB)} \equiv \text{(PL)} \to \text{(QG)}.$$

- Fix $x \in \mathbb{R}^n$ and take $y = \Pi_S(x)$; from the quadratic lower bound (1),

$$f^\star \geq f(x) + \langle v, \Pi_S(x) - x \rangle - \frac{\rho}{2}\|\Pi_S(x) - x\|^2, \quad \forall v \in \partial f(x)$$

- Choose v as the minimal norm element in $\partial f(x)$, completing (EB) $\to$ (PL)

$$\begin{aligned} f(x) - f^\star &\leq \operatorname{dist}(0, \partial f(x)) \operatorname{dist}(x, S) + \frac{\rho}{2}\operatorname{dist}^2(x, S) && \text{Cauchy-Schwartz} \\ &\leq \mu_{\mathrm{e}} \cdot \operatorname{dist}^2(0, \partial f(x)) + \frac{\rho \mu_{\mathrm{e}}^2}{2}\operatorname{dist}^2(0, \partial f(x)) && \text{Applying (EB)} \\ &= \left((2\mu_{\mathrm{e}} + \rho\mu_{\mathrm{e}}^2)/2\right) \operatorname{dist}^2(0, \partial f(x)). \end{aligned}$$

- The proof from (PL) $\to$ (EB) is much more involved: the slope technique [Drusvyatskiy et al., 2021], and Ekeland's variational principle [Ekeland, 1974].

Proof sketches

Let f be a proper closed ρ-weakly convex function. if 1) $f(x)$ is convex (i.e., $\rho = 0$), or 2) the (QG) coefficient satisfies $\mu_{\mathrm{q}} > \frac{\rho}{2} \geq 0$, then

$$\text{(RSI)} \equiv \text{(EB)} \equiv \text{(PL)} \equiv \text{(QG)}.$$

- We only need to prove (QG) $\to$ (EB) when $\mu_{\mathrm{q}} > \frac{\rho}{2} \geq 0$
- Indeed, we have

$$\mu_{\mathrm{q}} \cdot \operatorname{dist}^2(x, S) \leq f(x) - f^\star \leq \langle g, x - \Pi_S(x) \rangle + \frac{\rho}{2} \operatorname{dist}^2(x, S).$$

  Choosing g as the minimal norm element, yields

$$\begin{aligned} \left(\mu_{\mathrm{q}} - \frac{\rho}{2}\right) \cdot \operatorname{dist}^2(x, S) &\leq \langle g, x - \Pi_S(x) \rangle \\ &\leq \operatorname{dist}(0, \partial f(x)) \times \operatorname{dist}(x, S). \quad \text{Cauchy-Schwartz} \end{aligned}$$

- Cancelling a factor, we have (EB)

$$(\mu_{\mathrm{q}} - \rho/2) \cdot \operatorname{dist}(x, S) \leq \operatorname{dist}(0, \partial f(x)).$$

Outline

Motivation: linear convergence of GD methods

EB, PL, and QG for weakly convex functions

Linear convergences of proximal point methods

Conclusions

Proximal point method

Consider the optimization problem

$$f^\star = \min_x \; f(x),$$

where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper closed convex function.

- Define the proximal mapping

$$\mathrm{prox}_{\alpha,f}(x) := \operatorname*{argmin}_{x\in\mathbb{R}^n} \; f(x) + \frac{1}{2\alpha}\left\|x - x_k\right\|^2,$$

- The PPM generates iterates by

$$x_{k+1} = \mathrm{prox}_{c_k,f}(x_k), \quad k = 0, 1, 2, \ldots$$

  where $\{c_k\}_{k\geq 0}$ is a sequence of positive real numbers.
- Conceptually very simple algorithm; historically used for guiding the design/analysis of other algorithms
  - Proximal bundle methods (Lemarechal et al., 1981), augmented Lagrangian methods (Rockafellar, 1976a).
  - Increasing applications in modern machine learning (Drusvyatskiy, 2017)

Linear convergences

- The convergence of PPM for (nonsmooth) convex optimization has been studied since 1970s (Rockafellar, 1976b).
- The sublinear convergence of cost gaps is relatively easy to establish,
- Different assumptions exist for linear convergences: [Rockafellar, 1976b] [Luque, 1984] [Leventhal, 2009] [Cui et al. 2016] [Drusvyatskiy and Lewis, 2018].
  - The classical result by Rockafellar, 1976b requires that $(\partial f)^{-1}$ is locally Lipschitz at 0 (implying a unique solution).

Theorem (Linear convergence)

Let $f:\mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function, and $S \neq \emptyset$. Suppose f satisfies (PL) *(or* (EB), (QG)*) over the sublevel set $[f \leq f^\star + \nu]$. Then, the PPM iterates enjoy linear convergence rates, i.e.,*

$$f(x_{k+1}) - f^\star \leq \omega_k \cdot (f(x_k) - f^\star),$$
$$\operatorname{dist}(x_{k+1}, S) \leq \theta_k \cdot \operatorname{dist}(x_k, S),$$

for all $k \geq k_0$, where $\omega_k = 2/(2 + \mu_{\mathrm{p}} c_k) < 1, 0 < \theta_k < 1$.

Inexact PPM and Linear convergences

- Inexact PPM: consider an inexact update

$$x_{k+1} \approx \mathsf{prox}_{c_k, f}(x_k).$$

- Two classical criteria in Rockafellar's seminal work [Rockafellar, 1976b]

$$\|x_{k+1} - \mathsf{prox}_{c_k, f}(x_k)\| \le \epsilon_k, \quad \textstyle\sum_{k=0}^{\infty} \epsilon_k < \infty, \tag{A}$$

$$\|x_{k+1} - \mathsf{prox}_{c_k, f}(x_k)\| \le \delta_k \|x_{k+1} - x_k\|, \quad \textstyle\sum_{k=0}^{\infty} \delta_k < \infty. \tag{B}$$

Theorem (Linear convergence of inexact PPM)

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed convex function, and $S \neq \emptyset$. Suppose f satisfies (EB) *(or* (QG), (PL)*) over the sublevel set $[f \le f^\star + \nu]$. Let $\{x_k\}$ be any sequence generated by inexact PPM; There exists a nonnegative $\theta_k < 1$ and a large $\bar{k} > 0$ such that for all $k \ge \bar{k}$, we have*

$$\mathrm{dist}(x_{k+1}, S) \le \hat{\theta}_k \mathrm{dist}(x_k, S), \ \textit{where } \hat{\theta}_k = \frac{\theta_k + 2\delta_k}{1 - \delta_k} \ \textit{and} \ \lim_{k \to \infty} \hat{\theta}_k = \theta_k < 1.$$

Numerical examples

Machine Learning instances

- Linear support vector machine (SVM) (Zhang and Lin (2015)),
- Lasso (Tibshirani (1996)),
- Elastic-net (Zou and Hastie (2005))

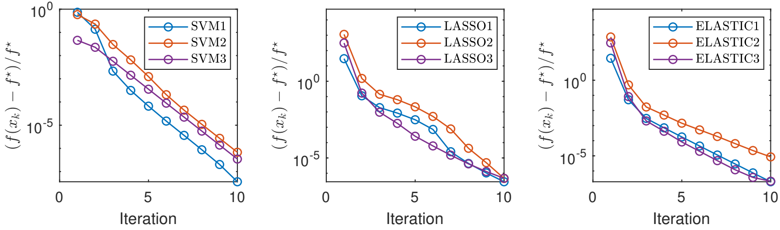


Figure: Linear convergences of cost value gaps for linear SVM (left), lasso (middle), and elastic-net (right).

Outline

Motivation: linear convergence of GD methods

EB, PL, and QG for weakly convex functions

Linear convergences of proximal point methods

Conclusions

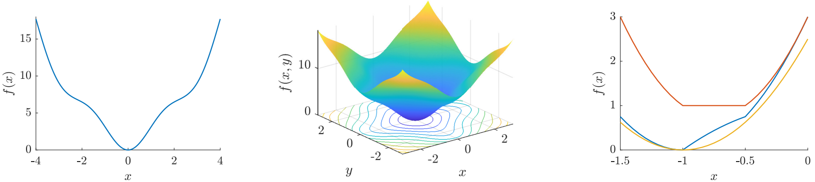
Summary

- **Relationship and equivalency for ρ-weakly convex functions:**

Let f be a proper closed ρ-weakly convex function. We have

$$\text{(SC)} \to \text{(RSI)} \to \text{(EB)} \equiv \text{(PL)} \to \text{(QG)}.$$

Furthermore, if 1) $f(x)$ is convex (i.e., $\rho = 0$), or 2) the (QG) coefficient satisfies $\mu_{\mathrm{q}} > \frac{\rho}{2}$, then we have (RSI) $\equiv$ (EB) $\equiv$ (PL) $\equiv$ (QG).



- **Linear convergences** of PPM and inexact PPM under (EB), (PL), (QG).
- **Ongoing work**: Applications in conic optimizations using the augmented Lagrangian method.

Thank you for your attention!

Q & A

- Liao, Feng-Yi, Lijun Ding, and Yang Zheng. "Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods." arXiv preprint arXiv:2312.16775 (2023).

Supported by
NSF ECCS-2154650;
NSF CMMI-2320697

References

- Boris T Polyak. Gradient methods for the minimisation of functionals. USSR Computational Mathematics and Mathematical Physics, 3(4):864–878, 1963.
- Luo, Zhi-Quan, and Paul Tseng. "Error bounds and convergence analysis of feasible descent methods: a general approach." Annals of Operations Research 46.1 (1993): 157-178.
- Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. arXiv preprint arXiv:1303.4645, 2013.
- Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximalgradient methods under the polyak-łojasiewicz condition. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16, pages 795–811. Springer, 2016.
- Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. Mathematics of Operations Research, 43(3):919–948, 2018.
- FJ Aragon Artacho and Michel H Geoffroy. Characterization of metric regularity of subdifferentials. Journal of Convex Analysis, 15(2):365, 2008.
- Jerome Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. Mathematical Programming, 165:471–507, 2017
- Jane J Ye, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. Set-Valued and Variational Analysis, pages 1–35, 2021.

References

- Daoli Zhu, Lei Zhao, and Shuzhong Zhang. A unified analysis for the subgradient methods minimizing composite nonconvex, nonsmooth and non-lipschitz functions. arXiv preprint arXiv:2308.16362, 2023.
- Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. Mathematical Programming, 185:357–383, 2021.
- R Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of operations research, 1(2):97–116, 1976a.
- R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. SIAM journal on control and optimization, 14(5):877–898, 1976b.
- Dmitriy Drusvyatskiy. The proximal point method revisited. arXiv preprint arXiv:1712.06038, 2017.
- Fernando Javier Luque. Asymptotic convergence analysis of the proximal point algorithm. SIAM Journal on Control and Optimization, 22(2):277–293, 1984
- Ying Cui, Defeng Sun, and Kim-Chuan Toh. On the asymptotic superlinear convergence of the augmented lagrangian method for semidefinite programming with multiple solutions. arXiv preprint arXiv:1610.00875, 2016.

Extra slides

Restricted Secant Inequality

Restricted Secant Inequality (RSI): there exists a positive constant $\mu_\mathrm{r} > 0$ such that

$$\mu_\mathrm{r} \cdot \operatorname{dist}^2(x, S) \le \langle g, x - \Pi_S(x) \rangle, \quad \forall g \in \partial f(x). \tag{RSI}$$

1. **Strong Convexity (SC)**: there exists a positive constant $\mu_\mathrm{s} > 0$ such that

$$f(x) + \langle g, y - x \rangle + \mu_\mathrm{s} \cdot \|y - x\|^2 \le f(y), \quad \forall g \in \partial f(x). \tag{SC}$$

2. **Polyak-Łojasiewic (PL) inequality**: there exists a constant $\mu_\mathrm{p} > 0$ such that

$$\mu_\mathrm{p} \cdot (f(x) - f^\star) \le \operatorname{dist}^2(0, \partial f(x)) \tag{PL}$$

3. **Error bound (EB)**: there exists a constant $\mu_\mathrm{e} > 0$ such that

$$\operatorname{dist}(x, S) \le \mu_\mathrm{e} \cdot \operatorname{dist}(0, \partial f(x)) \tag{EB}$$

4. **Quadratic Growth (QG)**: there exists a constant $\mu_\mathrm{q} > 0$ such that

$$\mu_\mathrm{q} \cdot \operatorname{dist}^2(x, S) \le f(x) - f^\star \tag{QG}$$