Recent Progress on Error Bounds for Structured Convex Programming

Zirui Zhou

Joint work with Anthony Man-Cho So

Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong

September 3, 2014, Beijing

Outline

- overview of error bound
- associated solution mapping
- upper Lipschitzian continuity of multifunctions
- a sufficient condition for error bound
- strongly convex functions
- convex functions with polyhedral epigraph
- group-lasso regularizer
- conclusion

Structured Convex Programming

Consider the structured problem:

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + \tau P(x),$$

$\tau > 0$ given, optimal value v^*, optimal solution set $\mathcal{X}$.

- f: convex and continuously differentiable;
- P: lower semicontinuous and convex, like
  - indicator function of a non-empty closed convex set,
  - various regularizers in application, i.e., ℓ_1, group-lasso.

Residual Function

Define a residual function $R : \mathbb{R}^n \to \mathbb{R}^n$,

$$R(x) := \arg\min_{d \in \mathbb{R}^n} \left\{ \ell_F(x + d; x) + \frac{1}{2}\|d\|^2 \right\},$$

where $\|\cdot\|$ is the usual vector 2-norm and ℓ_F is the linearization of F,

$$\ell_F(y; x) := f(x) + \langle \nabla f(x), y - x \rangle + \tau P(y).$$

- $x \in \mathcal{X} \Leftrightarrow \|R(x)\| = 0$,
- easy to compute.

Residual Function: Examples

- $P(x) \equiv 0, \quad R(x) = -\nabla f(x);$
- $P(x) = \mathcal{I}_D(x), \quad R(x) = x - [x - \nabla f(x)]_D^+;$
- $P(x) = \|x\|_1, \quad R(x) = x - s_\tau(x - \nabla f(x));$

where $[\,\cdot\,]_D^+$ is the projection operator, $s_\tau(\cdot)$ is the vector shrinkage operator.

Let $v = s_\tau(x)$,

$$v_i = \begin{cases} x_i - \tau, & x_i \geq \tau; \\ 0, & -\tau < x_i < \tau; \\ x_i + \tau, & x_i \leq -\tau. \end{cases}$$

Error Bound: Definition

- Forward error: $\text{dist}(x, \mathcal{X})$.

- Backward error: $\|R(x)\|$.

Error Bound Condition: there exists $\kappa > 0$ and a closed set $\mathcal{U} \subseteq \mathbb{R}^n$, such that

$$\text{dist}(x, \mathcal{X}) \leq \kappa \|R(x)\|, \quad \text{whenever } x \in \mathcal{U}.$$

- Global error bound: $\mathcal{U} = \mathbb{R}^n$.

- Local error bound: $\mathcal{U}$ is the closure of a neighbourhood of $\mathcal{X}$.

What If Error Bound Holds

- **Stopping criterion:** estimate $\text{dist}(x^k, \mathcal{X})$,

$$\text{dist}(x^k, \mathcal{X}) \le \kappa \|R(x^k)\|.$$

- **Linear convergence:** for example, under mild assumptions,

$$\|R(x^k)\| \le \kappa_1 \|x^{k+1} - x^k\|, \quad k = 1, 2, \ldots,$$

  This gives a key step for linear convergence,

$$\text{dist}(x^k, \mathcal{X}) \le \kappa \|R(x^k)\| \le \kappa\kappa_1 \|x^{k+1} - x^k\|,$$

  – global error bound $\Rightarrow$ global linear rate;
  – local error bound $\Rightarrow$ asymptotic linear rate.

Conditions for Error Bounds: Existing Results

(a) f is strongly convex **[Pang'87]**;

(b) $f(x) = h(Ax)$, $P(x)$ is of polyhedral epigraph **[Luo-Tseng'92]**;

(c) $f(x) = h(Ax)$, $P(x)$ is the group-lasso or sparse group-lasso regularizer **[Tseng'09, Zhang-Jiang-Luo'13]**.

Notations in case (b) and (c),

- A is any matrix;
- h is strongly (strictly) convex differentiable function with ∇h Lipschitz continuous;
- group-lasso: for $x \in \mathbb{R}^n$, $P(x) = \sum_{J \in \mathcal{J}} \omega_J \|x_J\|_2$. $\mathcal{J}$ is a non-overlapping partition of $\{1, \ldots, n\}$.

Assumptions

Throughout, for the structured problem

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + \tau P(x), \tag{1}$$

we make the following assumptions:

- f takes the form
$$f(x) = h(Ax),$$
where $A \in \mathbb{R}^{m\times n}$ is a matrix, $h : \mathbb{R}^m \to \mathbb{R}$ is σ-strongly convex and ∇h is L-Lipschitz continuous;

- $\mathcal{X}$ is non-empty.

Optimal Solution Set

First-order optimality condition,

$$\mathcal{X} = \{x \in \mathbb{R}^n \mid \mathbf{0} \in \nabla f(x) + \tau \partial P(x)\}.$$

Since h is strictly convex, we have

- there exists $\bar{y} \in \mathbb{R}^m$ such that $Ax = \bar{y}$, $\forall x \in \mathcal{X}$;
- $\nabla f(x) = A^T \nabla h(Ax)$, by letting $\bar{g} = A^T \nabla h(\bar{y})$, then $\nabla f(x) = \bar{g}$, $\forall x \in \mathcal{X}$.

Thus, by assuming $\bar{y}$ and $\bar{g}$ are known, $\mathcal{X}$ has the following characterization,

$$\mathcal{X} = \{x \in \mathbb{R}^n \mid Ax = \bar{y},\ -\bar{g} \in \tau \partial P(x)\}.$$

Solution Mapping

- Let $\Sigma : \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ be a multifunction (set-valued function) defined as

$$\Sigma(t, e) := \{x \in \mathbb{R}^n \mid Ax = t,\ e \in \partial P(x)\}, \quad \forall t \in \mathbb{R}^m, e \in \mathbb{R}^n.$$

  We say Σ is the solution mapping associated with (1).

- Relationship with optimal solution set:

$$\mathcal{X} = \Sigma(\bar{y}, -\bar{g}/\tau) \ .$$

Upper Lipschitzian Continuity

For any solution mapping Σ and any $(\bar{t}, \bar{e}) \in \mathbb{R}^m \times \mathbb{R}^n$, we say

- Σ is globally upper Lipschitzian continuous (global-ULC) at $(\bar{t}, \bar{e})$ with modulus θ, if
$$\Sigma(t, e) \subseteq \Sigma(\bar{t}, \bar{e}) + \theta \|(t, e) - (\bar{t}, \bar{e})\| \mathcal{B}, \quad \forall (t, e) \in \mathbb{R}^m \times \mathbb{R}^n.$$

- Σ is locally upper Lipschitzian continuous (local-ULC) at $(\bar{t}, \bar{e})$ with modulus θ, if there exists a constant $\delta > 0$ such that
$$\Sigma(t, e) \subseteq \Sigma(\bar{t}, \bar{e}) + \theta \|(t, e) - (\bar{t}, \bar{e})\| \mathcal{B}, \quad \text{whenever } \|(t, e) - (\bar{t}, \bar{e})\| \leq \delta.$$

Here $\mathcal{B}$ is the unit ball of $\mathbb{R}^m \times \mathbb{R}^n$.

A Sufficient Condition for Error Bound

Proposition. *Let* Σ *be the associated solution mapping of* (1)*, then*

(a) Σ *is global-ULC at* $(\bar{y}, -\bar{g}/\tau) \Longrightarrow$ *global error bound holds.*

(b) Σ *is local-ULC at* $(\bar{y}, -\bar{g}/\tau) \Longrightarrow$ *local error bound holds.*

Remark. In case (b), the strongly convex assumption on h can be relaxed to strictly convex, i.e., strongly convex on any compact subset of dom$h$.

Proof of Global Error Bound

For any $x \in \mathbb{R}^n$, by optimality condition of $R(x)$,

$$\mathbf{0} \in \nabla f(x) + R(x) + \tau \partial P(x + R(x)).$$

This gives us

$$x + R(x) \in \Sigma \left(A(x + R(x)),\ -\frac{\nabla f(x) + R(x)}{\tau} \right).$$

Since Σ is global-ULC at $(\bar{y}, -\bar{g}/\tau)$ and $\Sigma(\bar{y}, -\bar{g}/\tau) = \mathcal{X}$.

$$\begin{aligned} \text{dist}(x + R(x), \mathcal{X}) &\leq \theta \left\| \left(A(x + R(x)), -\frac{\nabla f(x) + R(x)}{\tau} \right) - (\bar{y}, -\bar{g}/\tau) \right\| \\ &\leq \tilde{\theta} \left(\|Ax - \bar{y}\| + \|R(x)\| \right). \end{aligned}$$

The second inequality utilizes Lipschitz continuity of ∇f.

Suppose $\bar{x}$ is the projection of x onto $\mathcal{X}$, and $\bar{x}^R$ is the projection of $x + R(x)$.

$$\begin{aligned} \text{dist}(x, \mathcal{X}) &\leq \|x - \bar{x}^R\| = \|x + R(x) - \bar{x}^R - R(x)\| \\ &\leq \text{dist}(x + R(x), \mathcal{X}) + \|R(x)\|. \end{aligned}$$

Thus by choosing proper constant κ_0, we obtain

$$\text{dist}(x, \mathcal{X}) \leq \kappa_0 \left(\|Ax - \bar{y}\| + \|R(x)\|\right).$$

Using the inequality that for any $a, b \in \mathbb{R}$, $(a+b)^2 \leq 2(a^2 + b^2)$, we have

$$\text{dist}^2(x, \mathcal{X}) \leq 2\kappa_0^2(\|Ax - \bar{y}\|^2 + \|R(x)\|^2). \tag{2}$$

Since h is strongly convex with factor σ,

$$\sigma\|Ax - \bar{y}\|^2 \leq \langle \nabla h(Ax) - \nabla h(\bar{y}), Ax - \bar{y} \rangle = \langle \nabla f(x) - \bar{g}, x - \bar{x} \rangle. \tag{3}$$

Using Fermat's rule for $R(x)$ and standard arguments, there exists constant $\kappa_1 > 0$ such that

$$\langle \nabla f(x) - \bar{g}, x - \bar{x} \rangle \leq \kappa_1 \|x - \bar{x}\| \cdot \|R(x)\|.$$

Combining the above equality with (3) and (2), there exists $\kappa_2 > 0$ satisfying

$$\mathrm{dist}^2(x, \mathcal{X}) \leq \kappa_2(\|x - \bar{x}\| \cdot \|R(x)\| + \|R(x)\|^2).$$

Solving this quadratic inequality, we obtain a constant κ such that

$$\mathrm{dist}(x, \mathcal{X}) \leq \kappa\|R(x)\|.$$

This establishes the global error bound. $\square$

ULC Property of Solution Mapping

Solution mapping:

$$\Sigma(t, e) = \{x \in \mathbb{R}^n \mid Ax = t,\ e \in \partial P(x)\}, \quad \forall t \in \mathbb{R}^m, e \in \mathbb{R}^n.$$

Next, we will study the ULC property of Σ for the following three cases.

- f is strongly convex and P is any lower-semicontinuous convex function;
- f is non-strongly convex and P is of polyhedral epigraph;
- f is non-strongly convex and P is group-lasso regularizer.

f **Strongly Convex**

- A is surjective, and has inverse A^{-1}.
- For any $(t, e) \in \mathbb{R}^m \times \mathbb{R}^n$,

$$\Sigma(t, e) = \{A^{-1}(t)\}, \quad \text{or} \quad \Sigma(t, e) = \emptyset.$$

- If Σ is non-empty at $(\bar{t}, \bar{e})$, then

$$\Sigma(t, e) \subseteq \Sigma(\bar{t}, \bar{e}) + \|A^{-1}\| \cdot \|t - \bar{t}\|\mathcal{B}, \quad \forall (t, e) \in \mathbb{R}^m \times \mathbb{R}^n.$$

So in this case, Σ is **global-ULC** at $(\bar{t}, \bar{e})$ and **global error bound** holds.

f Non-Strongly Convex and P Polyhedral

- P is of polyhedral epigraph.

$$\text{epi}P = \{(z, w) \in \mathbb{R}^n \times \mathbb{R} \mid C_z z + C_w w \le d\},$$

  where $C_w, d \in \mathbb{R}^l$, $C_z \in \mathbb{R}^l \times \mathbb{R}^n$.

- **Proposition:** for any $e \in \mathbb{R}^n$, $e \in \partial P(x)$ if and only if there exists $s \in \mathbb{R}$ such that (x, s) is the optimal solution of the following LP:

$$\begin{array}{ll} \min & -e^T z + w \\ \text{s.t.} & C_z z + C_w w \le d \end{array} \tag{4}$$

  Proof: Indeed, if $e \in \partial P(x)$, by definition of subgradient,

$$P(z) \ge P(x) + e^T(z - x), \quad \forall z \in \text{dom}P.$$

  Upon rearranging,

$$P(x) - e^T x \le P(z) - e^T z \le w - e^T z, \quad \forall (z, w) \in \text{epi}P.$$

This implies $(x, P(x))$ is an optimal solution of (4).

On the other hand, if (x, s) is an optimal solution, then $s = P(x)$. If not, since $(x, s), (x, P(x)) \in \text{epi}P$, $P(x) < s$ and $-e^T x + P(x) < -e^T x + s$. So

$$P(x) - e^T x \leq P(z) - e^T z, \quad \forall z \in \text{dom}P.$$

By definition of subgradient, $e \in \partial P(x)$. □

- **Optimality Condition for LP:** $e \in \partial P(x)$ if and only if there exist $s \in \mathbb{R}, \gamma \in \mathbb{R}^l$ such that (x, s, γ) is the solution of the following system,

$$\mathcal{S}(e) := \left\{ (z, w, \lambda) \;\middle|\; \begin{array}{rcl} C_z^*(\lambda) & = & e, \\ 1 + \langle C_w, \lambda \rangle & = & 0, \\ \lambda & \geq & \mathbf{0}, \\ C_z z + C_w \cdot w & \leq & d, \\ \langle \lambda, C_z z + C_w \cdot w - d \rangle & = & 0. \end{array} \right\}$$

- The solution mapping Σ can be expressed as

$$\Sigma(t, e) = \left\{ x \in \mathbb{R}^n \mid Ax = t,\ (x, s, \gamma) \in \mathcal{S}(e) \text{ for some } s \in \mathbb{R}, \gamma \in \mathbb{R}^l \right\}.$$

Polyhedral Multifunction

- A multifunction $\Gamma : \mathcal{X} \rightrightarrows \mathcal{Y}$ is said to be a polyhedral multifunction if Graph(Γ) is a finite union of polyhedral sets, where

$$\text{Graph}(\Gamma) := \{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid y \in \Gamma(x)\}.$$

- Polyhedral multifunctions are local-ULC **[Robinson'81]**.

- Σ is a polyhedral multifunction and thus Σ is **local-ULC**.

So in this case, we have **local error bound**.

f Non-Strongly Convex and P Group-Lasso Regularizer

- Group-lasso regularizer:

$$P(x) = \sum_{J \in \mathcal{J}} \omega_J \|x_J\|_2,$$

- Solution mapping:

$$\Sigma(t, e) = \left\{ x \in \mathbb{R}^n \mid Ax = t,\ e \in \sum_{J \in \mathcal{J}} \omega_J \partial \|x_J\|_2 \right\}.$$

- **Theorem.** For any $(\bar{t}, \bar{e}) \in \mathbb{R}^m \times \mathbb{R}^n$, if Σ is non-empty and bounded at $(\bar{t}, \bar{e})$, then Σ is locally upper Lipschitzian continuous at $(\bar{t}, \bar{e})$.

So in this case, we have **local error bound**.

Proof of Theorem

For simplicity, we consider

$$\Sigma(t,e) = \{x \in \mathbb{R}^n \mid Ax = t,\ e \in \partial\|x\|_2\}\,.$$

By the definition of subgradient,

$$\partial\|z\|_2 = \begin{cases} \mathbb{B}(\mathbf{0},1) & \text{if } z = \mathbf{0}; \\ z/\|z\|_2 & \text{otherwise.} \end{cases}$$

- If $\|e\|_2 > 1$, $\Sigma(t,e)$ is empty;
- if $\|e\|_2 < 1$, $\Sigma(t,e)$, if not empty, equals $\{\mathbf{0}\}$;
- if $\|e\|_2 = 1$, $\Sigma(t,e)$, if not empty, has the expression

$$\Sigma(t,e) = \{x \in \mathbb{R}^n \mid Ax = t, x \text{ is a non-negative multiple of } e\}.$$

Suppose $(\bar{t}, \bar{e})$ satisfies that $\Sigma(\bar{t}, \bar{e})$ is non-empty and bounded. So $\|\bar{e}\|_2 \leq 1$. Consider the following two cases: (a) $\|\bar{e}\|_2 < 1$; (b) $\|\bar{e}\|_2 = 1$.

- **(a)** In this case $\Sigma(\bar{t}, \bar{e}) = \{\mathbf{0}\}$. Since $\|\bar{e}\|_2 < 1$, there exists $\delta_a > 0$ satisfying $\|e\|_2 < 1$ whenever $\|e - \bar{e}\|_2 \leq \delta_a$. So

$$\Sigma(t, e) = \emptyset \text{ or } \{\mathbf{0}\}, \quad \text{whenever } \|(t, e) - (\bar{t}, \bar{e})\|_2 \leq \delta_a.$$

It then satisfies

$$\Sigma(t, e) \subseteq \Sigma(\bar{t}, \bar{e}) + \theta \|(t, e) - (\bar{t}, \bar{e})\|_2 \mathcal{B}, \quad \text{whenever } \|(t, e) - (\bar{t}, \bar{e})\|_2 \leq \delta_a.$$

By definition, Σ is local-ULC at $(\bar{t}, \bar{e})$ if $(\bar{t}, \bar{e})$ is of case (a).

- **(b)** In this case,

$$\Sigma(\bar{t}, \bar{e}) = \{x \in \mathbb{R}^n \mid Ax = \bar{t}, x \text{ is a non-negative multiple of } \bar{e}\}.$$

Let $[\bar{e}, \bar{E}]$ be an orthonormal basis of $\mathbb{R}^n$. Then

$$x \text{ is a non-negative multiple of } \bar{e} \quad \Longleftrightarrow \quad \bar{e}^T x \geq 0, \bar{E}^T x = \mathbf{0}.$$

Thus we have the representation of Σ as

$$\Sigma(\bar{t}, \bar{e}) = \{x \in \mathbb{R}^n \mid Ax = \bar{t}, \bar{e}^T x \geq 0, \bar{E}^T x = \mathbf{0}\}.$$

This implies $\Sigma(\bar{t}, \bar{e})$ is a **polyhedral set**.

Applying the well-known **Hoffman's bound**, there exists $\kappa > 0$,

$$\text{dist}(x, \Sigma(\bar{t}, \bar{e})) \leq \kappa \left(\|Ax - \bar{t}\|_2 + [\bar{e}^T x]^- + \|\bar{E}^T x\|_2 \right), \quad \forall x \in \mathbb{R}^n.$$

For any scalar z, we denote $[z]^- = \max\{0, -z\}$.

Now consider $x \in \Sigma(t,e)$ with $(t,e) \neq (\bar{t},\bar{e})$.

– If $\|e\|_2 < 1$, then $x = \mathbf{0}$ and $Ax = t$. We obtain

$$\text{dist}(x, \Sigma(\bar{t},\bar{e})) \leq \kappa\|t-\bar{t}\|_2 \leq \kappa(\|t-\bar{t}\|_2 + \|e-\bar{e}\|_2), \quad \forall x \in \Sigma(t,e). \quad (5)$$

– If $\|e\|_2 = 1$, then $Ax = t$ and x is a non-negative multiple of e.

Fact. There exists a matrix E such that $[e, E]$ is an orthonormal basis of $\mathbb{R}^n$ and $\|E_i - \bar{E}_i\|_2 \leq \|e-\bar{e}\|_2, i = 1,\ldots,n-1$. E_i is the i-th column of E.

$$x \text{ is a non-negative multiple of } e \Longleftrightarrow e^T x \geq 0, \quad E^T x = \mathbf{0}.$$

Thus for any $x \in \Sigma(t,e)$,

$$\begin{aligned}
\text{dist}(x, \Sigma(\bar{t},\bar{e})) &\leq \kappa(\|t-\bar{t}\|_2 + [\bar{e}^T x]^- + \|\bar{E}^T x\|_2) \\
&\leq \kappa(\|t-\bar{t}\|_2 + [e^T x]^- + [(\bar{e}-e)^T x]^- + \|E^T x\|_2 + \|(\bar{E}-E)^T x\|_2) \\
&\leq \kappa\left(\|t-\bar{t}\|_2 + \|\bar{e}-e\|_2\|x\|_2 + \textstyle\sum_{i=1}^n \|\bar{E}_i - E_i\|_2\|x\|_2\right) \\
&\leq \kappa(\|t-\bar{t}\|_2 + n\|x\|_2\|\bar{e}-e\|_2)
\end{aligned}$$

Fact. If $\Sigma(\bar{t}, \bar{e})$ is bounded, there exists $\delta_b > 0$ such that $\Sigma(t, e)$ is bounded whenever $\|(t, e) - (\bar{t}, \bar{e})\|_2 \leq \delta_b$.

So there exists $R > 0$ such that for any $x \in \Sigma(t, e)$ with $\|(t, e) - (\bar{t}, \bar{e})\|_2 \leq \delta_b$, $\|x\|_2 \leq R$. Using the above relationship, we obtain that for any (t, e) satisfying $\|(t, e) - (\bar{t}, \bar{e})\|_2 \leq \delta_b$ and $\|e\|_2 = 1$,

$$\text{dist}(x, \Sigma(\bar{t}, \bar{e})) \leq \kappa(1 + nR)(\|t - \bar{t}\|_2 + \|e - \bar{e}\|_2), \quad \forall x \in \Sigma(t, e). \tag{6}$$

Combining (5) and (6), by letting $\theta = \kappa(1 + nR)$,

$$\Sigma(t, e) \subseteq \Sigma(\bar{t}, \bar{e}) + \theta\|(t, e) - (\bar{t}, \bar{e})\|_2 \mathcal{B}, \quad \text{whenever } \|(t, e) - (\bar{t}, \bar{e})\|_2 \leq \delta_b.$$

So Σ is local-ULC at $(\bar{t}, \bar{e})$ if $(\bar{t}, \bar{e})$ is of case (b).

Together with case (a), Σ is local-ULC at $(\bar{t}, \bar{e})$ is Σ is non-empty and bounded at $(\bar{t}, \bar{e})$. □

Conclusions and Future Work

Contributions:

- based on the ULC property of the associated solution mapping, we give a sufficient condition for error bound and unifies all the existing results.

- we give an alternative approach to error bound for group-lasso regularized optimization.

Some of the future directions:

- study the solution mapping for more cases, i.e., mixed norm, nuclear norm.

- error bounds beyond current assumptions.