A Unified Approach to Error Bounds for Structured Convex Optimization Problems

Anthony Man–Cho So

Joint Work with Zirui Zhou

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong (CUHK)

Splitting Algorithms, Modern Operator Theory, and Applications
CMO–BIRS Workshop
Oaxaca, Mexico

September 18, 2017

Table of Contents

1. Introduction

2. Sufficient Conditions for Error Bound

3. Error Bounds for $\ell_{1,p}$–Norm Regularization

4. Error Bounds for Trace Norm Regularization

5. Conclusion

Table of Contents

1. Introduction

2. Sufficient Conditions for Error Bound

3. Error Bounds for $\ell_{1,p}$–Norm Regularization

4. Error Bounds for Trace Norm Regularization

5. Conclusion

Sturctured Convex Optimization

- Many regularized loss minimization problems in machine learning take the form

$$\min_{x \in \mathcal{E}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell([\mathcal{A}(x)]_i, b_i) + P(x) \right\}, \qquad \text{(RLM)}$$

  where

  - $\mathcal{E}$ is a finite–dimensional Euclidean space,
  - $\mathcal{A} : \mathcal{E} \to \mathbb{R}^N$ is a linear operator with $[\mathcal{A}(x)]_i$ representing the i–th prediction and $b_i \in \mathbb{R}$ is the i–th response,
  - $\ell : \mathbb{R} \to \mathbb{R}$ is a smooth convex loss function, and
  - $P : \mathbb{R}^n \to \mathbb{R}_+$ is a non–smooth, structure–inducing convex regularizer.

- For simplicity, we define

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \ell([\mathcal{A}(x)]_i, b_i).$$

Sturctured Convex Optimization

	$\ell(y, b)$	Domain of b
Linear Regression	$\frac{1}{2}(y-b)^2$	$\mathbb{R}$
Logistic Regression	$\log(1+\exp(-yb))$	$\{-1, 1\}$
Poisson Regression	$-yb+\exp(y)$	$\{0, 1, \ldots\}$

(a) Loss Functions

	$\mathcal{E}$	$P(x)$
LASSO	$\mathbb{R}^n$	$\|x\|_1$
Grouped LASSO	$\mathbb{R}^n$	$\sum_{J\in\mathcal{J}} \omega_J \|x_J\|_2,\ \omega_J \geq 0,$ $\mathcal{J}$ a partition of $\{1, \ldots, n\}$
Nuclear Norm	$\mathbb{R}^{m\times n}$	$\|x\|_*$

(b) Regularizers

Table: Some commonly used loss functions and regularizers.

How to Solve (RLM) Efficiently?

- Provided that the loss function ℓ and the regularizer P are efficiently representable, Problem (RLM) can be solved to arbitrary accuracy in polynomial time by interior point methods (IPMs) **[Nesterov-Nemirovski'94]**.

- Polynomial–time solvability used to be a big deal. However, it is no longer a practical measure of efficiency.

  - IPMs require solving a linear system in each iteration, which could be costly in the big data era.

- Recent applications lead to a renewal of interest in first–order methods (FOMs).

First–Order Methods for Solving (RLM)

- As the name suggests, various FOMs solve Problem (RLM) by finding a solution to its first–order necessary and sufficient optimality condition:

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

- It is well–known that this is equivalent to solving the fixed–point equation

$$x = \mathrm{prox}_P(x - \nabla f(x)), \tag{FP}$$

  where $\mathrm{prox}_P : \mathcal{E} \to \mathcal{E}$ is the proximity operator w.r.t. P defined by

$$\mathrm{prox}_P(x) = \arg\min_{y \in \mathcal{E}} \left\{ P(y) + \frac{1}{2}\|x - y\|_2^2 \right\}.$$

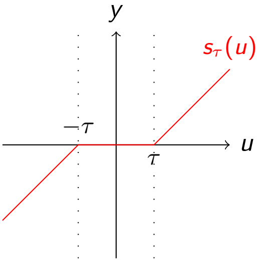
Examples of the Proximity Operator

- $P(x) = \mathcal{I}_{\mathcal{C}}(x)$: $\mathrm{prox}_P(u) = \Pi_{\mathcal{C}}(u)$, where $\mathcal{C} \subseteq \mathbb{R}^n$ is a non–empty closed convex set and

$$\mathcal{I}_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

- $P(x) = \tau \|x\|_1$: $\mathrm{prox}_P(u) = s_\tau(u)$, where $s_\tau : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$v = s_\tau(u), \quad v_i = \begin{cases} u_i - \tau & \text{if } u_i \geq \tau, \\ 0 & \text{if } u_i \in (-\tau, \tau), \\ u_i + \tau & \text{if } u_i \leq -\tau. \end{cases}$$



First–Order Methods for Solving (RLM)

- Currently, FOMs for solving (RLM) are quite well developed.
  - proximal gradient method **[folklore]**
  - incremental proximal methods **[Bertsekas'11]**
  - proximal stochastic dual coordinate ascent **[Shalev-Shwartz-Zhang'13]**
  - proximal stochastic gradient methods **[Nitanda'14, Xiao-Zhang'14, Lin-Lu-Xiao'15]**
  - ...

Analyzing First–Order Methods for Solving (RLM)

- One of the difficulties in understanding the convergence rates of these methods is the lack of strong convexity in Problem (RLM), which potentially results in multiple optimal solutions.
  - Example:

$$\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Ax - b\|_2^2 + \|x\|_1 \right\}.$$

- However, such difficulty can be circumvented by exploiting the nice properties of these methods and using error bounds to obtain a good estimate of an iterate's distance to the optimal set.

- Various error bound–based analysis frameworks have been developed; see, *e.g.*, **[Luo-Tseng'93, Attouch-Bolte-Svaiter'13, Bolte-Nguyen-Peypouquet-Suter'16, Li-Pong'17, S.-Zhou'17]**.

Analyzing First–Order Methods for Solving (RLM)

Let

- $F = f + P$ be the objective function,
- $F_{\min}$ be the optimal value of (RLM),
- $\mathcal{X}$ be the set of optimal solutions to (RLM), and
- $R(x) = x - \mathrm{prox}_P(x - \nabla f(x))$ be the residue (recall from (FP) that $R(x) = \mathbf{0}$ iff $x \in \mathcal{X}$).

Analyzing First–Order Methods for Solving (RLM)

- Suppose that Problem (RLM) possesses the following property:
  - **(Error Bound)** There exist $\kappa, \rho > 0$ such that

$$\mathrm{dist}(x, \mathcal{X}) \le \kappa \|R(x)\|_2 \quad \text{whenever } \mathrm{dist}(x, \mathcal{X}) \le \rho. \tag{EB}$$

- Suppose further that the FOM in question has the following properties:
  - **(Sufficient Decrease)** There exist $c_1 > 0$ such that

$$F(x^k) - F(x^{k+1}) \ge c_1 \|x^{k+1} - x^k\|_2^2. \tag{A1}$$

  - **(Cost–to–Go Estimate)** There exist $c_2 > 0$ such that

$$F(x^{k+1}) - F_{\min} \le c_2 \left(\mathrm{dist}(x^k, \mathcal{X})^2 + \|x^{k+1} - x^k\|_2^2 \right). \tag{A2}$$

  - **(Safeguard)** There exist $c_3 > 0$ such that

$$\|R(x^k)\|_2 \le c_3 \|x^{k+1} - x^k\|_2. \tag{A3}$$

- Then, the FOM in question converges linearly; *i.e.*, $\mathrm{dist}(x^{k+1}, \mathcal{X}) \le c \cdot \mathrm{dist}(x^k, \mathcal{X})$ for some $c \in (0, 1)$ **[Luo-Tseng'93]**.

Validity of Error Bounds

- Many FOMs are known to satisfy (variants) of (A1)–(A3).
- Thus, in this talk, we focus on the validity of (EB). Specifically, we are interested in the following:

Question

- Under what kind of conditions on f and P would (EB) hold?

Existing Results

Consider a general convex optimization problem of the form

$$\min_{x \in \mathcal{E}} \left\{ F(x) := f(x) + P(x) \right\}. \tag{NCP}$$

Then, Problem (NCP) possesses (EB) when

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

(b) $f(x) = h(\mathcal{A}(x))$, $P(x)$ is of polyhedral epigraph **[Luo-Tseng'92]**;

(c) $f(x) = h(\mathcal{A}(x))$, $P(x)$ is the grouped LASSO or sparse grouped LASSO regularizer **[Tseng'09, Zhang-Jiang-Luo'13]**.

Here, h is a smooth, strongly convex function whose gradient ∇h is Lipschitz continuous on any compact subset of $\mathbb{R}^N$.

Question

- Can these results be established in a unified manner?

Table of Contents

1. Introduction

2. **Sufficient Conditions for Error Bound**

3. Error Bounds for $\ell_{1,p}$–Norm Regularization

4. Error Bounds for Trace Norm Regularization

5. Conclusion

Setup

In the sequel, we focus on instances of Problem (NCP) that satisfy the following:

Assumptions

(A1). f takes the form

$$f(x) = h(\mathcal{A}(x)),$$

where $\mathcal{A} : \mathcal{E} \to \mathbb{R}^N$ is a linear operator, $h : \mathbb{R}^N \to \mathbb{R}$ is strongly convex and ∇h is Lipschitz continuous on any compact subset of $\mathrm{dom}(h)$.

(A2). $\mathcal{X}$ is non-empty and bounded.

Examples:

- Least square loss: $h(y) = \frac{1}{2}\|y - b\|_2^2$
- Logistic loss: $h(y) = \sum_{i=1}^m \log(1 + e^{y_i}) - \langle b, y \rangle, \quad b \in \{0,1\}^m$

Properties on Optimality

Recall that

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

Since h is strictly convex, we have

- **Invariance of $\mathcal{A}(x)$ over $\mathcal{X}$:**

$$\exists \bar{y} \in \mathbb{R}^m \text{ such that } \mathcal{A}(x) = \bar{y},\ \forall x \in \mathcal{X}.$$

- **Invariance of $\nabla f(x)$ over $\mathcal{X}$:**

$$\nabla f(x) = \bar{g} := \mathcal{A}^* \nabla h(\bar{y}),\ \forall x \in \mathcal{X}.$$

Proposition (Optimal Solution Set)

Suppose that Assumptions (A1) and (A2) are satisfied. There exists a pair $(\bar{y}, \bar{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ such that

$$\mathcal{X} = \{x \in \mathcal{E} \mid \mathcal{A}(x) = \bar{y},\ -\bar{g} \in \partial P(x)\}\,.$$

Error Bound with Alternative Residual Function

The characterization of $\mathcal{X}$ motivates the following alternative residual function, also known as the backward error:

$$\tilde{r}(x) := \|\mathcal{A}(x) - \bar{y}\|_2 + \mathrm{dist}(-\bar{g}, \partial P(x)).$$

Hence, we may consider the following error bound:

Error Bound with Alternative Residual Function

There exist constants $\kappa, \rho > 0$ such that

$$\mathrm{dist}(x, \mathcal{X}) \leq \kappa \cdot \tilde{r}(x) \quad \text{whenever } \mathrm{dist}(x, \mathcal{X}) \leq \rho. \tag{EBR}$$

(EB) and (EBR) are Equivalent

At first sight, (EBR) looks quite different from (EB):

$$\mathrm{dist}(x,\mathcal{X}) \le \kappa \|R(x)\|_2 \quad \text{whenever } \mathrm{dist}(x,\mathcal{X}) \le \rho. \tag{EB}$$

$$\mathrm{dist}(x,\mathcal{X}) \le \kappa \cdot \tilde{r}(x) \quad \text{whenever } \mathrm{dist}(x,\mathcal{X}) \le \rho. \tag{EBR}$$

However, we can show that

Theorem

Suppose that Assumptions (A1) and (A2) are satisfied. Then, the error bound (EB) *holds if and only if the error bound* (EBR) *holds.*

Hence, instead of dealing with (EB), it suffices to consider the validity of (EBR).

Validity of (EBR)

- Recall that the optimal solution set can be expressed as

$$\mathcal{X} = \{x \in \mathcal{E} \mid \mathcal{A}(x) = \bar{y}, \ -\bar{g} \in \partial P(x)\}$$

and the error bound (EBR) asks for the inequality

$$\mathrm{dist}(x, \mathcal{X}) \leq \kappa(\underbrace{\|\mathcal{A}(x) - \bar{y}\|_2 + \mathrm{dist}(-\bar{g}, \partial P(x))}_{\tilde{r}(x)}).$$

- Define

$$\Gamma_f(\bar{y}) := \{x \in \mathcal{E} \mid \mathcal{A}(x) = \bar{y}\}, \quad \Gamma_P(\bar{g}) := \{x \in \mathcal{E} \mid -\bar{g} \in \partial P(x)\}.$$

Then, $\mathcal{X} = \Gamma_f(\bar{y}) \cap \Gamma_P(\bar{g})$.

- Furthermore, define the residual functions for $\Gamma_f(\bar{y})$ and $\Gamma_P(\bar{g})$ by

$$r_f(x) := \|\mathcal{A}(x) - \bar{y}\|_2, \quad r_P(x) := \mathrm{dist}(-\bar{g}, \partial P(x)).$$

Obviously, it follows that $\tilde{r}(x) = r_f(x) + r_P(x)$.

Validity of (EBR)

Assume the following inequalities hold:

- $\mathrm{dist}(x, \Gamma_f(\bar{y}) \cap \Gamma_P(\bar{g})) \leq \kappa_1 \left[\mathrm{dist}(x, \Gamma_f(\bar{y})) + \mathrm{dist}(x, \Gamma_P(\bar{g}))\right]$;
- $\mathrm{dist}(x, \Gamma_f(\bar{y})) \leq \kappa_f \cdot r_f(x)$;
- $\mathrm{dist}(x, \Gamma_P(\bar{g})) \leq \kappa_P \cdot r_P(x)$.

Then, the error bound (EBR) holds:

$$\begin{aligned}
\mathrm{dist}(x, \mathcal{X}) &= \mathrm{dist}(x, \Gamma_f(\bar{y}) \cap \Gamma_P(\bar{g})) \\
&\leq \kappa_1 \left[\mathrm{dist}(x, \Gamma_f(\bar{y})) + \mathrm{dist}(x, \Gamma_P(\bar{g}))\right] \\
&\leq \kappa_1 \left[\kappa_f \cdot r_f(x) + \kappa_P \cdot r_P(x)\right] \\
&\leq \kappa_1 (\kappa_f + \kappa_P)(r_f(x) + r_P(x)) \\
&= \kappa \cdot \tilde{r}(x),
\end{aligned}$$

where we let $\kappa := \kappa_1(\kappa_f + \kappa_P)$.

However, the first and third inequalities does not come for free!

Sufficient Conditions for (EB)

Theorem

Suppose that Assumptions (A1) and (A2) are satisfied. If in addition the following two conditions are satisfied, then the error bound (EBR) *holds:*

(C1). *The collection of convex sets* $\{\Gamma_f(\bar{y}), \Gamma_P(\bar{g})\}$ *is* **boundedly linearly regular (BLR)***; i.e., there exists a constant* $\kappa > 0$ *along with a neighbourhood* $\mathcal{U}$ *of* $\Gamma_f(\bar{y}) \cap \Gamma_P(\bar{g})$ *such that for all* $x \in \mathcal{U}$*,*

$$\mathrm{dist}\,(x, \Gamma_f(\bar{y}) \cap \Gamma_P(\bar{g})) \leq \kappa\,[\mathrm{dist}\,(x, \Gamma_f(\bar{y})) + \mathrm{dist}\,(x, \Gamma_P(\bar{g}))]\,.$$

(C2). *For any* $\bar{x} \in \mathcal{X}$*, the subdifferential mapping* ∂P *is* **metrically sub–regular** *at* $\bar{x}$ *for* $-\bar{g}$*; i.e., there exists a constant* $\kappa > 0$ *along with a neighborbood* $\mathcal{U}$ *of* $\bar{x}$ *such that for all* $x \in \mathcal{U}$*,*

$$\mathrm{dist}\left(x, (\partial P)^{-1}(-\bar{g})\right) \leq \kappa \cdot \mathrm{dist}(-\bar{g}, \partial P(x)).$$

In other words, $(\partial P)^{-1}$ *is* **calm***.*

Consequently, if conditions (C1) and (C2) are satisfied, then the error bound (EB) *also holds.*

Remarks

- Note that

$$(\partial P)^{-1}(-\bar{g}) = \{x \in \mathcal{E} \mid -\bar{g} \in \partial P(x)\} = \Gamma_P(\bar{g}).$$

- Summary of the proof logic: Under Assumptions (A1) and (A2),

$$\text{(EB)} \quad \Longleftrightarrow \quad \text{(EBR)} \quad \Longleftarrow \quad \begin{cases} (C1) \\ (C2) \end{cases}$$

- In addition, we have

Fact [Bauschke-Borwein-Li'99, Corollary 3]

The collection $\{\Gamma_f(\bar{y}), \Gamma_P(\bar{g})\}$ is BLR if

$$\Gamma_f(\bar{y}) \cap \mathrm{ri}\,(\Gamma_P(\bar{g})) \neq \emptyset.$$

A Unified Framework for Establishing (EB)

- Our framework allows us to establish a number of existing error bound results in a unified manner.
- More interestingly, it leads to new error bound results.
- The validity of (EB) implies that $F = f + P$ is a so–called Kurdyka–Łojasiewicz (KL) function with exponent $1/2$. **[Li-Pong'17]**

Table of Contents

1. Introduction
2. Sufficient Conditions for Error Bound
3. **Error Bounds for $\ell_{1,p}$–Norm Regularization**
4. Error Bounds for Trace Norm Regularization
5. Conclusion

$\ell_{1,p}$–Regularization

We explore the error bound for

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

where $P : \mathbb{R}^n \to \mathbb{R}$ is the so-called $\ell_{1,p}$–norm regularizer with $p \in [1, \infty]$:

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

Here, $\|\cdot\|_p$ is the vector p-norm:

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^{l} |x_i|^p\right)^{1/p} & \text{if } p \in [1,\infty); \\ \max_i\{|x_i|\} & \text{if } p = \infty \end{cases}, \quad \forall x \in \mathbb{R}^l.$$

$\ell_{1,p}$–Regularization: Applications

Applications of $\ell_{1,p}$–regularization include:

- LASSO **[Tibshirani'96]**:

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

- Grouped LASSO **[Yuan-Lin'06, Meier-van der Geer-Bühlmann'08]**: *e.g.*, with logistic loss

$$\min_{x\in\mathbb{R}^n}\left\{\sum_{i=1}^{m}\log(1+e^{a_i^Tx})-\langle b_i,a_i^Tx\rangle+\sum_{J\in\mathcal{J}}\omega_J\|x_J\|_2\right\},\quad b\in\{0,1\}^m$$

- Multi–task feature learning **[Zhang et al.'10]**
- Multiple Kernel Learning **[Tomioka-Suzuki'10]**

Conditions (C1) and (C2)

Lemma 1 [Bounded Linear Regularity]

Let P be the $\ell_{1,p}$–norm regularizer. For any $\bar{g} \in \mathbb{R}^n$, the set $\Gamma_P(\bar{g})$ is polyhedral. Consequently, the collection $\{\Gamma_f(\bar{y}), \Gamma_P(\bar{g})\}$ is BLR for any $(\bar{y}, \bar{g}) \in \mathbb{R}^m \times \mathbb{R}^n$.

Lemma 2 [Metric Sub–Regularity]

Let P be the $\ell_{1,p}$–norm regularizer. Suppose that $p \in [1,2] \cup \{\infty\}$. Then, for any $(\bar{x}, \bar{g}) \in \mathbb{R}^n \times \mathbb{R}^n$ satisfying $-\bar{g} \in \partial P(\bar{x})$, ∂P is metrically sub–regular at $\bar{x}$ for $-\bar{g}$.

Summary: For $\ell_{1,p}$–norm regularization,

- Condition (C1) is always valid;
- Condition (C2) is valid if $p \in [1,2] \cup \{\infty\}$.

Validity of (EB)

Error Bound for $\ell_{1,p}$–Regularization

Suppose that Assumptions (A1) and (A2) are satisfied and P is the $\ell_{1,p}$–norm regularizer. Then,

- (EB) always holds when $p \in [1, 2]$ and $p = \infty$;
- (EB) fails in general when $p \in (2, \infty)$.

Remarks: "Fails in general" means that for any $p \in (2, \infty)$, we can construct an instance of $\ell_{1,p}$–regularization that satisfies Assumptions (A1) and (A2) but does not satisfy (EB).

Failure of Error Bound: An Example

Consider the following problem:

$$\min_{x\in\mathbb{R}^2} \left\{ \frac{1}{2}\|Ax-b\|^2 + \|x\|_p \right\} \tag{Q}$$

with $A = (1,0)$ and $b = 2$.

Proposition

Let $F_{\min}$ be the optimal value of Problem (Q) and $\mathcal{X}$ be its optimal solution set. Then, we have $F_{\min} = 1.5$ and

$$\mathcal{X} = \begin{cases} \{(1,0)^T\} & \textit{when } p \in [1,\infty), \\ \{(1,s)^T \mid -1 \le s \le 1\} & \textit{when } p = \infty. \end{cases}$$

- We will focus on $p \in (2,\infty)$ in the sequel.

Failure of Error Bound: An Example

- Let $\{\delta_k\}_{k\geq 0}$ be a sequence converging to zero; *i.e.*, $\delta_k = o(1)$. For simplicity, we assume that $\delta_k > 0$ for all $k \geq 0$.
- Consider the sequence $\{x^k\}_{k\geq 0}$ with

$$x_1^k := 2 - (1-\delta_k)^{\frac{1}{q}}, \quad x_2^k := \frac{2-(1-\delta_k)^{\frac{1}{q}}}{(1-\delta_k)^{\frac{1}{p}}} \cdot \delta_k^{\frac{1}{p}} + \delta_k^{\frac{1}{q}},$$

  where q is the Hölder conjugate of p. The sequence $\{x^k\}_{k\geq 0}$ converges to $\mathcal{X}$.
- For this particular sequence, one can prove that

$$x^k - \mathrm{prox}_P(x^k - \nabla f(x^k)) = (0, -\delta_k^{1/q})^T.$$

  Hence, $\|R(x^k)\|_2 = \|x^k - \mathrm{prox}_P(x^k - \nabla f(x^k))\|_2 = \delta_k^{1/q}$.

Failure of Error Bound: An Example

- Simple calculation leads to

$$\mathrm{dist}(x^k, \mathcal{X}) = \|x^k - (1,0)^T\|_2 = \Theta(\delta_k^{1/p}).$$

- The rate $\|R(x^k)\|_2$ approaches to 0 is $\Theta(\delta_k^{1/q})$.
- When $p \in (2, \infty)$, $1/p < 1/q$ and hence $\delta_k^{1/q} = o(\delta_k^{1/p})$.
- Therefore, there is no constant $\kappa > 0$ such that

$$\mathrm{dist}(x^k, \mathcal{X}) \leq \kappa \|R(x^k)\|_2.$$

Result

Error bound for Problem (Q) fails for any $p \in (2, \infty)$!

Failure of Error Bound: An Example

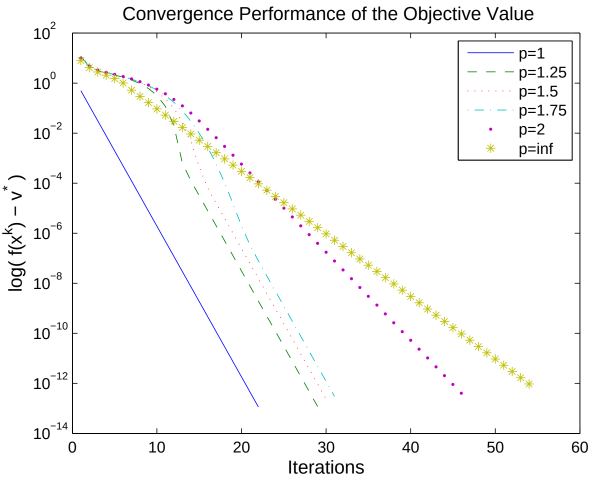


Figure: Proximal gradient method for solving (Q) with $p \in [1, 2] \cup \{\infty\}$.

Failure of Error Bound: An Example

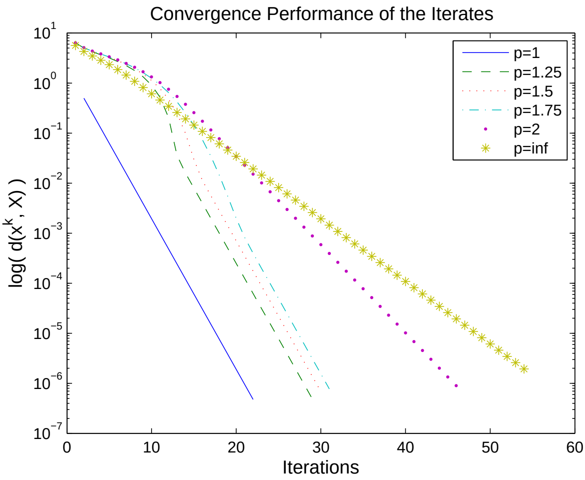


Figure: Proximal gradient method for solving (Q) with $p \in [1, 2] \cup \{\infty\}$.

Failure of Error Bound: An Example

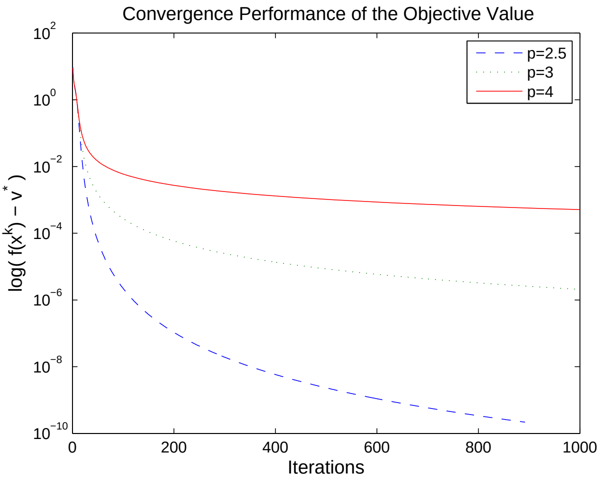


Figure: Proximal gradient method for solving (Q) with $p \in (2, \infty)$.

Failure of Error Bound: An Example

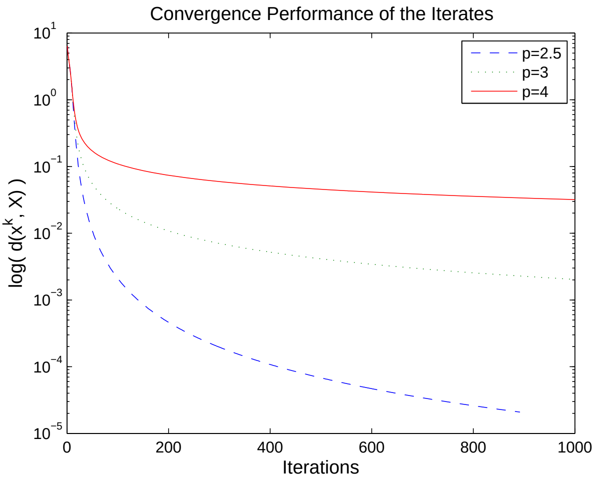


Figure: Proximal gradient method for solving (Q) with $p \in (2, \infty)$.

Table of Contents

1. Introduction
2. Sufficient Conditions for Error Bound
3. Error Bounds for $\ell_{1,p}$–Norm Regularization
4. **Error Bounds for Trace Norm Regularization**
5. Conclusion

Trace Norm Regularization

We explore the error bound for

$$\min_{X\in\mathbb{R}^{m\times n}} \left\{F(X) := f(X) + P(X)\right\},$$

where $P(X)$ is the trace norm regularizer; *i.e.*,

$$P(X) = \sum_{i=1}^{m} \sigma_i(X).$$

- WLOG, we assume that $m \leq n$.
- $\sigma_i(X)$ is the i–th singular value of X.

Conditions (C1) and (C2)

Lemma 1 [Bounded Linear Regularity]

Suppose that Assumptions (A1) and (A2) are satisfied and P is the trace norm regularizer. If there exists an $X^* \in \mathcal{X}$ such that

$$\mathbf{0} \in \nabla f(X^*) + \mathrm{ri}(\partial P(X^*)),$$

then the collection $\{\Gamma_f(\bar{y}), \Gamma_P(-\bar{g})\}$ is BLR.

Lemma 2: Metric Sub–Regularity of ∂P

Let P be the trace norm regularizer. For any two matrices $\bar{X}$ and $\bar{g}$ satisfying $-\bar{g} \in \partial P(\bar{X})$, the set-valued mapping ∂P is metrically sub–regular at $\bar{X}$ for $-\bar{g}$.

Validity of (EB)

Summary: For trace norm regularization:

- Condition (C1) is not always satisfied;
- Condition (C2) is always valid.

Error Bound for Trace Norm Regularization

Suppose that Assumptions (A1) and (A2) are satisfied. If there exists an $X^* \in \mathcal{X}$ such that

$$\mathbf{0} \in \nabla f(X^*) + \mathrm{ri}(\partial P(X^*)),$$

then the error bound for trace norm regularization holds.

Remarks on the condition:

- It can be viewed as a strict complementarity condition.
- Without such condition, examples for which the error bound fails can be constructed.

Failure of Error Bound: An Example

Consider the following problem:

$$\min_{X \in \mathbb{R}^{2\times 2}} \left\{ f(X) + \|X\|_* \right\}, \tag{R}$$

where $f(X) = h(\mathcal{A}(X))$. Moreover, we specify $\mathcal{A}$, h as follows:

- Linear operator $\mathcal{A} : \mathbb{R}^{2\times 2} \to \mathbb{R}^2$ is given by $\mathcal{A}(X) = (X_{11},\ X_{22})$.
- Strongly convex function $h : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$h(y) = \frac{1}{2}(y - B^{-1}d)^T B(y - B^{-1}d)$$

with

$$B = \begin{bmatrix} 3/2 & -2 \\ -2 & 3 \end{bmatrix} \text{ and } d = \begin{bmatrix} 5/2 \\ -1 \end{bmatrix}.$$

It is easy to prove that the optimal solution set $\mathcal{X} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\}$.

Failure of Error Bound: An Example

Suppose that $\{\delta_k\} = o(1)$. Consider a sequence $\{X^k\}_{k \geq 0}$ given by

$$X^k = \begin{bmatrix} 1 + 2\delta_k^2 & \delta_k \\ \delta_k & \delta_k^2 \end{bmatrix}.$$

- The rate at which $\{X^k\}_{k \geq 0}$ approaches $\mathcal{X}$:

$$\mathrm{dist}(X^k, \mathcal{X}) = \|X^k - \bar{X}\|_F = \Theta(\delta_k).$$

- The rate at which $\{\|R(X^k)\|_F\}_{k \geq 0}$ approaches 0:

$$\|R(X^k)\|_F = \Theta(\delta_k^2).$$

Therefore, $\|R(X^k)\|_F = o(\mathrm{dist}(X^k, \mathcal{X}))$; *i.e.*, error bound fails.

Failure of Error Bound: An Example

Reason of failure: The optimal solution set is $\{\bar{X}\}$, where $\bar{X} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

- $\nabla f(\bar{X}) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$.
- $\partial \|\bar{X}\|_* = \{Z \in \mathbb{R}^{2\times 2} \mid Z_{11} = 1,\ Z_{12} = Z_{21} = 0,\ Z_{22} \in [-1, 1]\}$.
- $\mathbf{0} \in \nabla f(\bar{X}) + \partial \|\bar{X}\|_*$. However, $\mathbf{0} \notin \nabla f(\bar{X}) + \mathrm{ri}(\partial \|\bar{X}\|_*)$.

Failure of Error Bound: An Example

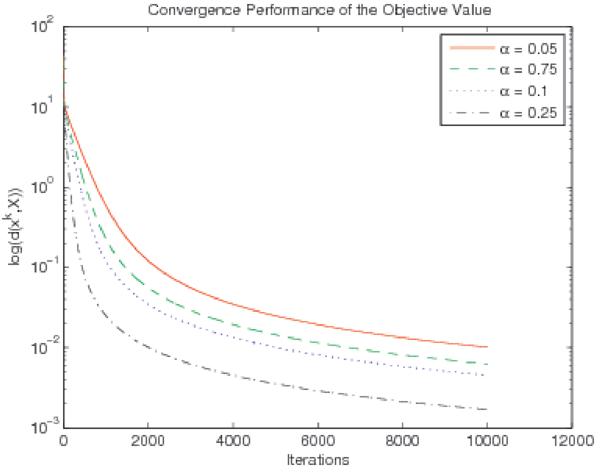


Figure: Proximal gradient method for solving (R).

Failure of Error Bound: An Example

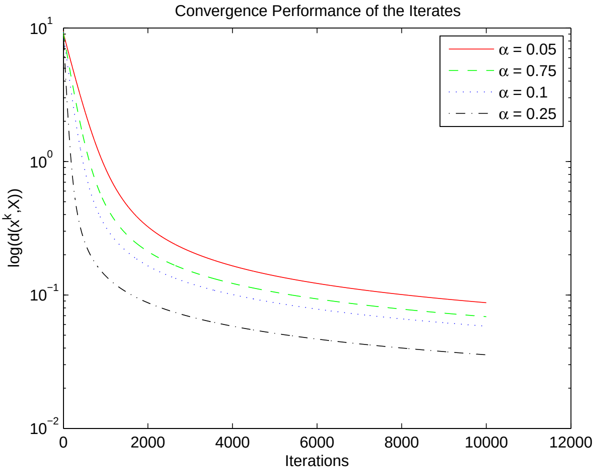


Figure: Proximal gradient method for solving (R).

Table of Contents

1. Introduction

2. Sufficient Conditions for Error Bound

3. Error Bounds for $\ell_{1,p}$–Norm Regularization

4. Error Bounds for Trace Norm Regularization

5. **Conclusion**

Take–Away

Key points of this talk:

- Error bounds provide a useful handle for analyzing the convergence rates of first–order methods, especially when the problems in question do not have a strongly convex objective.

- We establish sufficient conditions for error bound, which reduces the validity of (EB) to considering the conditions (C1) and (C2).

- Our framework provides a unified treatment of a number of existing error bound results.

- For $\ell_{1,p}$–regularization, we show that error bound holds when $p \in [1,2] \cup \{\infty\}$ while it fails in general when $p \in (2,\infty)$.

- For trace norm regularization, we prove that the error bound holds as long as a strict complementarity condition is satisfied.

Discussion

- The error bounds discussed in this talk can also be used to establish the superlinear convergence of a certain Newton–type method for solving (RLM) **[Yue-Zhou-S.'16]**.
- The error bound-based convergence analysis framework is applicable to the non–convex setting as well, though error bounds for non–convex optimization problems are generally very difficult to establish. Some recent results include:
  - Matrix completion **[Keshavan-Montanari-Oh'10]**
  - Phase retrieval **[Candès-Li-Soltanolkotabi'15]**
  - Phase synchronization **[Liu-Yue-S.'16]**
  - Quadratic optimization with orthogonality constraint **[Liu-Wu-S.'16]**

This talk is based on

- Z. Zhou, A. M.–C. So. *A Unified Approach to Error Bounds for Structured Convex Optimization Problems*. Mathematical Programming, Series A, to appear, 2016.

Thank You!