On Solutions of Sparsity Constrained Optimization

Lili Pan†, Naihua Xiu†, Shenglong Zhou†

† Beijing Jiaotong University

Speaker: Naihua Xiu

March 21, 2015

Outline

Optimality conditions for unconstrained optimization

Let us review optimality conditions for optimization problem:

$$\min f(x) \text{ s.t. } x \in \mathbb{R}^N,$$

where $f(x)$ is a first- or second-order continuously differentiable function.

First-order necessary condition: If x^* is local minimizer, then$\nabla f(x^*) = 0$. If $f(x)$ is convex, vice versa.

Second-order sufficient condition: If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$, then x^* is a local minimizer.

Second-order necessary condition: If x^* is local minimizer, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$.

Optimality conditions for convex constrained optimization

Also, we consider convex constrained problem:

$$\min f(x) \text{ s.t. } x \in \Omega,$$

where $f(x)$ is first-order continuously differentiable and Ω is convex. The following are the first-order necessary conditions.

fixed-point equation	$x^* = P_\Omega(x^* - \frac{1}{L}\nabla f(x^*)), L > 0$
Variational Inequality	$\langle x - x^*, \nabla f(x^*)\rangle \geq 0, \forall x \in \Omega$
Critical Point	$0 \in \nabla f(x^*) + N_\Omega(x^*)$
Projected gradient	$\nabla_\Omega f(x^*) = 0$

Note: They are equivalent for the above problem.

Optimality conditions for convex constrained optimization

Second-order conditions

Let $\Omega = \{g_i(x) \geq 0, i = 1, \cdots, m; h_j(x) = 0, j = 1, \cdots, l\}$ is convex, and g_i and h_j are twice continuously differentiable.
$L(x, \lambda, \mu) = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x) - \sum_{j=l}^{l} \mu_j h_j(x), \lambda_i \geq 0, i = 1, \cdots, m.$

Second-order necessary condition: If x^* is a local minimizer under some constraint qualification and there exists $(x^*, \bar{\lambda}, \bar{\mu})$ satisfying KKT system, then it holds

$$d^\top \nabla^2 L(x^*, \bar{\lambda}, \bar{\mu}) d \geq 0, \forall d \in T_\Omega(x^*).$$

Second-order sufficient condition: $x^* \in \Omega$ and there exist $(x^*, \bar{\lambda}, \bar{\mu})$ satisfying KKT system. If

$$d^\top \nabla^2 L(x^*, \bar{\lambda}, \bar{\mu}) d > 0, \forall d \in T_\Omega(x^*),$$

Then x^* is a strictly local minimizer.

Sparsity Constrained Optimization

- **Sparsity Constrained Optimization (SCO)**

$$\min \quad f(x), \qquad \text{s.t.} \quad \|x\|_0 \leq s. \tag{1}$$

where $f(x) : \mathbb{R}^N \to \mathbb{R}$ is a continuously differentiable or twice differentiable function, $\|x\|_0$ is the l_0-norm of x.

- Let $S \triangleq \{x \in \mathbb{R}^N | \; \|x\|_0 \leq s\}$. Then $S = \cup S_i$ is nonconvex, where S_i is the s-dimensional subspace. So, this problem has combinational character and is NP-hard.

Sparsity Constrained Optimization

Some first-order optimality conditions have been built for SCO [BE].

- A s-sparse vector $x^* \in S$ is called an L-stationary point of (1), if

$$x^* \in P_S\left(x^* - \frac{1}{L}\nabla f(x^*)\right), L > 0. \tag{2}$$

- A s-sparse vector $x^* \in S$ is a basic feasible vector of (1), if

$$\nabla_i f(x^*) = \begin{cases} 0, & \forall i, \text{if } \|x^*\|_0 < s, \\ 0, & i \in \text{supp}(x^*), \text{if } \|x^*\|_0 = s. \end{cases}$$

- A s-sparse vector $x^* \in S$ is called a CW-minimum of (1), if

$$f(x^*) \begin{cases} = \min_{t\in\mathbb{R}} f(x^* + te_i), & \forall i, & \text{if } \|x^*\|_0 < s, \\ \leq \min_{t\in\mathbb{R}} f(x^* - x_i^* e_i + te_j), & i \in \text{supp}(x^*),\ \forall j, & \text{if } \|x^*\|_0 = s. \end{cases}$$

[BE] Beck, A., Eldar, Y.: Sparsity constrained nonlinear optimization: optimality conditions and algorithms. SIAM J. Optim. **23**, 1480-509 (2013)

Optimality Conditions

Our questions:

- Is there any other first-order optimality conditions for SCO? If yes, What is the relationship among them?
- What are the second-order optimality conditions for SCO?

In this talk, our answer is "yes". The tools we used are tangent cone and normal cone.

Tangent Cone and Normal Cone

Definitions of Bouligand Tangent Cone and Normal Cone

For any nonempty set $\Omega \subseteq \mathbb{R}^N$, its *Bouligand Tangent Cone* $T_\Omega^B(\overline{x})$, and corresponding *Normal Cone* $N_\Omega^B(\overline{x})$ at point $\overline{x} \in \Omega$ are defined as:

$$T_\Omega^B(\overline{x}) := \left\{ d \in \mathbb{R}^N \;\middle|\; \begin{array}{l} \exists\ \{x^k\} \subset \Omega,\ \lim\limits_{k\to\infty} x^k = \overline{x},\ \lambda_k \geq 0, k = 1, \\ 2, \cdots, \text{such that} \lim\limits_{k\to\infty} \lambda_k (x^k - \overline{x}) = d \end{array} \right\},$$

$$N_\Omega^B(\overline{x}) := \left\{ d \in \mathbb{R}^N \mid \langle d, z \rangle \leq 0,\ \forall\, z \in T_\Omega^B(\overline{x}) \right\}.$$

Tangent Cone and Normal Cone

Definitions of Clarke Tangent Cone and Normal Cone

The *Clarke Tangent Cone* $T_\Omega^C(\overline{x})$ and corresponding *Normal Cone* $N_\Omega^C(\overline{x})$ at point $\overline{x} \in \Omega$ are defined as:

$$T_\Omega^C(\overline{x}) := \left\{ d \in \mathbb{R}^N \;\middle|\; \begin{array}{r} \forall\ \{x^k\} \subset \Omega,\ \forall\ \{\lambda_k\} \subset \mathbb{R}_+ \text{ with } \lim\limits_{k\to\infty} x^k = \overline{x}, \\ \lim_{k\to\infty} \lambda_k = 0, \exists\ \{y^k\} \text{ such that } \lim\limits_{k\to\infty} y^k = d \\ \text{and } x^k + \lambda_k y^k \in \Omega,\ \ k \in \mathbb{N} \end{array} \right\},$$

$$N_\Omega^C(\overline{x}) := \left\{\ d \in \mathbb{R}^N \mid \langle d, z\rangle \leq 0,\ \forall\ z \in T_\Omega^C(\overline{x})\ \right\}.$$

Tangent Cone and Normal Cone

Bouligand Tangent Cone and Normal Cone of Sparse Set

Theorem 1 For any $\overline{x} \in S$ and letting $\Gamma = \mathrm{supp}(\overline{x})$, the Bouligand tangent cone and corresponding normal cone of S at $\overline{x}$ are

$$T_S^B(\overline{x}) = \begin{cases} \mathrm{span}\{\, \mathrm{e}_i, \ \ i \in \Gamma \,\}, & \text{if } |\Gamma| = s \\ \bigcup \mathrm{span}\{\, \mathrm{e}_i, \ \ i \in \Upsilon \supseteq \Gamma, |\Upsilon| \le s \,\}, & \text{if } |\Gamma| < s \end{cases} \tag{3}$$

$$N_S^B(\overline{x}) = \begin{cases} \mathrm{span}\{\, \mathrm{e}_i, \ \ i \notin \Gamma \,\}, & \text{if } |\Gamma| = s \\ \{0\}, & \text{if } |\Gamma| < s \end{cases} \tag{4}$$

where $\mathrm{e}_i \in \mathbb{R}^N$ is a vector whose the ith component is one and others are zeros, $\mathrm{span}\{\mathrm{e}_i, i \in \Gamma\}$ denotes the subspace of $\mathbb{R}^N$ spanned by $\{\, \mathrm{e}_i, \ i \in \Gamma\}$.

Tangent Cone and Normal Cone

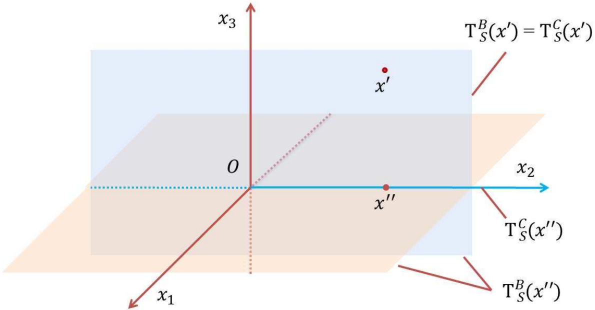
Clarke Tangent Cone and Normal Cone of Sparse Set

Theorem 2 For any $\overline{x} \in S$ and letting $\Gamma = \mathrm{supp}(\overline{x})$, then the Clarke tangent cone and corresponding normal cone of S at $\overline{x}$ are

$$T_S^C(\overline{x}) = \mathrm{span}\{\, \mathrm{e}_i, \ \ i \in \Gamma \,\}, \tag{5}$$

$$N_S^C(\overline{x}) = \mathrm{span}\{\, \mathrm{e}_i, \ \ i \notin \Gamma \,\}. \tag{6}$$

Tangent Cone and Normal Cone



Bouligand tangent cone and Clarke tangent cone in three dimensional space, where $S = \{x \in \mathbb{R}^3 | \; \|x\|_0 \leq 2\}$ and $x' = (0,1,1)^\top, x'' = (0,1,0)^\top$. One can easily verify $\mathrm{T}^B_S(x') = \mathrm{T}^C_S(x') = \{x \in \mathbb{R}^3 | \; x_1 = 0\}$, $\mathrm{T}^B_S(x'') = \{x \in \mathbb{R}^3 | \; x_1 = 0\} \cup \{x \in \mathbb{R}^3 | \; x_3 = 0\}$ and $\mathrm{T}^C_S(x'') = \{x \in \mathbb{R}^3 | \; x_1 = x_3 = 0\}$.

First-Order Optimality Conditions

N-Stability and T-Stability

Definition

A vector $x^* \in S$ is called an $N^\sharp$-stationary point and $T^\sharp$-stationary point of (1) if it respectively satisfies the relation

$$\begin{aligned} N^\sharp - \text{stationary point:} \quad & 0 \in \nabla f(x^*) + N^\sharp_S(x^*), \qquad (7)\\ T^\sharp - \text{stationary point:} \quad & 0 = \|\nabla^\sharp_S f(x^*)\|, \qquad (8)\end{aligned}$$

where $\nabla^\sharp_S f(x^*) = \arg\min\{\ \|x + \nabla f(x^*)\| \mid x \in T^\sharp_S(x^*)\ \}$, $\sharp \in \{B, C\}$ stands for the sense of Bouligand tangent cone or Clarke tangent cone.

First-Order Optimality Conditions

Theorem 3 Three kinds of stationary points under Bouligand tangent cone.

	$\|x^*\|_0 = s$	$\|x^*\|_0 < s$
L – stationary point	$\lvert(\nabla f(x^*))_i\rvert \begin{cases} = 0, & i \in \Gamma \\ \leq L\mathcal{M}_s(\lvert x^*\rvert) & i \notin \Gamma \end{cases}$	$\nabla f(x^*) = 0$
N^B – stationary point	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma \end{cases}$	$\nabla f(x^*) = 0$
T^B – stationary point	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma \end{cases}$	$\nabla f(x^*) = 0$

Remark N^B – stationary point coincides with the basic feasible vector for SCO.

First-Order Optimality Conditions

Theorem 4 Three kinds of stationary points under Clarke tangent cone.

	$\|x^*\|_0 = s$	$\|x^*\|_0 < s$
L – stationary point	$\lvert(\nabla f(x^*))_i\rvert \begin{cases} = 0, & i \in \Gamma \\ \leq L\mathcal{M}_s(\lvert x^*\rvert), & i \notin \Gamma \end{cases}$	$\nabla f(x^*) = 0$
N^C – stationary point	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma \end{cases}$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma \end{cases}$
T^C – stationary point	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma \end{cases}$	$(\nabla f(x^*))_i \begin{cases} = 0, & i \in \Gamma \\ \in \mathbb{R}, & i \notin \Gamma \end{cases}$

Remark N^C – stationary point is weaker than the basic feasible vector for SCO.

First-Order Optimality Conditions

Assumption 1 The gradient of the objective function $f(x)$ is Lipschitz with constant L_f over $\mathbb{R}^N$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \quad \forall\, x, y \in \mathbb{R}^N. \tag{9}$$

Theorem 5 If x^* is an optimal solution of (1),
(i) then x^* is an N^B-stationary point and hence N^C-stationary point.
(ii) Further, if Assumption 1 holds and $L > L_f$, then x^* is an L-stationary point of (1).

Second-Order Optimality Conditions

Now, we show second-order optimality conditions for SCO.

Theorem 6 (Second-Order Necessary Conditions) Assume $f(x)$ is twice continuously differentiable on $\mathbb{R}^N$. If $x^* \in S$ is the optimal solution of (1), we have

$$d^\top \nabla^2 f(x^*) d \geq 0, \quad \forall\, d \in \mathrm{T}_S^C(x^*) \tag{10}$$

where $\nabla^2 f(x^*)$ is the Hessian matrix of f at x^*.

Second-Order Optimality Conditions

Theorem 7 (Second-Order Sufficient Conditions) If $x^* \in S$ is an N^C-stationary point of (1) and $\nabla^2 f(x^*)$ is restricted positive definite, that is

$$d^\top \nabla^2 f(x^*) d > 0, \qquad \forall\ d \in \mathrm{T}_S^C(x^*), d \neq 0, \tag{11}$$

then x^* is the strictly local minimizer of (1). Moreover, there are $\eta > 0$ and $\delta > 0$, for any $x \in B(x^*, \delta) \cap S$, it holds

$$f(x) \geq f(x^*) + \eta \|x - x^*\|^2. \tag{12}$$

Special Case

If the problem (1) reduces to the compressed sensing, we have

$$\min f(x) := \frac{1}{2}\|Ax - b\|^2 \text{ s.t. } \|x\|_0 \leq s, \tag{13}$$

where $A \in \mathbb{R}^{M\times N}$, $b \in \mathbb{R}^M$, $L_f = \lambda_{\max}(A^\top A)$ is the largest eigenvalue of $A^\top A$.

Corollary 1 For the problem (13), if $x^* \in S$ is an N^C-stationary point and

$$d^\top A^\top Ad > 0, \quad \forall\ d \in \mathrm{T}_S^C(x^*), d \neq 0, \tag{14}$$

then x^* is the strictly local minimizer of (13).

Special Case

Definition: matrix A is s-**regular** if every s columns of A are linearly independent.

Thus, we have the following result.

Corollary 2 For the problem (13), if matrix A is s-regular, then the number of N^C-stationary points is finite, and every N^C-stationary point is uniquely local minimizer of problem (13).

Special Case

Corollary 3 For the problem (13), if matrix A is s-regular, and both A and b guarantee a unique solution to

$$\Gamma_0 \triangleq \arg\min_{|\Gamma|\leq s} \|\Pi_\Gamma b\|$$

where $\Pi_\Gamma b = A_\Gamma(A_\Gamma^\top A_\Gamma)^{-1}A_\Gamma^\top b$, then problem (13) has a unique solution.

Sufficient Conditions for Solution Uniqueness

Restricted Isometry Property (RIP)[CT]

- Matrix A obeys RIP for $0 < s < N$, if there exist a $0 < \delta_s < 1$ such that for all $\|x\|_0 \leq s$,

$$(1 - \delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_s)\|x\|_2^2. \tag{15}$$

- The RIP of matrix A makes the function $\|Ax - b\|^2$ is strongly convex and smooth in all s-dimensional subspaces.

- The RIP condition is a sufficient condition: the problem (13) has unique minimizer and is polynomially solvable.

[CT]E. J. Candés and T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203-4215.

Sufficient Conditions for Solution Uniqueness

The extension I of RIP

If $f(x)$ is continuously differentiable, we can extend RIP to RSC\ RSS:

- For any integer $s > 0$, we say $f(x)$ is restricted m_s-strongly convex and M_s-strongly smooth (RSC\RSS), if there exists m_s, $M_s > 0$ such that

$$\frac{m_s}{2}\|d\|^2 \leq f(x+d) - f(x) - \langle \nabla f(x), d \rangle \leq \frac{M_s}{2}\|d\|^2,$$
$$\forall |\mathrm{supp}(x) \cup \mathrm{supp}(d)| \leq s,$$

  where $\mathrm{supp}(x) = \{i \in \{1, \cdots, N\} \mid x_i \neq 0\}$.

[JJR]Jalali, A., Johnson, C. C., Ravikumar, P. K.: On learning discrete graphical models using greedy methods. Advances in Neural Information Processing Systems, **24**, 1935-1943. (2011)

Sufficient Conditions for Solution Uniqueness

Result I:

Suppose objective function $f(x)$ satisfies $RSC(\eta s^*)$ and $RSS(\eta s^*)$ with parameters m_s and M_s for some $\eta \geq 2 + 4\rho^2(\sqrt{(\rho^2-\rho)/s^*})^2$ with $\rho = M_s/m_s$. Moreover, suppose that the true solution x^* satisfy $\min_{j\in S^*} |x_j^*| > \sqrt{32\rho\epsilon_S/m_s}$ and $\epsilon_S \geq (s\rho\eta/m_s)s^*\lambda_n^2$, the output solution x satisfies:

$$\|x - x^*\|_2 \leq \frac{2}{m_s}\sqrt{s^*}(\lambda_n\sqrt{\eta} + \sqrt{\epsilon_S}\sqrt{2M_s}).$$

[JJR]Jalali, A., Johnson, C. C., Ravikumar, P. K.: On learning discrete graphical models using greedy methods. Advances in Neural Information Processing Systems, **24**, 1935-1943. (2011)

Sufficient Conditions for Solution Uniqueness

The extension II of RIP

When $f(x)$ is twice continuously differentiable, we can extend RIP to Stable Restricted Hessian:

- For any integer $s > 0$, we say $f(x)$ have a Stable Restricted Hessian (SRH) with constant μ_s, if
$$B_s(x)\|d\|^2 \leq d^\top \nabla^2 f(x) d \leq A_s(x)\|d\|^2,$$
$$\forall |\mathrm{supp}(x) \cup \mathrm{supp}(d)| \leq s.$$
and $1 \leq \frac{A_s(x)}{B_s(x)} \leq \mu_s$.

[BRB]Bahmani, S., Raj, B., Boufounos, P.: Greedy sparsity-constrained optimization. J. Mach. Learn. Res. **14**, 807-841 (2013)

Sufficient Conditions for Solution Uniqueness

Result II:

GraSP Algorithm:

- $\Gamma^{k+1} = supp(x^k) \cup \{the\ index\ of\ 2s\ largest\ element\ of\ \nabla f(x^k)\}$
- $\tilde{x}^{k+1} \in arg\min\{f(x),\ supp(x) \subseteq \Gamma^{k+1}\}$
- $x^{k+1} \in P_S(\tilde{x}^{k+1})$

Suppose $\{x^k\}_{k\geq 0}$ is generated by Algorithm GraSP, and the SRH holds at μ_{4s}-SRH with $\mu_{4s} \leq \frac{1+\sqrt{3}}{2}$. Furthermore, suppose that for some $\varepsilon > 0$ we have $\varepsilon \leq B_{4s}(x)$ for all $4s$-sparse vectors x. Then

$$\left\|x^k - x^*\right\|_2 \leq 2^{-k}\|x^*\|_2 + \frac{6+2\sqrt{3}}{\varepsilon}\|\nabla f(x^*)|_I\|_2,$$

where I is the position of the $3s$ largest entries of $\nabla f(x^*)$ in magnitude.

[BRB]Bahmani, S., Raj, B., Boufounos, P.: Greedy sparsity-constrained optimization. J. Mach. Learn. Res. **14**, 807-841 (2013)

Sufficient Conditions for Solution Uniqueness

The extension III of RIP

When $f(x)$ has restricted subgradient, we can extend RIP to Stable Restricted Linearization:

- We say $\nabla_s f(x)$ is a restricted subgradient of f at point x if

$$f(x+d) - f(x) \geq \langle \nabla_s f(x), d \rangle, \qquad \|d\|_0 \leq s.$$

- We say $f(x)$ have a Stable Restricted Linearization (SRL) with constant μ_s, if

$$\frac{\beta_s(x)}{2}\|d\|^2 \leq f(x+d) - f(x) - \langle \nabla_s f(y), d \rangle \leq \frac{\alpha_s(x)}{2}\|d\|^2,$$

  for $|\mathrm{supp}(x) \cup \mathrm{supp}(d)| \leq s$, and $1 \leq \frac{\alpha_s(x)}{\beta_s(x)} \leq \mu_s$.

[BRB]Bahmani, S., Raj, B., Boufounos, P.: Greedy sparsity-constrained optimization. J. Mach. Learn. Res. **14**, 807-841 (2013)

Sufficient Conditions for Solution Uniqueness

Result III:

Suppose $\{x^k\}_{k\geq 0}$ is generated by Algorithm GraSP, and the SRL holds at μ_{4s}-SRL, $\mu_{4s} \leq \frac{3+\sqrt{3}}{4}$. Furthermore, suppose that for $\varepsilon > 0$, for all $4s$-sparse x, $\varepsilon \leq \beta_{4s}(x)$. Then

$$\left\|x^k - x^*\right\|_2 \leq 2^{-k}\|x^*\|_2 + \frac{6+2\sqrt{3}}{\varepsilon}\|\nabla_I f(x^*)\|\|_2,$$

where I is the position of the $3s$ largest entries of $\nabla f(x^*)$ in magnitude.

[BRB]Bahmani, S., Raj, B., Boufounos, P.: Greedy sparsity-constrained optimization. J. Mach. Learn. Res. **14**, 807-841 (2013)

Extensions and Future Work

- Like [LZ], [BH] and [BLP], we have used our tangent and normal cone technique to study the optimality conditions of the following problems:

$$\min c^\top x, \ \ s.t. \ \ Ax \leq b, \|x\|_0 \leq s$$

$$\min \frac{1}{2} x^\top Gx + g^\top x, \ \ s.t. \ \ Qx \leq q, \|x\|_0 \leq s$$

$$\min f(x), \ \ s.t. \ \ Ax - b \in K, \|x\|_0 \leq s$$

- we will consider the dual theory of the above problems.

[LZ]Zhaosong Lu, Yong Zhang: Sparse Approximation via Penalty Decomposition Methods, SIAM Journal on Optimization, 2014.

[BH] A. Beck and N. Hallak, On the minimization over sparse symmetric sets, 2014.

[BLP]H. H. Bauschke, D. R. Luke, H. M. Phan and X. Wang, Restricted normal cones and the method of alternating projections: theory, J. Set-Valued and Variational Analysis 21:431–473 (2013).

Thank you !