Pinning Down the Łojasiewicz Exponent for Structured Non-Convex Optimization

Anthony Man-Cho So (苏文藻)
Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong

(Joint Work w/ Huikang Liu (刘慧康) & Weijie Wu (吴玮婕))

MLA'15

7 November 2015

Non-Convex Opt: Something to Avoid?

- Non-convex optimization
  - typically considered difficult
  - algorithms may get stuck in stationary points
  - convergence rates of algorithms seem difficult to establish

Non-Convex Opt: Something to Avoid?

- Non-convex optimization
  - typically considered difficult
  - algorithms may get stuck in stationary points
  - convergence rates of algorithms seem difficult to establish
- Structured non-convex optimization models
  - arise in many applications (examples follow shortly)
  - structure could be exploited for algorithm design and possibly also for analysis
  - sometimes admit faster/cheaper computation than the corresponding convex approximations

Structured Non-Convex Opt Models

- Low-Rank Matrix Recovery

$$\min_{X \in R^{m \times r},\ Y \in R^{n \times r}} \{ \| A(XY^T) - b \|_F^2 + r(X, Y) \}$$

  - Can be tackled by alternating optimization
  - Compare with the convex approximation:

$$\min_{Z \in R^{m \times n}} \{ \| A(Z) - b \|_F^2 + \| Z \|_* \}$$

    Standard algorithms (e.g., proximal gradient) require an SVD per iteration
  - Wide range of applications

Structured Non-Convex Opt Models

- Non-negative Tensor Factorization

$$\min_{X_1,\ldots,X_n \geq 0} \{ \| T - X_1 \circ \cdots \circ X_n \|_F^2 + r(X_1, \ldots, X_n) \}$$

  - Here, T is a tensor (multidimensional array) and $\circ$ represents a particular tensor decomposition (e.g., CANDECOMP/PARAFAC or Tucker)
  - Can be tackled by block coordinate descent methods
  - Applications in signal and image processing

Structured Non-Convex Opt Models

- Principal Component Analysis

$$\max_{X \in R^{m \times n} : X^T X = I} \|A^T X\|_F^2$$

  - Here, $m \gg n$ and we aim at recovering the top n left singular vectors of A
  - A fundamental matrix computation problem with long history and numerous fast algorithms

Analysis of Non-Convex Opt Problems

- In view of
  - prevalence of structured non-convex optimization models
  - availability of fast algorithms for solving them

a natural question is to understand the convergence behavior of those algorithms

- convergence (rate)
- properties of the limit point

Existing Analysis Approach

- Find conditions that guarantee fast convergence of certain algorithms to global optimum
  - assumptions on data (e.g., RIP, incoherence, etc.) and initialization of the algorithms typically required
  - need to understand the behavior of the non-convex objective function around the global optima

Existing Analysis Approach

- Find conditions that guarantee fast convergence of certain algorithms to global optimum
  - assumptions on data (e.g., RIP, incoherence, etc.) and initialization of the algorithms typically required
  - need to understand the behavior of the non-convex objective function around the global optima
  - focus of many recent efforts
    - (matrix completion) Jain et al.'13, Hardt'14, Sun-Luo'15
    - (dictionary learning) Agarwal et al.'14
    - (PCA/SVD) Shamir'15
    - …

Existing Analysis Approach

- Use the Łojasiewicz inequality to study the convergence behavior of descent methods
  - apply to a wide range of objective functions (convex or non-convex) and descent methods
  - no assumptions on the data or initialization of the algorithms required
  - typically can only establish convergence to stationary points

Existing Analysis Approach

- Use the Łojasiewicz inequality to study the convergence behavior of descent methods
  - apply to a wide range of objective functions (convex or non-convex) and descent methods
  - no assumptions on the data or initialization of the algorithms required
  - typically can only establish convergence to stationary points
  - focus of this talk

Łojasiewicz Inequality

- Let f be a real analytic function and x^* be one of its stationary points (i.e., $\nabla f(x^*) = 0$). Then, there exist $\delta, \eta > 0, \theta \in (0, 1/2]$ such that for all $y \in B(x^*, \delta)$,

$$|f(y) - f(x^*)|^{1-\theta} \leq \eta \|\nabla f(y)\|_2$$

- Thus, the Łojasiewicz inequality gives local growth information around stationary points of f.
- The growth rate is determined by the Łojasiewicz exponent θ.

Łojasiewicz Inequality: An Example

- Let $f: R^n \to R$ be given by $f(x) = \|x\|_2^2$ and $x^* = 0$. Then, $\nabla f(x^*) = 0$ and

$$\|y\|_2 = |f(y) - f(x^*)|^{1-(1/2)} \leq \frac{1}{2}\|\nabla f(y)\|_2 = \|y\|_2$$

for all $y \in R^n$

Łojasiewicz Inequality: Implications

- (Absil et al.'05) Suppose that a bounded sequence of iterates $\{x_k\}$ satisfies
  - Primary Descent: There exists $\sigma > 0$ such that for sufficiently large k ,
    $$f(x_{k+1}) - f(x_k) \leq -\sigma \|\nabla f(x_k)\|_2 \|x_{k+1} - x_k\|_2$$
  - Stationarity: For sufficiently large k ,
    $$[f(x_{k+1}) = f(x_k)] \Rightarrow [x_{k+1} = x_k]$$
  - Safeguard: There exists $\kappa > 0$ such that for sufficiently large k ,
    $$\|x_{k+1} - x_k\|_2 \geq \kappa \|\nabla f(x_k)\|_2$$

Łojasiewicz Inequality: Implications

- Then, together with the Łojasiewicz inequality, the sequence $\{x_k\}$ converges to a stationary point x^* of f .

Łojasiewicz Inequality: Implications

- Then, together with the Łojasiewicz inequality, the sequence $\{x_k\}$ converges to a stationary point x^* of f.
- Moreover, the convergence rate can be estimated as follows:
  - (Sublinear convergence) If $\theta \in (0,1/2)$, then
$$\|x_k - x^*\|_2 = O\left(k^{-\theta/(1-2\theta)}\right)$$
  - (Linear convergence) If $\theta = 1/2$, then
$$\|x_k - x^*\|_2 = e^{-O(k)}$$

Łojasiewicz Inequality: Discussion

- Many first-order methods for the unconstrained minimization of f generate iterates that satisfy the three properties.
- Real analytic functions include, e.g., all polynomials, regardless of their convexity.

Łojasiewicz Inequality: Discussion

- Many first-order methods for the unconstrained minimization of f generate iterates that satisfy the three properties.
- Real analytic functions include, e.g., all polynomials, regardless of their convexity.
- However, the Łojasiewicz exponent is difficult to estimate, even for some simple f !
  - In fact, the exponent is not known in most cases.

Łojasiewicz Inequality: Discussion

- (D'Acunto-Kurdyka'05) For a real polynomial $f: R^n \to R$ of degree $d \geq 2$, we have

$$\theta = \frac{1}{d(3d-3)^{n-1}}$$

- This leads to very weak convergence rate result.

Łojasiewicz Inequality: Discussion

- (D'Acunto-Kurdyka'05) For a real polynomial $f: R^n \to R$ of degree $d \geq 2$, we have

$$\theta = \frac{1}{d(3d-3)^{n-1}}$$

- This leads to very weak convergence rate result.

- Question: For structured problems, could we get a sharp estimate of the Łojasiewicz exponent?
  - This will lead to meaningful convergence rate results for a host of first-order methods.

Orthogonality-Constrained QPs

- Consider the following problem (denoted OCQP):

$$\min_{X \in R^{m \times n} : X^T X = I} \mathrm{Tr}(X^T A X B)$$

  - non-convex objective function and constraint
  - includes the PCA formulation as special case, as

$$\|A^T X\|_F^2 = \mathrm{Tr}(X^T A A^T X)$$

Orthogonality-Constrained QPs

- Consider the following problem (denoted OCQP):

$$\min_{X \in R^{m \times n} : X^T X = I} \mathrm{Tr}(X^T A X B)$$

  - non-convex objective function and constraint
  - includes the PCA formulation as special case, as

$$\|A^T X\|_F^2 = \mathrm{Tr}(X^T A A^T X)$$

- The feasible set is known as the Stiefel manifold, denoted by $\mathrm{St}(m, n)$.
- The manifold structure allows us to utilize techniques from manifold optimization.

Optimization on the Stiefel Manifold

- Instead of the usual gradient, we use the projected gradient, which is the usual gradient projected onto the tangent space to the manifold.
- For the Stiefel manifold, the projected gradient of $f(X) = \mathrm{Tr}(X^T AXB)$ is

$$\mathrm{grad}\, f(X) = 2AXB - XX^T AXB - XBX^T AX$$

Optimization on the Stiefel Manifold

- Instead of the usual gradient, we use the projected gradient, which is the usual gradient projected onto the tangent space to the manifold.
- For the Stiefel manifold, the projected gradient of $f(X) = \mathrm{Tr}(X^T AXB)$ is

$$\mathrm{grad}\, f(X) = 2AXB - XX^T AXB - XBX^T AX$$

- The set of stationary points is simply

$$\mathcal{X} = \{X \in \mathrm{St}(m,n) : \mathrm{grad}\, f(X) = 0\}$$

Optimization on the Stiefel Manifold

- The projected gradient allows us to treat (OCQP) like an unconstrained problem.
- (Schneider-Uschmajew'15) The convergence theorem of Absil et al.'05 carries over to the manifold setting if we replace the gradient in the Łojasiewicz inequality by the projected gradient.

Optimization on the Stiefel Manifold

- The projected gradient allows us to treat (OCQP) like an unconstrained problem.
- (Schneider-Uschmajew'15) The convergence theorem of Absil et al.'05 carries over to the manifold setting if we replace the gradient in the Łojasiewicz inequality by the projected gradient.

- Can be used to establish convergence rate results for a host of retraction-based line-search methods.

Łojasiewicz Inequality for (OCQP)

- (Liu-Wu-S.'15) There exist $\delta, \eta > 0$ such that for all $Y \in \mathrm{St}(m,n), X^* \in \mathcal{X}$ with $Y \in B(X^*, \delta)$,

$$|f(Y) - f(X^*)|^{1/2} \leq \eta \|\mathrm{grad}\, f(Y)\|_F$$

Łojasiewicz Inequality for (OCQP)

- (Liu-Wu-S.'15) There exist $\delta, \eta > 0$ such that for all $Y \in \mathrm{St}(m,n), X^* \in \mathcal{X}$ with $Y \in B(X^*, \delta)$,

$$|f(Y) - f(X^*)|^{1/2} \leq \eta \|\mathrm{grad}\, f(Y)\|_F$$

- The starting point of the proof is to show that every $X^* \in \mathcal{X}$ can be factorized as $X^* = PQ$, where columns of P (resp. Q) are eigenvectors of A (resp. B).

Łojasiewicz Inequality for (OCQP)

- (Liu-Wu-S.'15) There exist $\delta, \eta > 0$ such that for all $Y \in \mathrm{St}(m,n), X^* \in \mathcal{X}$ with $Y \in B(X^*, \delta)$,

$$|f(Y) - f(X^*)|^{1/2} \leq \eta \|\mathrm{grad}\, f(Y)\|_F$$

- The starting point of the proof is to show that every $X^* \in \mathcal{X}$ can be factorized as $X^* = PQ$, where columns of P (resp. Q) are eigenvectors of A (resp. B).
- The above inequality implies the linear convergence of any algorithm that satisfies primary descent, stationarity, and safeguard.

A Stiefel-SVRG for PCA

- The PCA formulation can be rewritten as

$$\min_{X \in R^{m \times n} : X^T X = I} -\mathrm{Tr}\left[X^T \left(\frac{1}{N} \sum_{i=1}^{N} a_i a_i^T \right) X \right]$$

where $a_1, \ldots, a_N$ are the columns of A.

- The structure of the problem motivates the design of an SVRG-type algorithm that works on the Stiefel manifold.

A Stiefel-SVRG for PCA

- Let $f_i(X) = -\mathrm{Tr}\left(X^T a_i a_i^T X\right)$. Initialize $\tilde{X}^0 \in \mathrm{St}(m,n)$.
- Outer loop $s = 1,2,\ldots$
  - set $\tilde{X} = \tilde{X}^{s-1}, \tilde{\mu} = \nabla f\left(\tilde{X}\right), X_0 = \tilde{X}$
  - Inner loop $k = 1,\ldots,T$
    - sample $i_k \in \{1,\ldots,N\}$ uniformly at random
    - compute $g_k = \nabla f_{i_k}(X_{k-1}) - \nabla f_{i_k}\left(\tilde{X}\right) + \tilde{\mu}$
    - project g_k onto tangent space at X_{k-1}
    - apply retraction to get X_k
  - set $\tilde{X}^s = X_k$, where $k \in \{1,\ldots,T\}$ is chosen uniformly at random

A Stiefel-SVRG for PCA

- Using the Łojasiewicz inequality, we can prove

(Liu-Wu-S.'15) Stiefel-SVRG converges linearly in expectation.

Conclusions

- Our work represents a very preliminary step towards understanding the Łojasiewicz exponents associated with structured non-convex optimization problems.
- Many intriguing questions remain
  - other structured non-convex (and non-smooth) models?
  - implications on algorithm design?
  - convergence to global optimum under the Łojasiewicz inequality-based framework?

Thank You!