On the Linear Convergence of the ADMM for Regularized Low-Rank Matrix Recovery

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

(Joint Work with Caihua Chen, Huikang Liu, Qi Zhang, and Zirui Zhou)

2018 International Workshop on
Modern Optimization and Applications

18 June 2018

Low-Rank Matrix Recovery

- A prototypical form of low-rank matrix recovery problems:

$$\min_{X \in \mathbb{R}^{m \times n}} f(X) \quad \text{subject to} \quad \mathrm{rank}(X) \leq r,$$

  where we assume that $f : \mathbb{R}^{m \times n} \to \mathbb{R}_+$ is a smooth convex function.

- Various applications (see the survey **[Davenport-Romberg'16]**)
  - multi-label classification
  - multi-task learning
  - network localization
  - recommender systems

- A non-convex, NP-hard problem in general.

Low-Rank Matrix Recovery: Convex Approaches

- A popular approach is to replace $\mathrm{rank}(\cdot)$ by a convex surrogate, thus yielding a convex optimization problem (see, e.g., **[Recht-Fazel-Parrilo'10, Gross'11]**). Examples include:

$$\min_{X \in \mathbb{R}^{m \times n}} f(X) \quad \text{subject to} \quad \|X\|_* \leq r,$$

$$\min_{X \in \mathbb{R}^{m \times n}} \{f(X) + \lambda \|X\|_*\}.$$

- Advantages
  - polynomial-time solvable
  - exact recovery results under certain assumptions on f
- Issue
  - can be expensive to solve when problem size is large

Low-Rank Matrix Recovery: Non-Convex Approaches

- An alternative approach is to write $X = UV^T$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$.
  - enforce the rank constraint explicitly
  - commonly used in practice

- This gives the following factored form of the low-rank matrix recovery problem:

$$\min_{U\in\mathbb{R}^{m\times r},\, V\in\mathbb{R}^{n\times r}} f(UV^T).$$

- Advantages
  - smaller variable size
  - can often be tackled by standard methods (e.g., gradient descent, alternating minimization)

- Issues
  - ambiguities caused by invertible transformation: if (U, V) is a solution, then so is $\left(UM, V(M^{-1})^T\right)$ for any invertible M
  - a non-convex formulation

Low-Rank Matrix Recovery: Non-Convex Approaches

- To address the first issue, one approach is to include a regularizer in the formulation (see, e.g., **[Koren-Bell-Volinsky'09, Ge-Lee-Ma'16, Sun-Luo'16, Tu-Boczar-Simchowitz-Soltanolkotabi-Recht'16]**):

$$\min_{U\in\mathbb{R}^{m\times r},\, V\in\mathbb{R}^{n\times r}} \left\{F(U,V) := f(UV^T) + g(U,V)\right\}, \qquad \text{(MR-F)}$$

  where $g : \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \to \mathbb{R}_+$ is a smooth regularizer. Such regularizer can also be used to induce certain desirable structure in the solution.

- Some examples of g include:
  - $g(U,V) = \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$
  - $g(U,V) = \frac{\lambda}{4}\left\|U^TU - V^TV\right\|_F^2$

Low-Rank Matrix Recovery: Non-Convex Approaches

- However, the second issue (i.e., the non-convexity of Problem (MR-F)) remains.
  - In fact, even the regularizer g is non-convex in some cases.
- Nevertheless, standard local search heuristics (e.g., gradient descent, alternating optimization) tend to work well on Problem (MR-F).
  - convergence to high-quality solution
  - fast convergence rate
- Question: Can such phenomenon be rigorously justified?

Understanding Non-Convex Low-Rank Matrix Recovery

Current approaches in the literature essentially follow two lines.

- Characterize the growth behavior of F around the set of global optima.
  - basins of attraction
  - yield convergence rate results for suitably initialized standard methods
  - Examples: **[Jain-Netrapalli-Sanghavi'13, Hardt'14, Zheng-Lafferty'15, Zhao-Wang-Liu'15, Sun-Luo'16, Park-Kyrillidis-Caramanis-Sanghavi'16, Tu-Boczar-Simchowitz-Soltanolkotabi-Recht'16]**

- Characterize the global geometry of F.
  - no spurious local minima
  - yield global convergence of suitably modified gradient descent methods
  - Example: **[Bhojanapalli-Neyshabur-Srebro'16, Ge-Lee-Ma'16, Li-Lu-Arora-Haupt-Liu-Zhao'16, Ge-Jin-Zheng'17, Park-Kyrillidis-Caramanis-Sanghavi'17, Zhu-Li-Tang-Wakin'17, Li-Zhu-Tang'18]**

ADMM for Non-Convex Low-Rank Matrix Recovery

- Despite the significant recent advances, the convergence behavior of certain practically efficient methods is still not well understood.
  - Case in point: Alternating Direction Method of Multipliers (ADMM)

- To tackle Problem (MR-F) by ADMM, we rewrite it as

$$\min_{\substack{X\in\mathbb{R}^{m\times n}\\ U\in\mathbb{R}^{m\times r},\, V\in\mathbb{R}^{n\times r}}} \{f(X)+g(U,V)\} \quad \text{subject to} \quad X=UV^T. \qquad \text{(MR-C)}$$

- **Observation**: The constraint is bi-affine in (X,U) and V. If the objective function is bi-convex in (X,U) and V, then Problem (MR-C) admits exact ADMM updates (see, e.g., **[Boyd-Parikh-Chu-Peleato-Eckstein'11]**).

ADMM for Non-Convex Low-Rank Matrix Recovery

- The augmented Lagrangian associated with Problem (MR-C) is given by

$$L_\beta(X,U,V;\Lambda) = f(X) + g(U,V) - \langle \Lambda, X - UV^T\rangle + \frac{\beta}{2}\|X - UV^T\|_F^2,$$

  where $\beta > 0$ is a given parameter.

- The ADMM updates in the k-th iteration is then given by

$$\begin{cases} U^{k+1} = \arg\min\limits_{U\in\mathbb{R}^{m\times r}} \left\{ L_\beta(X^k, U, V^k; \Lambda^k) + \frac{1}{2}\|U - U^k\|_P^2 \right\}, & \text{(1a)} \\ V^{k+1} = \arg\min\limits_{V\in\mathbb{R}^{n\times r}} \left\{ L_\beta(X^k, U^{k+1}, V; \Lambda^k) + \frac{1}{2}\|V - V^k\|_Q^2 \right\}, & \text{(1b)} \\ X^{k+1} = \arg\min\limits_{X\in\mathbb{R}^{m\times n}} L_\beta(X, U^{k+1}, V^{k+1}; \Lambda^k), & \text{(1c)} \\ \Lambda^{k+1} = \Lambda^k - \beta\left(X^{k+1} - U^{k+1}(V^{k+1})^T\right). & \text{(1d)} \end{cases}$$

  Here, $P \in \mathbb{S}_+^m$, $Q \in \mathbb{S}_+^n$ are chosen such that for some $\nu > 0$, $U \mapsto g(U,V) + \frac{1}{2}\|U\|_P^2$ and $V \mapsto g(U,V) + \frac{1}{2}\|V\|_Q^2$ are ν-strongly convex; cf. **[Fazel-Pong-Sun-Tseng'13, Han-Sun-Zhang'17]**.

Prior Convergence Analyses of the ADMM

- Existing analyses of ADMM in non-convex settings do not fully exploit the structural properties of the regularized low-rank matrix recovery problem (MR-F).
  - The works **[Hong-Luo-Razaviyayn'16, Wang-Yin-Zeng'15, Yang-Pong-Chen'17]** deal with affine, not bi-affine, constraints.
  - The work **[Hajinezhad-Shi'18]** tackles the bi-affine constraints but establishes only global subsequential convergence of the iterates to critical points.

Convergence Analysis of the ADMM: Basic Assumptions

We focus on instances of Problem (MR-F) with square loss—i.e.,

$$\min_{U\in\mathbb{R}^{m\times r},\, V\in\mathbb{R}^{n\times r}} \left\{ F(U,V) := \frac{1}{2}\|\mathcal{A}(UV^T) - b\|_2^2 + g(U,V) \right\} \qquad \text{(MR-F)}$$

for given linear operator $\mathcal{A}$ and vector b—and satisfy the following assumptions:

- The regularizer $g : \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \to \mathbb{R}_+$ is differentiable and semi-algebraic and satisfies
  - (Orthogonal Invariance) $g(U,V) = g(UR,VR)$ for any $R \in \mathbb{O}^r$.
  - (Lipschitz Continuity of Gradient) For any compact subset $\mathcal{C}$ of $\mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ and any $(U,V),(U',V') \in \mathcal{C}$, there exists a constant $L_{\mathcal{C}} > 0$ such that

$$\|\nabla g(U,V) - \nabla g(U',V')\|_F \le L_{\mathcal{C}} \|(U,V) - (U',V')\|_F.$$

- The objective function $F : \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \to \mathbb{R}_+$ is level bounded; i.e., for any $\alpha \in \mathbb{R}$, the set $\mathcal{L}_F(\alpha) := \{(U,V) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \mid F(U,V) \le \alpha\}$ is bounded (possibly empty).

Remarks on the Basic Assumptions

- Note that g is not required to be convex, though the ADMM (1) requires the existence of $P \in \mathbb{S}^m_+$, $Q \in \mathbb{S}^n_+$, and $\nu > 0$ such that

$$U \mapsto g(U,V) + \frac{1}{2}\|U\|_P^2 \quad \text{and} \quad V \mapsto g(U,V) + \frac{1}{2}\|V\|_Q^2$$

  are ν-strongly convex, so that the sub-problems can (in principle) be efficiently solved.

- The rotational invariance of g serves to remove part of the ambiguities in the factorization $X = UV^T$.

- The level boundedness of F allows us to establish the boundedness of the sequence generated by the ADMM (1).

- The semi-algebraicity of g implies that of F. As such, we can utilize the Łojasiewicz inequality-based convergence theory to establish the convergence of the ADMM (1).

Remarks on the Basic Assumptions

Question: Are there instances of Problem (MR-F) satisfying the basic assumptions?

- (S1). Matrix Factorization with Squared Frobenius Norm Regularizer

$$F(U,V) = \frac{1}{2}\|UV^T - M\|_F^2 + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right).$$

  Here, $M \in \mathbb{R}^{m\times n}$ is given.

- (S2). Matrix Sensing with Balancing Regularizer

$$F(U,V) = \frac{1}{2}\|\mathcal{A}(UV^T) - b\|_2^2 + \frac{\lambda}{4}\|U^TU - V^TV\|_F^2.$$

  Here, $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ is a given linear operator satisfying the (r,δ_r)-restricted isometry property (RIP) for some constant $\delta_r \in (0,1)$ (i.e., the inequalities $(1-\delta_r)\|X\|_F^2 \le \|\mathcal{A}(X)\|_2^2 \le (1+\delta_r)\|X\|_F^2$ hold for any matrix $X \in \mathbb{R}^{m\times n}$ of rank at most r), $b \in \mathbb{R}^p$ is a given vector.

Remarks on the Basic Assumptions

- When the objective function takes the form

$$F(U,V) = \frac{1}{2}\|UV^T - M\|_F^2 + g(U,V)$$

and $\beta = 1$, the ADMM (1) reduces to the two-block BCD method, as the updates (1a) and (1b) become

$$\begin{cases} U^{k+1} = \arg\min\limits_{U\in\mathbb{R}^{m\times r}} \left\{ \frac{1}{2}\|U(V^k)^T - M\|_F^2 + g(U,V^k) + \frac{1}{2}\|U-U^k\|_P^2 \right\}, \\ V^{k+1} = \arg\min\limits_{V\in\mathbb{R}^{n\times r}} \left\{ \frac{1}{2}\|U^{k+1}V^T - M\|_F^2 + g(U^{k+1},V) + \frac{1}{2}\|V-V^k\|_Q^2 \right\}. \end{cases}$$

- Our convergence results apply to this case as well.

Global Convergence Analysis of the ADMM

- Let $U^0 \in \mathbb{R}^{m\times r}$ and $V^0 \in \mathbb{R}^{n\times r}$ be arbitrary. Consider the initialization

$$X^0 = U^0(V^0)^T, \quad \Lambda^0 = \mathcal{A}^*(\mathcal{A}(X^0) - b). \tag{2}$$

  Let $Z^k = (X^k, U^k, V^k; \Lambda^k)$, $k = 0, 1, \ldots,$ be the iterates generated by the ADMM (1).

- Our immediate goal is to establish the convergence of $\{Z^k\}_{k\geq 0}$.

- Based on the assumptions made, one can show

  Proposition. (Sequence Boundedness) Suppose that $\beta \geq \|\mathcal{A}\|^2$. Then, $\{Z^k\}_{k\geq 0}$ is bounded.

- To proceed, we utilize the convergence theory developed in **[Attouch-Bolte-Svaiter'13]**.

Global Convergence Analysis of the ADMM

- **Fact: [Attouch-Bolte-Svaiter'13]** Let $h : \mathbb{R}^\ell \to \mathbb{R}$ be a differentiable function satisfying the Kurdyka-Łojasiewicz property at its critical points. Furthermore, suppose that $\{y^k\}_{k\geq 0}$ is a bounded sequence satisfying the following properties:
  - (Sufficient Decrease) There exists a constant $a > 0$ such that for $k = 0, 1, \ldots,$
  $$h(y^{k+1}) - h(y^k) \leq -a\|y^{k+1} - y^k\|_2^2.$$
  - (Safeguard) There exists a constant $b > 0$ such that for $k = 0, 1, \ldots,$
  $$\|\nabla h(y^{k+1})\|_2 \leq b\|y^{k+1} - y^k\|_2.$$

  Then, the sequence $\{y^k\}_{k\geq 0}$ converges to a critical point of h.

- Task: We need to find an appropriate h for our setting.

Global Convergence Analysis of the ADMM

- By letting $L_\beta^k = L_\beta(Z^k)$, one can show

  Proposition. (Sufficient Decrease) Suppose that $\beta \geq \|\mathcal{A}\|^2$. There exist constants $p, \tau > 0$ such that for $k = 1, 2, \ldots,$

$$\left(L_\beta^{k+1} + \frac{p}{2}\|X^{k+1} - X^k\|_F^2\right) - \left(L_\beta^k + \frac{p}{2}\|X^k - X^{k-1}\|_F^2\right)$$
$$\leq -\frac{\tau}{2}\left(\|X^{k+1} - X^k\|_F^2 + \|X^k - X^{k-1}\|_F^2 + \|U^{k+1} - U^k\|_F^2 + \|V^{k+1} - V^k\|_F^2\right).$$

- This motivates us to define the following h:

$$h(X, X', U, V; \Lambda) = L_\beta(X, U, V; \Lambda) + \frac{p}{2}\|X - X'\|_F^2,$$

  where $p > 0$ is the constant in the above proposition.

- Note that h is differentiable. Moreover, our assumption implies that h is semi-algebraic, thus satisfying the Kurdyka-Łojasiewicz property at its critical points.

Global Convergence Analysis of the ADMM

- Let $y^k = (X^k, X^{k-1}, U^k, V^k; \Lambda^k)$, where $k = 1, 2, \ldots$.

- Using the optimality conditions of the ADMM updates (1a)–(1d), one can show that $\Lambda^{k+1} - \Lambda^k = \mathcal{A}^*\mathcal{A}(X^{k+1} - X^k)$. Hence, the above sufficient decrease property can be expressed as

$$h(y^{k+1}) - h(y^k) \leq -a\|y^{k+1} - y^k\|_F^2$$

  for some constant $a > 0$.

- Furthermore, one can show that the safeguard property holds for the function h we defined.

- By utilizing the convergence theory developed in **[Attouch-Bolte-Svaiter'13]** and comparing the critical points of F and h, we obtain the following

  Theorem. (Global Convergence of the ADMM) Suppose that $\beta \geq \|\mathcal{A}\|^2$ and $(X^0, U^0, V^0; \Lambda^0)$ is initialized according to (2). Then, the sequence $(U^k, V^k)_{k\geq 0}$ converges to a critical point of F.

Local Convergence Analysis of the ADMM

- Our next goal is to study the local convergence behavior of the ADMM when the initial point lies in a suitably chosen neighborhood of the optimal solution set $\mathcal{W}$.

- Motivated by the orthogonal invariance of F, we use the following to measure the distance between a point (U, V) and an optimal solution $(U^\star, V^\star)$:

$$\mathrm{dist}((U,V),(U^\star,V^\star)) = \min_{R\in\mathbb{O}^r} \|(U,V) - (U^\star R, V^\star R)\|_F.$$

  We can then define neighborhoods of $\mathcal{W}$ via

$$\mathrm{dist}((U,V),\mathcal{W}) = \inf_{(U^\star,V^\star)\in\mathcal{W}} \mathrm{dist}((U,V),(U^\star,V^\star)).$$

Local Convergence Analysis of the ADMM: Assumption

- To determine the convergence rate of the ADMM, one typically needs to impose a growth condition on certain function related to F.

- We assume that F satisfies the Łojasiewicz inequality with exponent $1/2$ at any $(U^\star, V^\star) \in \mathcal{W}$; i.e., there exist constants $\delta, c > 0$ such that for any $(U, V) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ and $(U^\star, V^\star) \in \mathcal{W}$ satisfying $\mathrm{dist}((U, V), (U^\star, V^\star)) \leq \delta$,

$$|F(U, V) - F(U^\star, V^\star)|^{1/2} \leq c\|\nabla F(U, V)\|_F.$$

Remarks on the Łojasiewicz Inequality Assumption

- It is important to note that we require the objective function F to satisfy the Łojasiewicz inequality, not the augmented Lagrangian L_β.

- It is possible to establish the Łojasiewicz inequality with exponent $1/2$ (and explicitly given constants $\delta, c > 0$) for various F's, such as the ones below:

$$F(U,V) = \frac{1}{2}\|UV^T - M\|_F^2 + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right), \tag{S1}$$

$$F(U,V) = \frac{1}{2}\|\mathcal{A}(UV^T) - b\|_2^2 + \frac{\lambda}{4}\|U^TU - V^TV\|_F^2. \tag{S2}$$

  The Łojasiewicz inequality with exponent $1/2$ for (S1) is new, while that for (S2) can be deduced essentially from the results in **[Zhu-Li-Tang-Wakin'17]**.

- However, it is typically much more difficult to establish the Łojasiewicz inequality with an explicit exponent for the corresponding L_β's, as they involve the dual variable Λ.

Remarks on the Łojasiewicz Inequality Assumption

- Since the ADMM updates involve both the primal variables (X, U, V) and the dual variable Λ, why is it sufficient to just assume that F satisfies the Łojasiewicz inequality?

- Key Observation: The cost-to-go of the augmented Lagrangian L_β can be controlled by that of the objective function F. Specifically,

  Proposition. (Cost-to-Go Estimate) Let $(U^\star, V^\star) \in \mathcal{W}$ and set $X^\star = U^\star (V^\star)^T$, $\Lambda^\star = \mathcal{A}^*(\mathcal{A}(X^\star) - b)$, $Z^\star = (X^\star, U^\star, V^\star; \Lambda^\star)$. Then, for $k = 0, 1, \ldots$,

$$L_\beta^{k+1} - L_\beta(Z^\star) \leq F(U^{k+1}, V^{k+1}) - F(U^\star, V^\star) + \frac{\|\mathcal{A}\|^4}{2\beta} \|X^{k+1} - X^k\|_F^2.$$

Local Convergence Analysis of the ADMM

- Consequently, we obtain the following

  Theorem. (Linear Convergence of the ADMM) Suppose that $\beta \geq \|\mathcal{A}\|^2$. Furthermore, suppose that $(X^0, U^0, V^0; \Lambda^0)$ is initialized according to (2) and satisfies $\mathrm{dist}((U^0, V^0), \mathcal{W}) \leq \delta_0$ for some constant $\delta_0 > 0$. Then, there exist constants $\gamma > 0$, $\rho \in (0, 1)$ such that for $k = 0, 1, \ldots,$

$$\mathrm{dist}((U^k, V^k), \mathcal{W}) \leq \gamma \cdot \rho^k;$$

  i.e., the sequence $\{(U^k, V^k)\}_{k \geq 0}$ will converge linearly to an optimal solution to Problem (MR-F).

Closing Remarks

- We identified several structural properties of a class of non-convex regularized low-rank matrix recovery problems that would imply the global convergence and local linear convergence of the ADMM.

- We also exhibited two concrete instances that possess such properties.

- An interesting direction is to study the geometric properties of other structured non-convex optimization problems (such as those that arise in machine learning and signal processing) and exploit them in the design and analysis of fast methods.

Thank You!