Inertial block majorization minimization for multiblock composite optimization problems

Duy Nhat Phan
University of Massachusetts Lowell

Joint work with Le Thi Khanh Hien and Nicolas Gillis

24th Midwest Optimization Meeting, Waterloo, Ontario, Oct. 28-29, 2022

1. Multiblock optimization problem and BCD
2. TITAN - an inertial block majorization minimization framework
3. TITAN with composite surrogate and its application to matrix completion
4. Numerical results

Matrix completion

We would like to predict how much someone is going to like a product based on their product preferences:

John	5	1	3	5
Tom	?	?	?	2
Alice	4	?	3	?

Matrix completion

Given a data matrix $A \in \mathbb{R}^{m \times n}$ and a positive integer r (factorization rank), we would like to find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ by

$$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{r \times n}} \left\{ \frac{1}{2} \|\mathcal{P}(A - UV)\|_F^2 + \mathcal{R}(U, V) \right\},$$

where $\mathcal{R}$ is a regularization term, and $\mathcal{P}(Z)_{ij} = Z_{ij}$ if A_{ij} is observed and is equal to 0 otherwise.

Nonnegative Matrix Factorization (NMF)

Given a data matrix $X \in \mathbb{R}^{m \times n}$ and a positive integer r, find

$$\min_{W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}} \quad \frac{1}{2}||X - WH||_F^2 := \frac{1}{2}\sum_{ij}(X - WH)_{ij}^2.$$

Nonnegative Matrix Factorization (NMF)

Given a data matrix $X \in \mathbb{R}^{m \times n}$ and a positive integer r, find

$$\min_{W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}} \quad \frac{1}{2}||X - WH||_F^2 := \frac{1}{2}\sum_{ij}(X - WH)_{ij}^2.$$

- NMF can be rewritten as $\min_{W,H} f(W,H) + g_1(W) + g_2(H)$, where $f(W,H) = \frac{1}{2}\|X - WH\|_F^2$, $g_1(W) = \mathcal{I}_{\mathbb{R}_+^{m \times r}}(W)$, and $g_2(H) = \mathcal{I}_{\mathbb{R}_+^{r \times n}}(H)$.

Nonnegative Matrix Factorization (NMF)

Given a data matrix $X \in \mathbb{R}^{m \times n}$ and a positive integer r, find

$$\min_{W \in \mathbb{R}_+^{m \times r}, H \in \mathbb{R}_+^{r \times n}} \quad \frac{1}{2}||X - WH||_F^2 := \frac{1}{2}\sum_{ij}(X - WH)_{ij}^2.$$

- NMF can be rewritten as $\min_{W,H} f(W,H) + g_1(W) + g_2(H)$,
  where $f(W,H) = \frac{1}{2}\|X - WH\|_F^2$, $g_1(W) = \mathcal{I}_{\mathbb{R}_+^{m \times r}}(W)$, and $g_2(H) = \mathcal{I}_{\mathbb{R}_+^{r \times n}}(H)$.
- NMF: $\min_{W_{:i}, H_{i:}} f(W_{:i}, H_{i:}) + \sum_{i=1}^{r} g_i(W_{:i}) + \sum_{i=r+1}^{2r} g_i(H_{i:})$,
  where $f(W_{:i}, H_{i:}) = \frac{1}{2}\left\|X - \sum_{i=1}^{r} W_{:i}H_{i:}\right\|_F^2$, $g_i(W_{:i}) = \mathcal{I}_{\mathbb{R}_+^m}(W_{:i})$, $i = 1, \ldots, r$, and $g_{i+r}(H_{i:}) = \mathcal{I}_{\mathbb{R}_+^n}(H_{i:})$, $i = 1, \ldots, r$.

Multiblock optimization problem

$$\begin{aligned} \min_{x} \quad & F(x) := f(x_1, \ldots, x_m) + \sum_{i=1}^{m} g_i(x_i) \\ \text{subject to} \quad & x_i \in \mathcal{X}_i \text{ for } i = 1, \ldots, m, \end{aligned} \tag{1}$$

where

- $\mathcal{X}_i$ is a closed convex set of a finite dimensional real linear space $\mathbb{E}_i$,
- x can be decomposed into m blocks $x = (x_1, \ldots, x_m)$ with $x_i \in \mathcal{X}_i$,
- $f : \mathcal{X} \to \mathbb{R}$ is continuous but possibly non-smooth non-convex,
- $g_i(\cdot)$ is a proper and lower semi-continuous function (possibly with extended values), and
- we assume $\mathrm{dom} g_i \cap \mathcal{X}_i$ is a non-empty closed set and F is bounded from below.

We denote by $\mathcal{X} := \prod_{i=1}^{m} \mathcal{X}_i$.

Block Coordinate Descent Methods

Multiblock optimization problem

$$\begin{aligned} \min_{x} \quad & F(x) := f(x_1, \ldots, x_m) + \sum_{i=1}^{m} g_i(x_i) \\ \text{subject to} \quad & x_i \in \mathcal{X}_i \text{ for } i = 1, \ldots, m. \end{aligned}$$

1: **Initialize**: Choosing initial point x^0 and other parameters.
2: **for** $k = 0, \ldots$ **do**
3: **for** $i = 1, \ldots, m$ **do**
4: Fix the latest values of the blocks $j \neq i$: $\left(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^{k}, \ldots, x_m^{k}\right)$
5: Update block i to get $\left(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^{k}, \ldots, x_m^{k}\right)$
6: **end for**
7: **end for**

Block Coordinate Descent Methods

Denote $f_i^k(x_i) := f\left(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k\right)$. BCD methods can typically be classified into three categories:

1. Classical BCD (also known as Alternating Optimization) methods update each block of variables as follows

$$x_i^{k+1} \in \underset{x_i \in \mathcal{X}_i}{\operatorname{argmin}} f_i^k(x_i) + g_i(x_i).$$

2. Proximal BCD methods update each block of variables as follows

$$x_i^{k+1} \in \underset{x_i \in \mathcal{X}_i}{\operatorname{argmin}} f_i^k(x_i) + \frac{1}{2\beta_i^k}\left\|x_i - x_i^k\right\|^2 + g_i(x_i).$$

3. Proximal gradient BCD methods update each block of variables as follows

$$x_i^{k+1} \in \underset{x_i \in \mathcal{X}_i}{\operatorname{argmin}} f_i^k\left(x_i^k\right) + \left\langle \nabla f_i^k\left(x_i^k\right), x_i - x_i^k \right\rangle + \frac{1}{2\beta_i^k}\left\|x_i - x_i^k\right\|^2 + g_i(x_i).$$

Block surrogate function

Given $y \in \mathcal{X}$, a function $u_i(\cdot, y) : \mathcal{X}_i \to \mathbb{R}$ is called a block i surrogate function of f if $u_i(x_i, y)$ is lower semi-continuous in x_i and the following conditions are satisfied:

(a) $u_i(y_i, y) = f(y)$,

(b) $u_i(x_i, y) \geq f(x_i, y_{\neq i})$ for all $x_i \in \mathcal{X}_i$, where

$$f(x_i, y_{\neq i}) := f(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_m).$$

Block surrogate function

Given $y \in \mathcal{X}$, a function $u_i(\cdot, y) : \mathcal{X}_i \to \mathbb{R}$ is called a block i surrogate function of f if $u_i(x_i, y)$ is lower semi-continuous in x_i and the following conditions are satisfied:

(a) $u_i(y_i, y) = f(y)$,

(b) $u_i(x_i, y) \geq f(x_i, y_{\neq i})$ for all $x_i \in \mathcal{X}_i$, where

$$f(x_i, y_{\neq i}) := f(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_m).$$

Majorization-Minimization for Multiblock Optimization Problems:

$$\min_{x_i \in \mathcal{X}_i} f(x_1, \ldots, x_m) + \sum_{i=1}^{m} g_i(x_i).$$

Block MM update: $x_i^{k+1} \in \operatorname{argmin}_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$,
where $x^{k,i} = (x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k)$ for $i = 1, \ldots, m$, $x^{k,0} = x^k$.

Example. Suppose the block function $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-smooth. The descent lemma gives us

$$f(x_i, y_{\neq i}) \leq \underbrace{f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{L_i^{(y)}}{2} \|x_i - y_i\|^2}_{u_i(x_i, y)-\text{Lipschitz gradient surrogate}}$$

Example. Suppose the block function $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-smooth. The descent lemma gives us

$$f(x_i, y_{\neq i}) \leq \underbrace{f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{L_i^{(y)}}{2} \|x_i - y_i\|^2}_{u_i(x_i, y)-\text{Lipschitz gradient surrogate}}$$

Proximal gradient BCD method: $x_i^{k+1} \in \operatorname{argmin}_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

Example. Suppose the block function $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-smooth. The descent lemma gives us

$$f(x_i, y_{\neq i}) \leq \underbrace{f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{L_i^{(y)}}{2} \|x_i - y_i\|^2}_{u_i(x_i, y)-\text{Lipschitz gradient surrogate}}$$

Proximal gradient BCD method: $x_i^{k+1} \in \operatorname{argmin}_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

Proximal gradient BCD for NMF

$\min_{W,H} f(W,H) + \mathcal{I}_{\mathbb{R}_+^{m \times r}}(W) + \mathcal{I}_{\mathbb{R}_+^{r \times n}}(H)$, where $f(W,H) = \frac{1}{2}\|X - WH\|_F^2$.

Example. Suppose the block function $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-smooth. The descent lemma gives us

$$f(x_i, y_{\neq i}) \leq \underbrace{f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{L_i^{(y)}}{2} \|x_i - y_i\|^2}_{u_i(x_i, y)-\text{Lipschitz gradient surrogate}}$$

Proximal gradient BCD method: $x_i^{k+1} \in \operatorname{argmin}_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

Proximal gradient BCD for NMF

$\min_{W,H} f(W,H) + \mathcal{I}_{\mathbb{R}_+^{m \times r}}(W) + \mathcal{I}_{\mathbb{R}_+^{r \times n}}(H)$, where $f(W,H) = \frac{1}{2}\|X - WH\|_F^2$.

Proximal gradient BCD for NMF:

$$W^{k+1} = \max\left\{W^k - \frac{1}{L_1^k}\nabla_W f(W^k, H^k), 0\right\},$$

$$H^{k+1} = \max\left\{H^k - \frac{1}{L_2^k}\nabla_H f(W^{k+1}, H^k), 0\right\}$$

Proximal gradient BCD for MCP

MCP:

$$\min_{U \in \mathbb{R}^{\mathbf{m}\times\mathbf{r}}, V \in \mathbb{R}^{\mathbf{r}\times\mathbf{n}}} \psi(U,V) + \mathcal{R}(U,V),$$

where $\psi(U,V) = \frac{1}{2}\|\mathcal{P}(A - UV)\|_F^2$, and we take

$$\mathcal{R}(U,V)^1 = \lambda\Big(\sum_{ij}\big(1 - \exp(-\theta|u_{ij}|)\big) + \sum_{ij}\big(1 - \exp(-\theta|v_{ij}|)\big)\Big).$$

[1] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In Proceeding of international conference on machine learning ICML'98, 1998.

Proximal gradient BCD for MCP

MCP:

$$\min_{U\in\mathbb{R}^{\mathbf{m}\times\mathbf{r}},V\in\mathbb{R}^{\mathbf{r}\times\mathbf{n}}} \psi(U,V)+\mathcal{R}(U,V),$$

where $\psi(U,V)=\frac{1}{2}\|\mathcal{P}(A-UV)\|_F^2$, and we take

$$\mathcal{R}(U,V)^1=\lambda\Big(\sum_{ij}\big(1-\exp(-\theta|u_{ij}|)\big)+\sum_{ij}\big(1-\exp(-\theta|v_{ij}|)\big)\Big).$$

Proximal gradient BCD for MCP:

$$U^{k+1}\in\arg\min_U\left\langle\nabla_U\psi(U^k,V^k),U\right\rangle+\frac{L_1^k}{2}\|U-U^k\|^2+\lambda\sum_{ij}\Big(1-\exp(-\theta|u_{ij}|)\Big)$$

$$V^{k+1}\in\arg\min_V\left\langle\nabla_V\psi(U^{k+1},V^k),V\right\rangle+\frac{L_2^k}{2}\|V-V^k\|^2+\lambda\sum_{ij}\Big(1-\exp(-\theta|v_{ij}|)\Big)$$

[1] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In Proceeding of international conference on machine learning ICML'98, 1998.

Question

- Can we design a block MM algorithm with convergence guarantee, which applied to MCP have closed-form updates?
- Can we incorporate acceleration techniques into block MM algorithms?

TITAN with cyclic update

Require: Choose $x^{-1}, x^0 \in \mathcal{X}$ (x^{-1} can be chosen equal to x^0).
Ensure: x^k that approximately solves (1).

1: **for** $k = 0, 1, \ldots$ **do**
2: **for** $i = 1, ..., m$ **do**
3: Choose a block i surrogate function u_i of f and an extrapolation $\mathcal{G}_i^k(x_i^k, x_i^{k-1})$.
4: Update block i by

$$x_i^{k+1} \in \underset{x_i \in \mathcal{X}_i}{\operatorname{argmin}}\, u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i). \quad (2)$$

5: **end for**
6: Set $x^{k+1} = x^{k,m}$.
7: **end for**

Example. Suppose $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-Lipschitz smooth. Lipschitz gradient surrogate: $u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{\kappa_i L_i^{(y)}}{2} \|x_i - y_i\|^2$.

Example. Suppose $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-Lipschitz smooth. Lipschitz gradient surrogate: $u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{\kappa_i L_i^{(y)}}{2} \|x_i - y_i\|^2$.

Proximal gradient BCD: $x_i^{k+1} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

TITAN: $x_i^{k,i} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i)$.

Example. Suppose $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-Lipschitz smooth. Lipschitz gradient surrogate: $u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{\kappa_i L_i^{(y)}}{2} \|x_i - y_i\|^2$.

Proximal gradient BCD: $x_i^{k+1} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

TITAN: $x_i^{k,i} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i)$.

Let $\mathcal{X}_i = \mathbb{E}_i$ and
$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \nabla_i f(x^{k,i-1}) - \nabla_i f(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1})$,
where $\bar{x}_i^k = x_i^k + \tau_i^k (x_i^k - x_i^{k-1})$, τ_i^k, β_i^k are some extrapolation parameters and $L_i^k = L_i^{(x^{k,i-\mathbf{1}})}$.

Example. Suppose $x_i \mapsto f(x_i, y_{\neq i})$ is $L_i^{(y)}$-Lipschitz smooth. Lipschitz gradient surrogate: $u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{\kappa_i L_i^{(y)}}{2} \|x_i - y_i\|^2$.

Proximal gradient BCD: $x_i^{k+1} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

TITAN: $x_i^{k,i} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i)$.

Let $\mathcal{X}_i = \mathbb{E}_i$ and
$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \nabla_i f(x^{k,i-1}) - \nabla_i f(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1})$,
where $\bar{x}_i^k = x_i^k + \tau_i^k (x_i^k - x_i^{k-1})$, τ_i^k, β_i^k are some extrapolation parameters and $L_i^k = L_i^{(x^{k,i-\mathbf{1}})}$.

Proximal gradient BCD:

$$x_i^{k+1} \in \operatorname{argmin}_{x_i} \left\langle \nabla_i f(x_i^k, x_{\neq i}^{k,i-1}), x_i \right\rangle + \frac{\kappa_i L_i^k}{2} \left\| x_i - x_i^k \right\|^2 + g_i(x_i).$$

TITAN:

$$x_i^{k+1} \in \operatorname{argmin}_{x_i} \left\langle \nabla_i f(\bar{x}_i^k, x_{\neq i}^{k,i-1}), x_i \right\rangle + \frac{\kappa_i L_i^k}{2} \left\| x_i - \hat{x}_i^k \right\|^2 + g_i(x_i),$$

where $\bar{x}_i^k = x_i^k + \tau_i^k (x_i^k - x_i^{k-1})$ and $\hat{x}_i^k = x_i^k + \beta_i^k (x_i^k - x_i^{k-1})$.

Lipschitz gradient surrogate:

$$u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{\kappa_i L_i^{(y)}}{2} \|x_i - y_i\|^2.$$

Proximal gradient BCD: $x_i^{k+1} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

TITAN: $x_i^{k,i} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i)$.

Lipschitz gradient surrogate:

$$u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{\kappa_i L_i^{(y)}}{2} \|x_i - y_i\|^2.$$

Proximal gradient BCD: $x_i^{k+1} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

TITAN: $x_i^{k,i} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i)$.

Let $\mathcal{X}_i = \mathbb{E}_i$ and

$$\mathcal{G}_i^k = \alpha_i^k \left(\nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1}) - \nabla_i f(x^{k,i-1}) \right) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1}),$$

where α_i^k and β_i^k are some extrapolation parameters.

Lipschitz gradient surrogate:

$$u_i(x_i, y) = f(y) + \langle \nabla_i f(y), x_i - y_i \rangle + \frac{\kappa_i L_i^{(y)}}{2} \|x_i - y_i\|^2.$$

Proximal gradient BCD: $x_i^{k+1} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) + g_i(x_i)$.

TITAN: $x_i^{k,i} \in \arg\min_{x_i \in \mathcal{X}_i} u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i)$.

Let $\mathcal{X}_i = \mathbb{E}_i$ and

$$\mathcal{G}_i^k = \alpha_i^k \left(\nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1}) - \nabla_i f(x^{k,i-1}) \right) + \kappa_i L_i^k \beta_i^k (x_i^k - x_i^{k-1}),$$

where α_i^k and β_i^k are some extrapolation parameters.

TITAN: Inertial Block Proximal with Hessian damping [2] [a]

$$x_i^{k+1} \in \operatorname*{argmin}_{x_i} \left\langle \nabla_i f(x^{k,i-1}) + \alpha_i^k \left(\nabla_i f(x^{k,i-1}) - \nabla_i f(x_i^{k-1}, x_{\neq i}^{k,i-1}) \right), x_i \right\rangle + \frac{\kappa_i L_i^k}{2} \left\| x_i - (x_i^k + \beta_i^k (x_i^k - x_i^{k-1})) \right\|^2 + g_i(x_i).$$

[a] [2] S. Adly and H. Attouch. Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping. SIAM Journal on Optimization, 30(3):2134– 2162, 2020.

Main results (see ▸ the paper for details)

- Sub-sequential convergence.
- Global convergence.
- TITAN with proximal surrogate functions.
- TITAN with Lipschitz gradient surrogate functions.
  - TITAN recovers the Nesterov type acceleration as in [5, 6]. [2] [3]
  - TITAN recovers the inertial block proximal gradient method that uses two different extrapolation points in [4].
  - TITAN leads to a new inertial block proximal gradient algorithm with Hessian damping, that is a multiblock version of [7]. [4]
- TITAN with Bregman surrogate functions.
- TITAN with quadratic surrogate functions.
- TITAN leads to new inertial multi-block algorithms with composite surrogate functions.

---

[2] [5] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM Journal on Imaging Sciences, 6(3):1758–1789, 2013.

[3] [6] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. Journal of Scientific Computing, 72(2):700–734, 2017.

[4] [7] S. Adly and H. Attouch. Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping. SIAM Journal on Optimization, 30(3):2134– 2162, 2020.

TITAN with composite surrogate

Problem: $\min_x f(x) + g(x)$, where $f(x) = \psi(x) + \phi(r(x))$, and

TITAN with composite surrogate

Problem: $\min_x f(x) + g(x)$, where $f(x) = \psi(x) + \phi(r(x))$, and

- $\psi : \mathcal{X} \to \mathbb{R}$ is possibly nonsmooth nonconvex, which has block surrogate functions $u_i^{\psi}(x_i, y)$, $i = 1, \ldots, m$,

TITAN with composite surrogate

Problem: $\min_x f(x) + g(x)$, where $f(x) = \psi(x) + \phi(r(x))$, and

- $\psi : \mathcal{X} \to \mathbb{R}$ is possibly nonsmooth nonconvex, which has block surrogate functions $u_i^{\psi}(x_i, y)$, $i = 1, \ldots, m$,
- $r = (r_1, ..., r_m)$, where $r_i : \mathcal{X}_i \to \mathcal{Y}_i \subset \mathbb{F}_i$ are Lipschitz continuous (that is, $\|r_i(x_i) - r_i(y_i)\| \leq L_{r_i}\|x_i - y_i\|$ for $x_i, y_i \in \mathcal{X}_i$) and $\mathbb{F}_i$ $(i = 1, \ldots, m)$ are finite dimensional real linear spaces, and

TITAN with composite surrogate

Problem: $\min_x f(x) + g(x)$, where $f(x) = \psi(x) + \phi(r(x))$, and

- $\psi : \mathcal{X} \to \mathbb{R}$ is possibly nonsmooth nonconvex, which has block surrogate functions $u_i^{\psi}(x_i, y)$, $i = 1, \ldots, m$,
- $r = (r_1, ..., r_m)$, where $r_i : \mathcal{X}_i \to \mathcal{Y}_i \subset \mathbb{F}_i$ are Lipschitz continuous (that is, $\|r_i(x_i) - r_i(y_i)\| \leq L_{r_i}\|x_i - y_i\|$ for $x_i, y_i \in \mathcal{X}_i$) and $\mathbb{F}_i$ $(i = 1, \ldots, m)$ are finite dimensional real linear spaces, and
- $\phi : \mathcal{Y} := \mathcal{Y}_1 \times ... \times \mathcal{Y}_m \to \mathbb{R}_+$ is a continuously differentiable and block-wise concave function with Lipschitz gradient on the image of r.

TITAN with composite surrogate

Problem: $\min_x f(x) + g(x)$, where $f(x) = \psi(x) + \phi(r(x))$, and

- $\psi : \mathcal{X} \to \mathbb{R}$ is possibly nonsmooth nonconvex, which has block surrogate functions $u_i^{\psi}(x_i, y)$, $i = 1, \ldots, m$,
- $r = (r_1, ..., r_m)$, where $r_i : \mathcal{X}_i \to \mathcal{Y}_i \subset \mathbb{F}_i$ are Lipschitz continuous (that is, $\|r_i(x_i) - r_i(y_i)\| \leq L_{r_i}\|x_i - y_i\|$ for $x_i, y_i \in \mathcal{X}_i$) and $\mathbb{F}_i$ $(i = 1, \ldots, m)$ are finite dimensional real linear spaces, and
- $\phi : \mathcal{Y} := \mathcal{Y}_1 \times ... \times \mathcal{Y}_m \to \mathbb{R}_+$ is a continuously differentiable and block-wise concave function with Lipschitz gradient on the image of r.

Composite surrogate function for f:

$$u_i(x_i, y) = u_i^{\psi}(x_i, y) + \phi(r(y)) + \langle \nabla_i \phi(r(y)), r_i(x_i) - r_i(y_i) \rangle.$$

Problem: $\min_x f(x) + g(x)$, where $f(x) = \psi(x) + \phi(r(x))$.
Composite surrogate for f:

$$u_i(x_i, y) = u_i^{\psi}(x_i, y) + \phi(r(y)) + \langle \nabla_i \phi(r(y)), r_i(x_i) - r_i(y_i) \rangle.$$

Problem: $\min_x f(x) + g(x)$, where $f(x) = \psi(x) + \phi(r(x))$.
Composite surrogate for f:

$$u_i(x_i, y) = u_i^{\psi}(x_i, y) + \phi(r(y)) + \langle \nabla_i \phi(r(y)), r_i(x_i) - r_i(y_i) \rangle.$$

TITAN for the case ψ is block-wise Lipschitz smooth

Suppose $x_i \mapsto \psi(x_i, y_{\neq i})$ is $L_i^{(y)}$-Lipschitz smooth. We choose Lipschitz gradient surrogate for ψ and take $\mathcal{G}_i^k$ as

$$\mathcal{G}_i^k(x_i^k, x_i^{k-1}) = \nabla_i \psi(x^{k,i-1}) - \nabla_i \psi(\bar{x}_i^k, x_{\neq i}^{k,i-1}) + L_i^k \beta_i^k (x_i^k - x_i^{k-1}).$$

TITAN:

$$x_i^{k+1} = \operatorname*{argmin}_{x_i} \left\langle \nabla_i \psi(\bar{x}_i^k, x_{\neq i}^{k,i-1}), x_i \right\rangle + \frac{L_i^k}{2} \left\| x_i - \hat{x}_i^k \right\|^2 + \langle \nabla_i \phi(r(x^{k,i-1})), r_i(x_i) \rangle + g_i(x_i),$$

where $\bar{x}_i^k = x_i^k + \tau_i^k (x_i^k - x_i^{k-1})$ and $\hat{x}_i^k = x_i^k + \beta_i^k (x_i^k - x_i^{k-1})$.

Applying TITAN to solve MCP.

Recall - MCP

MCP solves $\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{r \times n}} \psi(U, V) + \mathcal{R}(U, V)$,

where $\psi(U, V) = \frac{1}{2}\|\mathcal{P}(A - UV)\|_F^2$, and

$$\mathcal{R}(U, V) = \lambda\Big(\sum_{ij}\big(1 - \exp(-\theta|u_{ij}|)\big) + \sum_{ij}\big(1 - \exp(-\theta|v_{ij}|)\big)\Big).$$

Applying TITAN to solve MCP.

Recall - MCP

MCP solves $\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{r\times n}} \psi(U,V)+\mathcal{R}(U,V)$,

where $\psi(U,V)=\frac{1}{2}\|\mathcal{P}(A-UV)\|_F^2$, and

$$\mathcal{R}(U,V)=\lambda\Big(\sum_{ij}\big(1-\exp(-\theta|u_{ij}|)\big)+\sum_{ij}\big(1-\exp(-\theta|v_{ij}|)\big)\Big).$$

- We observe that ψ is block-wise Lipschitz smooth and $\mathcal{R}=\phi\circ r$, where ϕ and r are given by

$$\begin{aligned}\phi(U,V) &= \lambda\Big(\textstyle\sum_{ij}\big(1-\exp(-\theta u_{ij})\big)+\sum_{ij}\big(1-\exp(-\theta v_{ij})\big)\Big),\\ r(U,V) &= (r_1(U),r_2(V))=(|U|,|V|),\end{aligned} \tag{3}$$

- Hence, we select the Lipschitz gradient surrogate for ψ and the composite surrogate function for $f=\psi+\phi\circ r$.

TITAN for MCP.

Update of U:

$$U^{k+1} = \operatorname*{argmin}_U \left\langle \nabla_U \psi(\bar{U}^k, V^k), U \right\rangle + \frac{L_1^k}{2}\|U - \bar{U}^k\|^2 + \langle \nabla_U \phi(r(U^k, V^k)), |U| \rangle,$$
$$= \mathcal{S}_{1/L_1^k}\left(P^k, \nabla_U \phi\big(r(U^k, V^k)\big)\right),$$

where

$$L_1^k = \|V^k (V^k)^\top\|, \bar{U}^k = U^k + \beta_1^k (U^k - U^{k-1}), P^k = \bar{U}^k - \frac{1}{L_1^k}\nabla_U \psi(\bar{U}^k, V^k)$$

and $\mathcal{S}_\tau$ is the soft-thresholding with parameter τ:

$$\mathcal{S}_\tau(P, W)_{ij} = [|p_{ij}| - \tau w_{ij}]_+ \text{sign}(p_{ij}).$$

Update of V:

$$V^{k+1} = \mathcal{S}_{1/L_2^k}\left(Q^k, \nabla_V \phi\big(r(U^{k+1}, V^k)\big)\right),$$

where

$$L_2^k = \|(U^{k+1})^T U^{k+1}\|, Q^k = \bar{V}^k - \frac{1}{L_2^k}\nabla_V \psi(U^{k+1}, \bar{V}^k), \bar{V}^k = V^k + \beta_2^k (V^k - V^{k-1}).$$

In our experiments, we choose

$$C = 0.9999^2, \mu_0 = 1, \mu_k = \tfrac{1}{2}(1 + \sqrt{1 + 4\mu_{k-1}^2}),$$
$$\beta_i^k = \min\left\{\tfrac{\mu_k - 1}{\mu_k}, \sqrt{C(1-\nu)L_i^{k-1}/L_i^k}\right\}.$$

Datasets and code

- The code is available from `https://github.com/nhatpd/TITAN`.

Table: The number of users, items, and ratings used in each data set.

Data set		#users	#items	#ratings
MovieLens	1M	6,040	3,449	999,714
	10M	69,878	10,677	10,000,054
Netflix		480,189	17,770	100,480,507

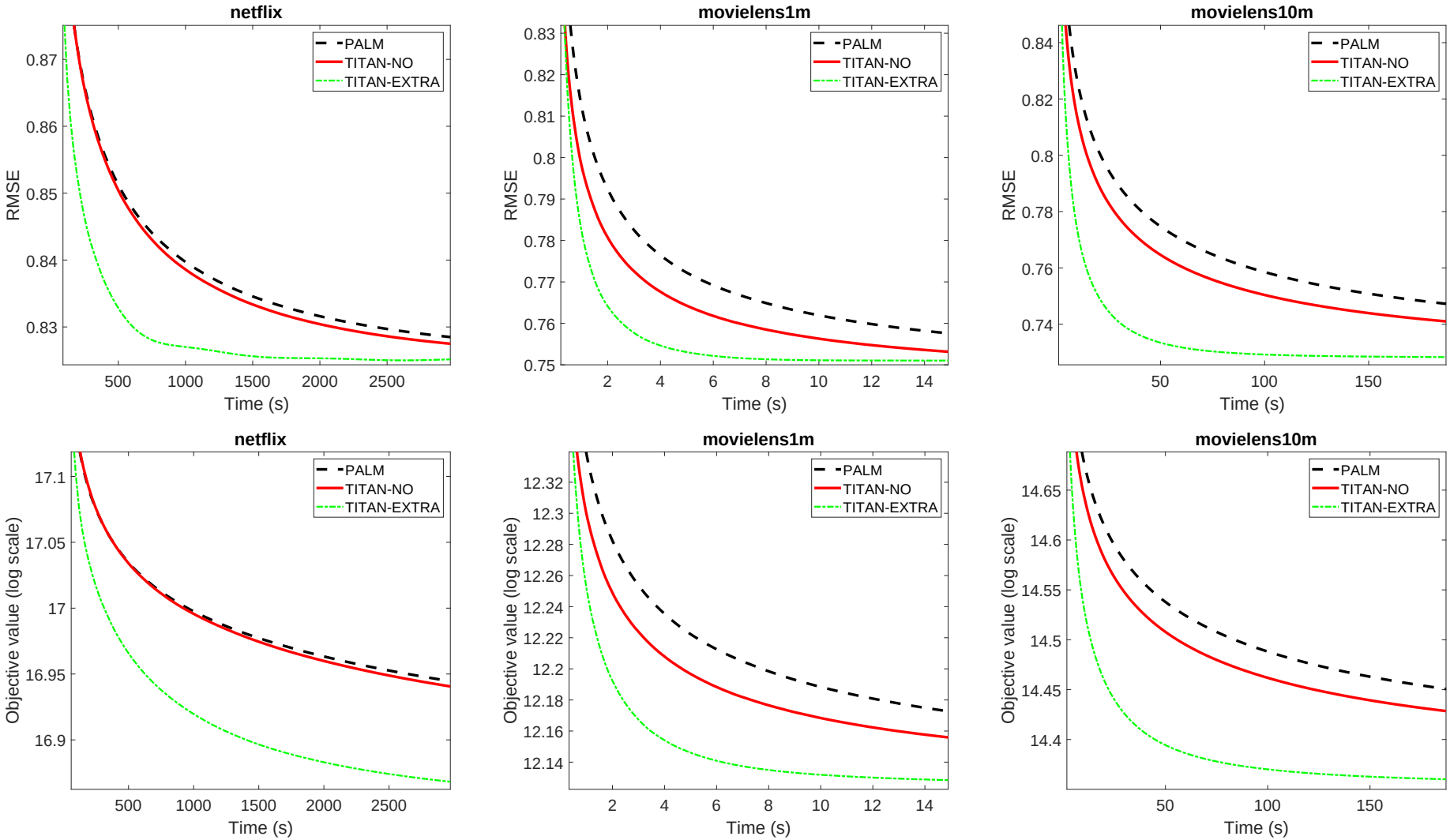

Figure: TITAN and PALM applied on the MCP. Evolution of the average value of the root mean squared error $RMSE = \sqrt{\|\mathcal{P}_T(A - UV)\|^2 / N_T}$, where $\mathcal{P}_T(Z)_{ij} = Z_{ij}$ if A_{ij} belongs to the test set and 0 otherwise, N_T is the number of ratings in the test set.

Summary

- TITAN for solving multiblock optmization problems:

$$x_i^{k,i} \in \underset{x_i \in \mathcal{X}_i}{\operatorname{argmin}}\, u_i(x_i, x^{k,i-1}) - \langle \mathcal{G}_i^k(x_i^k, x_i^{k-1}), x_i \rangle + g_i(x_i).$$

- TITAN unifies the convergence analysis of many known accelerated block coordinate descent methods.
- TITAN leads to new accelerated block coordinate descent methods.
- Extend TITAN for solving multiblock optimization problems with linear coupling constraint.
- Applications to NMF, MCP and a latent low-rank representation problem strongly confirm the acceleration effect of inertial technique.

- L. T. K. Hien, D. N. Phan, N. Gillis, "An Inertial Block Majorization Minimization Framework for Nonsmooth Nonconvex Optimization", arXiv:2010.12133
- L. T. K. Hien, D. N. Phan, N. Gillis, "Inertial Alternating Direction Method of Multipliers for Non-Convex Non-Smooth Optimization", Computational Optimization and Applications, 83, pp. 247-285, 2022.
- L. T. K. Hien, D. N. Phan, N. Gillis, M. Ahookhosh, P. Patrinos, "Block Bregman Majorization Minimization with Extrapolation", SIAM Journal on Mathematics of Data Science 4, 1-25, 2022.

THANK YOU FOR YOUR ATTENTION.