Block Coordinate Descent (BCD) Methods/ Proximal Alernating Linearized (PALM) Method

Acknowledegement: part of the lecture slides by Yangyang Xu and Wotao Yin

References

- Y. Xu and W. Yin. A block coordinate descent method for regularized multi-convex optimization with applications to nonnegative tensor factorization and completion. SIAM Journal on imaging sciences, 6(3), pp. 1758–1789, 2013.
- Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. `http://arxiv.org/abs/1410.1386`
- Jerome Bolte, Shoham Sabach, Marc Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, mathematical programming

Outline

Regularized multi-convex optimization

Model

$$\min_x F(x_1, x_2, \ldots, x_s) = f(x_1, x_2, \ldots, x_s) + \sum_{i=1}^{s} r_i(x_i)$$

where

1. f is differentiable and multi-convex, generally non-convex; e.g. $f(x_1, x_2) = x_1^2 x_2^2 + 2x_1^2 + x_2$;

2. each r_i is convex, possibly non-smooth; e.g. $r_i(x_i) = \|x_i\|_1$;

3. r_i is defined on $\mathbb{R} \cup \infty$; it can enforce $x_i \in \mathcal{X}_i$ by setting

$$r_i(x_i) = \delta_{\mathcal{X}_i}(x_i) = \begin{cases} 0, & \text{if } x_i \in \mathcal{X}_i, \\ \infty, & \text{otherwise.} \end{cases}$$

Applications

- Low-rank matrix recovery (Recht et. al, 2010)

$$\min_{X,Y} \|\mathcal{A}(XY) - \mathcal{A}(M)\|^2 + \alpha\|X\|_F^2 + \beta\|Y\|_F^2$$

- Sparse dictionary learning (Mairal et. al, 2009)

$$\min_{D,X} \frac{1}{2}\|DX - Y\|_F^2 + \lambda \sum_i \|x_i\|_1, \textit{subject to } \|d_j\|_2 \leq 1, \forall j;$$

- Blind source separation (Zibulevsky and Pearlmutter, 2001)

$$\min_{A,Y} \frac{1}{2}\|AYB - X\|_F^2 + \lambda\|Y\|_1, \textit{subject to } \|a^j\|_2 \leq 1, \forall j;$$

Applications

- Nonnegative matrix factorization (Lee and Seung, 1999)

$$\min_{X,Y} \|M - XY\|_F^2, \textit{subject to } X \geq 0; Y \geq 0;$$

- Nonnegative tensor factorization (Welling and Weber, 2001)

$$\min_{A_1,\ldots,A_N \geq 0} \|\mathcal{M} - A_1 \circ A_2 \circ \ldots \circ A_N\|_F^2;$$

Challenges

Non-convexity and non-smoothness cause

1. tricky convergence analysis;

2. expensive updates to all variables simultaneously.

Goal: to develop an efficient algorithm with simple update and global convergence (of course, to a stationary point)

Outline

Framework of block coordinate descent (BCD [1])

$$\min_x F(x_1, x_2, \ldots, x_s) = f(x_1, x_2, \ldots, x_s) + \sum_{i=1}^{s} r_i(x_i)$$

Algorithm 1: Block coordinate descent
Initialization:choose$(x_1^0, \ldots, x_s^0)$
for $k = 1, 2, \ldots,$ **do**
  for $i = 1, 2, \ldots, s$ **do**
    update x_i^k with all other blocks fixed
  end for
  if stopping criterion is satisfied **then**
  return $(x_1^k, \ldots, x_s^k)$
  end if
end for

Throughout iterations, each block x_i is updated by one of the three update schemes (coming next...)

[1] block coordinate update (BCU) is perhaps a more accurate name

Scheme 1: block minimization

The most-often used update:

$$x_i^k = \underset{x_i}{\operatorname{argmin}} F(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^{k-1}, \ldots, x_s^{k-1})$$

Existing results for differentiable convex F:

- Differentiable F and bounded level set $\Rightarrow$ objective converges to optimal value(Warga'63);
- Further with strict convexity $\Rightarrow$ sequence converges(Luo and Tseng'92);

Scheme 1: block minimization

Existing results for non-differentiable convex F:

- Non-differentiable F can cause stagnation at a non-critical point(Warga'63):
- Non-smooth part is separable $\Rightarrow$ subsequence convergence($i.e.$, exists a limit point) (Tseng'93)

Scheme 1: block minimization

Existing results for non-convex F:
May cycle or stagnate at a non-critical point (Powell'73):

$$F(x_1, x_2, x_3) = -x_1x_2 - x_2x_3 - x_3x_1 + \sum_{i=1}^{3}[(x_i - 1)_+^2 + (-x_i - 1)_+^2]$$

Each $F(x_i)$ has the form $(-a)x_i + [(x_i - 1)_+^2 + (-x_i - 1)_+^2]$

it minimizer $x_i^* = sign(a)(1 + 0.5\,|a|)$

Scheme 1: block minimization

Starting from $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$ with $\epsilon \geq 0$, minimizing F over $x_1, x_2, x_3, x_1, x_2, x_3, \ldots$ produces:

$$\xrightarrow{x_1} (1+\tfrac{1}{8}\epsilon, 1+\tfrac{1}{2}\epsilon, -1-\tfrac{1}{4}\epsilon) \quad \xrightarrow{x_2} (1+\tfrac{1}{8}\epsilon, -1-\tfrac{1}{16}\epsilon, -1-\tfrac{1}{4}\epsilon)$$
$$\xrightarrow{x_3} (1+\tfrac{1}{8}\epsilon, -1-\tfrac{1}{16}\epsilon, 1+\tfrac{1}{32}\epsilon) \quad \xrightarrow{x_1} (-1-\tfrac{1}{64}\epsilon, -1-\tfrac{1}{16}\epsilon, 1+\tfrac{1}{32}\epsilon)$$
$$\xrightarrow{x_2} (-1-\tfrac{1}{64}\epsilon, 1+\tfrac{1}{128}\epsilon, 1+\tfrac{1}{32}\epsilon) \quad \xrightarrow{x_3} (-1-\tfrac{1}{64}\epsilon, 1+\tfrac{1}{128}\epsilon, -1-\tfrac{1}{256}\epsilon)$$

Scheme 1: block minimization

Remedies for non-convex F:

- F is differentiable and strictly quasiconvex over each block $\Rightarrow$ limit point is a critical point (Grippo and Sciandrone'00);

  quasiconvex: $F(\lambda x + (1-\lambda)y) \leq max(F(x), F(y)), \forall \lambda \in [0,1]$

- F is pseudoconvex over every two blocks and non-differentiable part is separable $\Rightarrow$ limit point is a critical point (Tseng'01);

  pseudoconvex: $\langle g, y-x \rangle \geq 0$,some $g \in \partial F(x) \Rightarrow F(x) \leq F(y)$

There is not global convergence result.

Scheme 2: block proximal descent

Adding $\|x_i - x_i^{k-1}\|_2^2$ gives better stability:

$$x_i^k = \underset{x_i}{\operatorname{argmin}}\, F(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^{k-1}, \ldots, x_s^{k-1}) + \frac{L_i^{k-1}}{2}\|x_i - x_i^{k-1}\|^2;$$

Convergence results require fewer assumptions on F:

- F is convex $\Rightarrow$ objective converges to optimal value (Auslender'92);
- F is non-convex $\Rightarrow$ limit point is stationary (Grippo and Sciandrone'00);

Non-smooth terms must still be separable. No global convergence for non-convex F.

Scheme 3: block proximal linear

Linearize f over block i and add $\frac{L_i^{k-1}}{2}\|x_i - \hat{x}_i^{k-1}\|^2$:

$$x_i^k = \operatorname*{argmin}_{x_i} \left\langle \nabla_i f, x_i - \hat{x}_i^{k-1} \right\rangle + r_i(x_i) + \frac{L_i^{k-1}}{2}\|x_i - \hat{x}_i^{k-1}\|^2;$$

where $f = f(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^{k-1}, \ldots, x_s^{k-1})$

Scheme 3: block proximal linear

- Extrapolate $\hat{x}_i^{k-1} = x_i^{k-1} + w_i^{k-1}(x_i^{k-1} - x_i^{k-2})$ with weight $w_i^{k-1} \geq 0$
- Much easier than schemes 1 & 2; may have closed-form solutions for simple r_i;
- Used in randomized BCD for differentiable convex problems (Nesterov'12);
- The update is less greedy than schemes 1 & 2, causes more iterations, but may save total time;
- Empirically, the "relaxation" tend to avoid "shallow-puddle" local minima better than schemes 1 & 2 .

Comparisons

1. Block coordinate minimization (scheme 1) is mostly used
   - May generally cycle or stagnate at a non-critical point (Powell'73);
   - Globally convergent for strictly convex problem (Luo and Tseng'92);
   - For non-convex problem, each limit point is a critical point if each subproblem has unique solution and objective is regular (Tseng'01);
   - Global convergence for non-convex problems is unknown;

Comparisons

2. Block proximal (scheme 2) can stabilize iterates
   - Each limit point is a critical point (Grippo and Sciandrone'00);
   - Global convergence for non-convex problems is unknown;

3. Block proximal linearization (scheme 3) is often easiest
   - Very few works use this scheme for non-convex problems yet;
   - Related to the coordinate gradient descent method (Tseng and Yun'09).

Why different update schemes?

- They deal with subproblems of different properties;
- Implementations are easier for many applications;
- Schemes 2 & 3 may save total time than scheme 1;
- Convergence can be analyzed in a unified way.

Example: sparse dictionary learning

$$\min_{D,X} \frac{1}{2}\|DX - Y\|_F^2 + \lambda \sum_i \|x_i\|_1, \textit{subject to } \|D\|_F \leq 1$$

apply scheme 1 to D and scheme 3 to X; both are closed-form.

Outline

Examples of global convergence by BCD

- Low-rank matrix recovery (Recht et. al, 2010)

$$\min_{X,Y} \|\mathcal{A}(XY) - \mathcal{A}(M)\|^2 + \alpha\|X\|_F^2 + \beta\|Y\|_F^2$$

- Sparse dictionary learning (Mairal et. al, 2009)

$$\min_{D,X} \frac{1}{2}\|DX - Y\|_F^2 + \|X\|_1 + \delta_{\mathcal{D}}(D), \mathcal{D} = \{D : \|d_j\|_2 \leq 1, \forall j\}$$

- Blind source separation (Zibulevsky and Pearlmutter, 2001)

$$\min_{A,Y} \frac{\lambda}{2}\|AYB - X\|_F^2 + \|Y\|_1 + \delta_{\mathcal{A}}(A), \mathcal{A} = \{A : \|a^j\|_2^2 \leq 1, \forall j\}$$

- Nonnegative matrix factorization (Lee and Seung, 1999)

$$\min_{X,Y} \|M - XY\|_F^2, \textit{subject to } X \geq 0; Y \geq 0;$$

Numerical results

Part I: nonnegative matrix factorization (NMF)
Model:

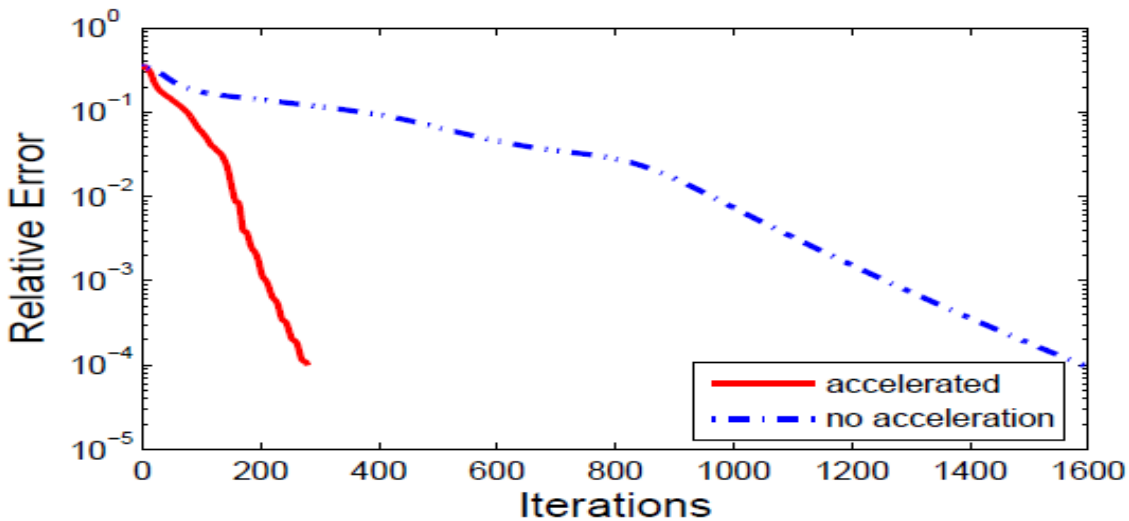
$$\min_{X,Y} \frac{1}{2}\|M - XY\|_F^2, \quad subject\ to\ X \in \mathbb{R}_+^{m\times r}, Y \in \mathbb{R}_+^{r\times n}$$

Algorithms compared:

1. APG-MF(proposed): BCD with scheme 3,$\omega_i^k = min(\hat{\omega}_k, \sqrt{\frac{L_i^{k-1}}{L_i^k}})$, i=1,2, where $\hat{\omega}_k = \frac{t_{k-1}-1}{t_k}$ and $t_0 = 1, t_k = \frac{1}{2}\sqrt{1+4t_{k-1}^2}$; $\hat{\omega}_k$ used in FISTA (Beck and Teboulle'09);
2. ADM-MF: alternating direction method for NMF (Y. Zhang'10);
3. Blockpivot-MF: BCD with block minimization (scheme 1); subproblems solved by block principle pivoting method (Kim and Park'08);
4. Als-MF and Mult-MF: Matlab's implementation.

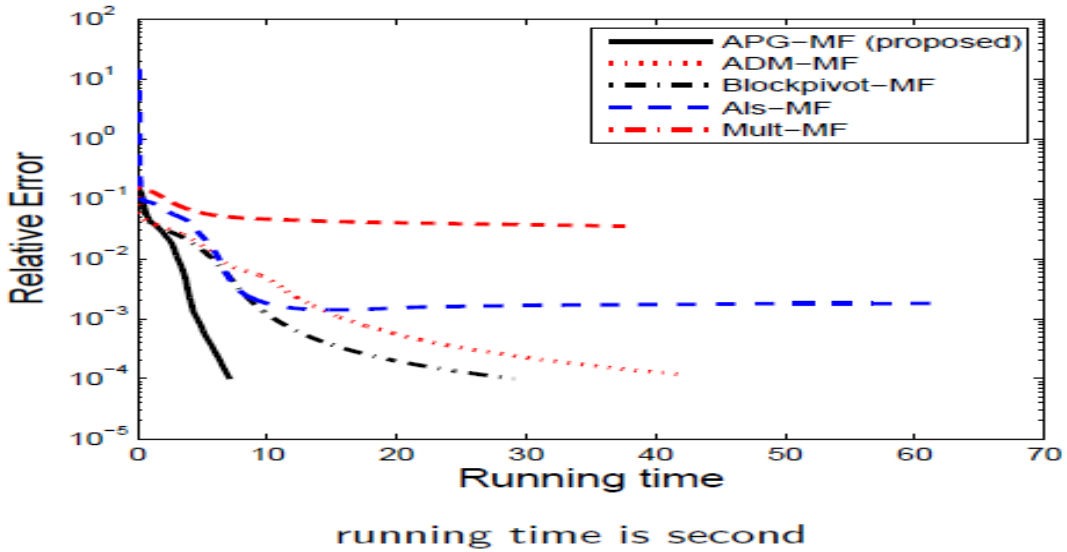
Extrapolation accelerates convergence

- Extrapolation acceleration: $\omega_i^k = min(\hat{\omega}_k, \sqrt{\frac{L_i^{k-1}}{L_i^k}})$, i=1,2, where $\hat{\omega}_k = \frac{t_{k-1}-1}{t_k}$ and $t_0 = 1, t_k = \frac{1}{2}\sqrt{1+4t_{k-1}^2}$
- No acceleration: $\omega_i^k = 0, i = 1, 2;$



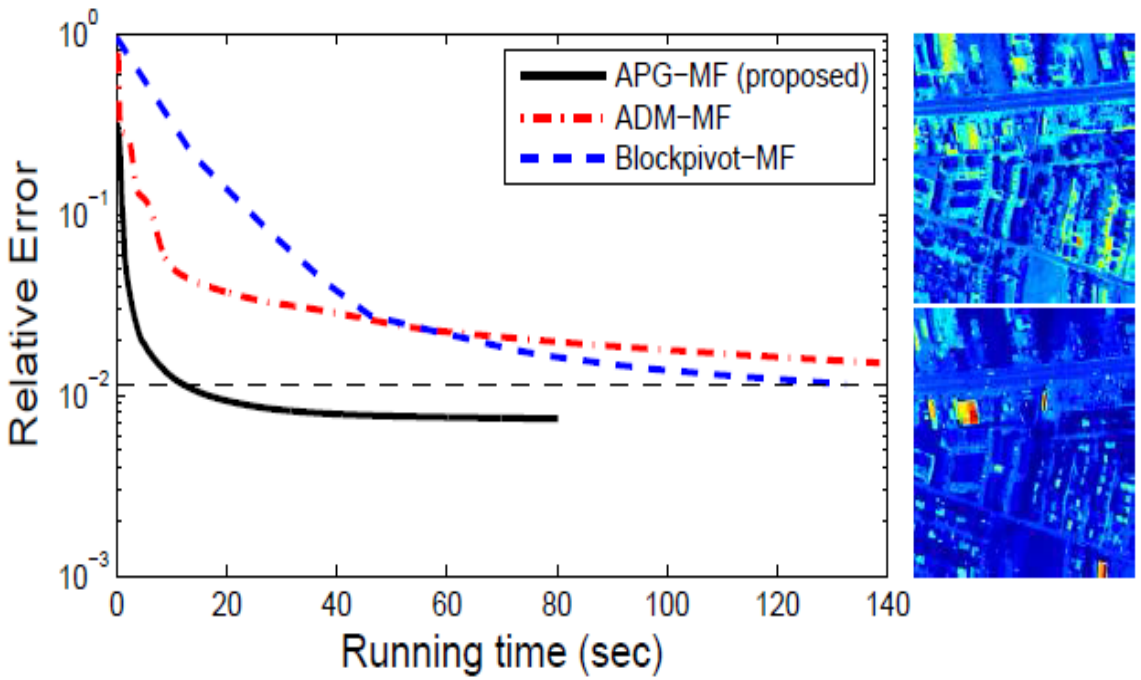
Comparison on synthetic data

- Random $M = LR$ and $L \in \mathbb{R}_+^{500\times 30}, R \in \mathbb{R}_+^{30\times 1000}$;
- relerr=$\frac{\|M-XY\|_F}{\|M\|_F}$ and running time(sec)



Comparison on hyperspectral data

- $163 \times 150 \times 150$ hyperspectral cube is reshaped to 22500×163 matrix M



Part II: Nonnegative 3-way tensor factorization

Model:

$$\min_{A_1,\ldots,A_3} \frac{1}{2}\|\mathcal{M} - A_1 \circ A_2 \circ A_3\|_F^2, \ \textit{subject to } A_n \in \mathbb{R}_+^{I_n \times r}, \forall n.$$

Compared algorithms:

1. APG-TF (proposed) : BCD with scheme 3, $\omega_i^k = min(\hat{\omega}_k, \sqrt{\frac{L_i^{k-1}}{L_i^k}})$, i=1,2,3, where $\hat{\omega}_k = \frac{t_{k-1}-1}{t_k}$ and $t_0 = 1, t_k = \frac{1}{2}\sqrt{1+4t_{k-1}^2}$
2. AS-TF: BCD with scheme 1 subproblems solved by active set method (Kim et.al, '08);
3. Blockpivot-TF: BCD with scheme 1; subproblems solved by block principle pivoting method (Kim and Park '12);

Swimmer dataset

Shashua and Hazan'05: NMF tends to form invariant parts as ghosts while NTF can correctly resolve all parts

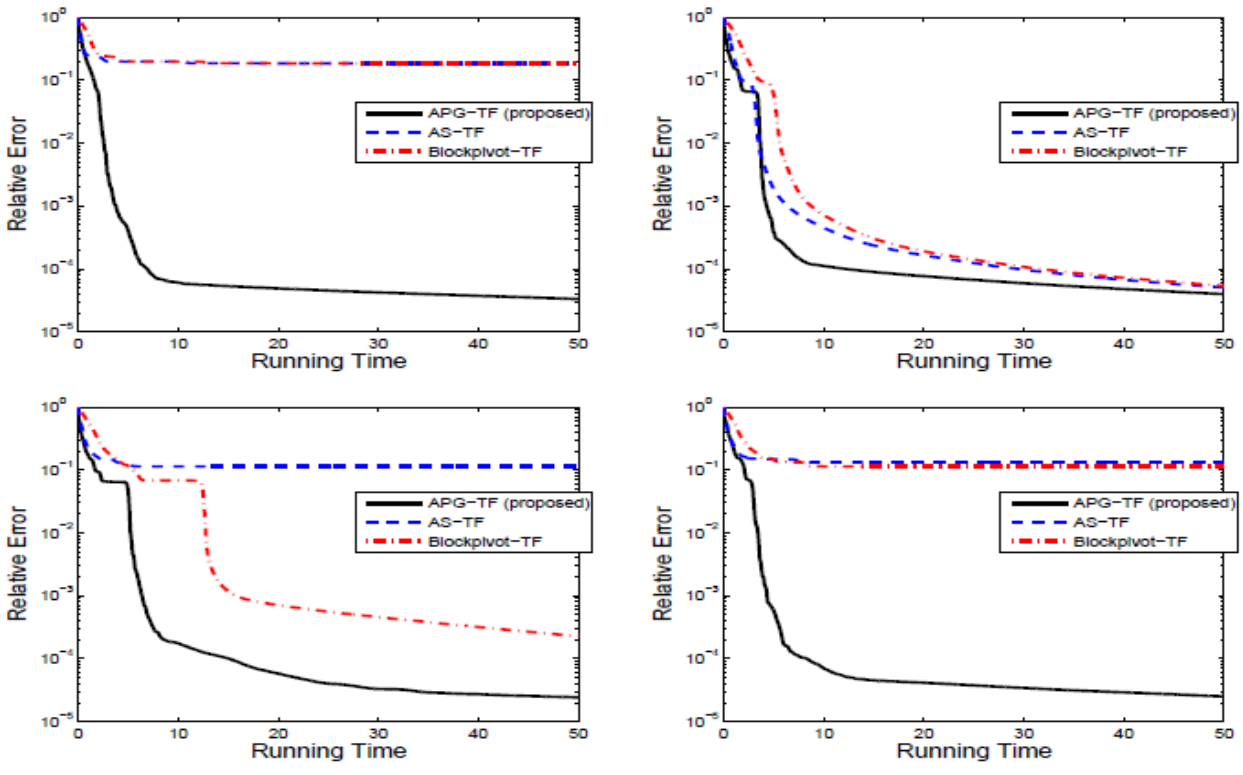
8 among 256 images

factors by NMF

factors by NTF

Comparison on the Swimmer dataset

$32 \times 32 \times 256$ nonnegative tensor $\mathcal{M}$; run to 50 seconds; r set to 60;



Outline

Framework of Bolte, Sabach, Teboulle

Consider the model:

$$(M) \qquad \min_{x,y} \quad \Psi(x,y) = f(x) + g(y) + H(x,y)$$

Let

$$\text{prox}_t^{\sigma}(x) = \arg\min_u \ \sigma(u) + \frac{t}{2}\|u - x\|^2$$

A single iteration of PALM:

- Take $\gamma_1 > 1$, set $c_k = \gamma_1 L_1(y^k)$ and compute

$$x^{k+1} = \text{prox}^f_{c_k}(x^k - \frac{1}{c_k}\nabla_x H(x^k, y^k))$$

- Take $\gamma_2 > 1$, set $d_k = \gamma_2 L_2(x^{k+1}$ and compute

$$y^{k+1} = \text{prox}^g_{d_k}(y^k - \frac{1}{d_k}\nabla_y H(x^{k+1}, y^k))$$

Nonconvex Proximal Operator

Let

$$m^{\sigma}(x,t) = \inf_{u} \ \sigma(u) + \frac{1}{2t}\|u-x\|^2$$

Well-definedness of proximal maps:

- Let $\sigma(u)$ be a proper and lower semicontinuous function with $\inf \ \sigma > -\infty$. Then, for every $t \in (0,\infty)$, the set $\operatorname{prox}^{\sigma}_{1/t}(x)$ is nonempty and compact. In addition, $m^{\sigma}(x,t)$ is finite and continuous in (x,t).
- When $\sigma = \delta_X$, the indicator function of a nonempty and closed set X, the proximal map reduces to the projection operator.

Subdifferentials of nonconvex and nonsmooth fun

$\sigma(u) : \mathbb{R}^d \to (-\infty, +\infty)$ be a proper and lower semicontinuous fun.

- For a given $x \in \mathbf{dom}\ \sigma$, the Fréchet subdifferential of σ at x, written $\hat{\partial}\sigma(x)$, is the set of all vectors $u \in \mathbb{R}^d$ which satisfy

$$\liminf_{y \neq x,\ y \to x} \frac{\sigma(y) - \sigma(x) - \langle u, y - x \rangle}{\|y - x\|} \geq 0$$

- The limiting-subdifferential, or simply the subdifferential, of σ at $x \in \mathbb{R}^n$, written $\partial\sigma(x)$, is defined through the following closure process

$$\partial\sigma(x) = \{u \in \mathbb{R}^d : \exists x^k \to x, \sigma(x^k) \to \sigma(x), \text{ and } u^k \in \hat{\partial}\sigma(x^k) \to u, k \to \infty\}$$

- Assume that the coupling function H in Problem (M) is continuously differentiable. Then for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$:

$$\partial\Psi(x, y) = (\nabla_x H(x, y) + \partial f(x), \nabla_y H(x, y) + \partial g(y))$$

Assumptions

- $\inf \Psi > -\infty$, $\inf f > -\infty$ and $\inf g > -\infty$
- For any fixed y, the partial gradient $\nabla_x H(x,y)$ is Lipschitz with moduli $L_1(y)$. For any fixed x, $\nabla_y H(x,y)$ is Lipschitz with moduli $L_2(x)$
- There exists $\lambda_i^-, \lambda_i^+ > 0$ such that

$$\inf\{L_1(y^k) : k \in N\} \geq \lambda_1^- \text{ and } \inf\{L_2(x^k) : k \in N\} \geq \lambda_2^-$$
$$\sup\{L_1(y^k) : k \in N\} \leq \lambda_1^+ \text{ and } \sup\{L_2(x^k) : k \in N\} \leq \lambda_2^+$$

- ∇H is Lipschitz continuous on bounded subsets of $\mathbb{R}^n \times \mathbb{R}^m$:

$$\|(\nabla_x H(x_1,y_1) - \nabla_x H(x_2,y_2), \nabla_y H(x_1,y_1) - \nabla_y H(x_2,y_2))\| \leq M\|(x_1 - x_2, y_1 - y_2)\|$$

Informal Proofs

- Sufficient decrease property: Find a positive constant ρ_1 such that
$$\rho_1 \|z^{k+1} - z^k\|^2 \leq \Psi(z^k) - \Psi(z^{k+1})$$
- A subgradient lower bound for the iterates gap: Assume that the sequence generated by the algorithm is bounded. Find another positive constant ρ_2, such that
$$\|w^{k+1}\| \leq \rho_2 \|z^{k+1} - z^k\|, \quad w^k \in \partial\Psi(z^k)$$
- Using the KL property: Assume that Ψ is a KL function and show that the generated sequence $\{z^k\}_{k\in N}$ is a Cauchy sequence.

Note that when the first two properties hold, then for any algorithm one can show that the set of accumulations points is a nonempty, compact and connected set.

The Kurdyka-Łojasiewicz (KL) property

- For any subset $S \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, $\mathrm{dist}(x, S) = \inf_y \ \|y - x\|$.
- Let $\eta \in (0, +\infty)$. Φ_η is the class of all concave and continuous functions $\psi : [0, \eta) \to \mathbb{R}_+$: (i) $\psi(0) = 0$, (ii) ψ is C^1 on $(0, \eta)$ and continuous at 0; (iii) for all $s \in (0, \eta)$, $\psi'(s) > 0$.
- Let σ be proper and lower semicontinuous.
- σ has KL property at $\bar{u} \in \mathbf{dom}\ \partial\sigma := \{u \in \mathbb{R}^d \mid \partial\sigma(u) \neq \emptyset\}$: if there exits $\eta \in (0, +\infty)$, a neighborhood U of $\bar{u}$ and a function $\psi \in \Phi_\eta$ such that for all $u \in U \cap [\sigma(\bar{u}) < \sigma(u) < \sigma(\bar{u}) + \eta]$, it holds:

$$\psi'(\sigma(u) - \sigma(\bar{u}))\mathrm{dist}(0, \partial\sigma(u)) \geq 1$$

- If σ satisfy the KL property at each point of $\mathbf{dom}\ \partial\sigma$, then σ is called a KL function

The Kurdyka-Łojasiewicz (KL) sets and functions

semi-algebraic, subanalytic and log-exp are KL functions

- A subset S of $\mathbb{R}^d$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^d \to \mathbb{R}$ such that

$$S = \cup_{j=1}^{p} \cap_{i=1}^{q} \{u \in \mathbb{R}^d : g_{ij}(u) = 0, h_{ij}(u) < 0\}$$

- A function $h : \mathbb{R}^d \to (-\infty, +\infty]$ is called semi-algebraic if its graph $\{(u,t) \in \mathbb{R}^{d+1} : h(u) = t\}$ is a semi-algebraic subset of $\mathbb{R}^{d+1}$
- Let $\sigma(u) : \mathbb{R}^d \to (-\infty, +\infty)$ be a proper and lower semicontinuous function. If σ is semi-algebraic then it satisfies the KL property at any point of $\mathbf{dom}\sigma$.

The Kurdyka-Łojasiewicz (KL) sets and functions

Examples:

- Real polynomial functions.
- Indicator functions of semi-algebraic sets.
- Finite sums and product of semi-algebraic functions.
- Composition of semi-algebraic functions.
- Sup/Inf type function, e.g., $\sup\{g(u,v) : v \in C\}$ is semi-algebraic when g is a semi-algebraic function and C a semi-algebraic set.
- In matrix theory, all the following are semi-algebraic sets: cone of PSD matrices, Stiefel manifolds and constant rank matrices.
- The function $x \to \operatorname{dist}(x,S)^2$ is semi-algebraic whenever S is a nonempty semialgebraic subset of $\mathbb{R}^d$.
- $\|\cdot\|_0$, $\|\cdot\|_p$ with a rational p are semi-algebraic

Proofs: Sufficient decrease property

- Let $h : \mathbb{R}^d \to R$ be a continuously differentiable function with gradient ∇h assumed L_h-Lipschitz continuous and let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function with $\inf_{\mathbb{R}^d} \sigma > -\infty$. Fix any $t > L_h$. Then, for any $u \in \mathbf{dom}\sigma$ and any

$$u^+ \in \mathrm{prox}_t^\sigma (u - \frac{1}{t}\nabla h(u)),$$

we have

$$h(u^+) + \sigma(u^+) \leq h(u) + \sigma(u) - \frac{1}{2}(t - L_h)\|u - u^+\|^2$$

Proofs: convergence property

- The sequence $\{\Psi(z^k)\}_{k\in N}$ is nonincreasing and with $\rho_1 = \min\{(\gamma_1 - 1)\lambda_1^-, (\gamma_2 - 1)\lambda_2^-\}$:

$$\frac{\rho_1}{2}\|z^{k+1} - z^k\|^2 \leq \Psi(z^k) - \Psi(z^{k+1}), \forall k \geq 0$$

- We have

$$\sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2 + \|y^{k+1} - y^k\|^2 = \sum_{k=1}^{\infty} \|z^{k+1} - z^k\|^2 < \infty$$

and hence $\lim_{k\to\infty} \|z^{k+1} - z^k\| = 0$

Proofs: subgradient lower bound for the iterates gap

- Define $\rho_2 = \max\{\gamma_1\lambda_1^+, \gamma_2\lambda_2^+\}$ and

$$\begin{aligned} A_x^k &= c_{k-1}(x^{k-1} - x^k) + \nabla_x H(x^k, y^k) - \nabla_x H(x^{k-1}, y^{k-1}) \\ A_y^k &= d_{k-1}(y^{k-1} - y^k) + \nabla_y H(x^k, y^k) - \nabla_y H(x^k, y^{k-1}) \end{aligned}$$

Then $(A_x^k, A_y^k) \in \partial\Psi(x^k, y^k)$ and there exists $M > 0$:

$$\|(A_x^k, A_y^k)\| \leq \|A_x^k\| + \|A_y^k\| \leq (2M + 3\rho_2)\|z^k - z^{k-1}\|$$

Proofs: Uniformized KL property

- Let Ω be compact and σ be proper and lower semicontinuous. Assume σ is constant on Ω and satisfy the KL property at each point of Ω. Then, $\exists \epsilon > 0, \eta > 0$ and $\psi \in \Psi_\eta$ such that for $\bar{u} \in \Omega$,

$$u \in \{u \in \mathbb{R}^d : \mathrm{dist}(u, \Omega) < \epsilon\} \cap [\sigma(\bar{u}) < \sigma(u) < \sigma(\bar{u}) + \eta]$$

one has

$$\psi'(\sigma(u) - \sigma(\bar{u}))\mathrm{dist}(0, \partial\sigma(u)) \geq 1$$

Convergence of PALM to critical points

Suppose that Ψ is a KL function. Let $\{z^k\}_{k\in N}$ be a sequence generated by PALM which is assumed to be bounded.

- The sequence $\{z^k\}_{k\in N}$ has finite length:

$$\sum_{k=1}^{\infty} \|z^{k+1} - z^k\| < \infty$$

- The sequence $\{z^k\}_{k\in N}$ converges to a critical point $z^* = (x^*, y^*)$ of Ψ

Extension of PALM for problems with $p \geq 1$ blocks

Convergence of PALM to critical points

Choose $\psi(s) = cs^{1-\theta}$, where $c > 0$ and $\theta \in [0, 1)$.

- If $\theta = 0$, then the sequence converges in a finite number of steps
- If $\theta \in (0, 1/2]$, then there exists $\omega > 0$ and $\tau \in [0, 1)$ such that $\|z^k - \bar{z}\| \leq \omega\tau^k$
- If $\theta \in (1/2, 1)$, then there exits $\omega > 0$ such that

$$\|z^k - \bar{z}\| \leq wk^{-\frac{1-\theta}{2\theta-1}}$$

Convergence results of Wotao and Yangyang

Assumptions:

$$\min_x F(x_1, x_2, \ldots, x_s) = f(x_1, x_2, \ldots, x_s) + \sum_{i=1}^{s} r_i(x_i)$$

Assumption 1. Continuous, lower-bounded, and $\exists$ a stationary point.
Assumption 2. Each block uses only one update scheme throughout, and

1. block using scheme 1: subproblem is strongly convex with modulus L_i^k;
2. block using scheme 3: subproblem has Lipschitz continuous gradient.

Assumption 3. $\exists 0 < l \leq L < \infty$ such that $l \leq L_i^k \leq L, \forall i, k$.

Assumptions 1-3 are assumed for all results below.

Convergence results

Lemma 2.2 Let $\{x^k\}$ be the sequence generated by BCD. If block i is updated by scheme 3, the extrapolation weight is controlled as $0 \leq \omega_i^k \leq \delta_\omega \sqrt{\frac{L_i^{k-1}}{L_i^k}}$ with $\delta_\omega < 1$ for all k.Then,

$$\sum_{i=1}^{\infty} \|x_k - x_{k+1}\|^2 < \infty$$

Theorem 2.1 (Limit point is stationary point) Under the assumptions of Lemma 2.2, any limit point of $\{x^k\}$ is a stationary point.

As a trivial extension:

Theorem 2.2 (Isolated stationary points) If $\{x^k\}$ is bounded and the stationary points are isolated, then x_k converges to a stationary point.

Global convergence and rate(using the Kurdyka-Lojasiewicz property)

Theorem 2.3 Let $\{x^k\}$ be the sequence generated by BCD. If block i is updated by scheme 3,assume $0 \leq \omega_i^k \leq \delta_\omega \sqrt{\frac{L_i^{k-1}}{L_i^k}}$ with $\delta_\omega < 1$ for all k. Assume $F(x_k) \leq F(x_{k-1})$. If $\{x^k\}$ has a finite limit point $\bar{x}$ and

$$\frac{|F(x)-F(\bar{x})|^\theta}{dist(0,\partial F(x))} \text{ is bounded around } \bar{x} \text{ for } \theta \in [0,1), \quad (1)$$

then

$$x^k \to \bar{x}$$

Theorem 2.4 (rate of convergence): In addition, in (1),

1. if θ= 0,x^k converges to $\bar{x}$ in finitely many iterations;
2. if $\theta \in (0, \frac{1}{2}], \|x_k - \bar{x}\| \leq C\tau^k, \forall k$, for certain $C > 0, \tau \in [0,1)$;
3. if $\theta \in (\frac{1}{2}, 1), \|x_k - \bar{x}\| \leq Ck^{\frac{-(1-\theta)}{2\theta-1}}, \forall k$, for certain $C > 0$.

The Kurdyka- Lojasiewicz (KL) property

Definition 2.9. (Lojasiewicz'93) $\psi(x)$ has the Kurdyka- Lojasiewicz (KL) property if there exists $\theta \in [0, 1)$ such that

$$\frac{|\psi(x)-\psi(\bar{x})|^{\theta}}{dist(0,\partial\psi(x))} \qquad (2)$$

is bounded around $\bar{x}$

History:

- Introduced by (Lojasiewicz'93) on real analytic functions, for which the term with $\theta \in [\frac{1}{2}, 1)$ in (2) is bounded around any critical point $\bar{x}$.
- (Kurdyka'98) extended the properties to functions on the o-minimal structure.
- (Bolte et. al '07) extended the property to nonsmooth sub-analytic functions.