Digest of the convergence analysis of APG - by Xu-Yin

Problem (1.1)

$$\min_{x \in \mathcal{X}} F(x_1, \cdots, x_S) \triangleq f(x_1, \cdots, x_S) + \sum_{i=1}^{S} r_i(x_i) \qquad (1.1)$$

- x_i block variable, x full variable, S-blocks
- $\mathcal{X}$ constraint on x, assumed:
  - close
  - block convex ... (def1)
- f assumed block convex, differentiable, smooth
- r_i extended valued ($r_i(x_i) \overset{\text{def}}{=} \infty$ if $x_i \notin \text{dom} x_i$)
  - r_i can be non-smooth

Note: r_i, f can be non-convex

Interpretation of (1.1)
- f models the joint function, r_i model individual block
- r_i models individual constraint
- $\mathcal{X}$ models joint constraint
- if no constraint on block i, $r_i \equiv 0$

Definitions

(def1) Set multi-convex (mc): the set $\mathcal{X}$ is mc if its projection to each sub-block is cvx.

i.e. $\forall i$, fixed S-1 block $\{x_j\}_{j \neq i}$, the set

$$\mathcal{X}_i(\{x_j\}_{j \neq i}) \triangleq \left\{ x_i \in \mathbb{R}^{n_i} \,\middle|\, (\underbrace{x_1, \cdots, x_{i-1}}_{\text{fix}}, \overset{\text{subject}}{\downarrow} x_i, \underbrace{x_{i+1}, \cdots, x_S}_{\text{fix}}) \in \mathcal{X} \right\}$$

is cvx.

(def2) Block multi-convex function : $\forall i$, f is convex of x_i while other block held fix.

i.e. for C.D, the function $F(x_1, \dots, x_s)$ is CVX w.r.t. one block while other block are fix.

BCD update of Gauss-Seidel Type / Cyclic update

"minimizing F cyclically over $x_1, x_2, \dots, x_s$ while fixing other blocks at their last updated values"

Notation x_i^k : x_i at k^{th} iteration

$$f_i^k(x_i) \triangleq f(\underbrace{x_1^k, \dots, x_{i-1}^k}_{\text{fix, at last updated value}}, \overset{\text{subject}}{x_i}, \underbrace{x_{i+1}^{k-1}, \dots x_s^{k-1}}_{\text{fix, at last updated value}})$$

note : k and k-1 !!

$g_i^k = \nabla f_i^k(x_i^{k-1})$ block partial gradient of f at x_i^{k-1}

note : g^{k} and x_i^{k} !!

$$\hat{x}_i^{k-1} = x_i^{k-1} + \omega_i^{k-1}(x_i^{k-1} - x_i^{k-2})$$

extrapolation point, $\omega_i^{k-1} \geq 0$ $\forall i,k$, the weight

Update forms

(a) original update $x_i^k = \underset{x_i \in \mathcal{X}_i^k}{\operatorname{argmin}} \; f_i^k(x_i) + r_i(x_i)$

"update x_i based on $f_i^k(x_i)$ and r_i"

(b) proximal update $x_i^k = \underset{x_i \in \mathcal{X}_i^k}{\operatorname{argmin}} \; f_i^k(x_i) + \frac{L_i^{k-1}}{2}\|x_i - x_i^{k-1}\|^2 + r_i(x_i)$

"update x_i based on original form plus a proximal term, where block-Lipschitz constant L_i^{k-1} is known"

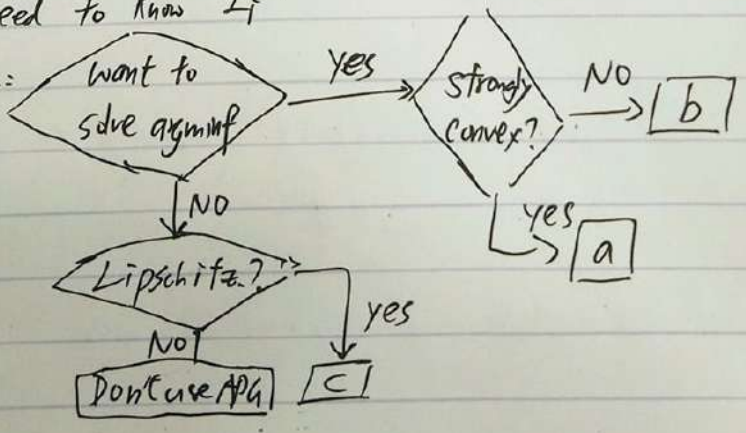
(c) prox-linear $x_i^k = \underset{x_i \in \mathcal{X}_i^k}{\operatorname{argmin}} \; \langle \hat{g}_i^k(x_i^{k-1}), x_i - x_i^{k-1} \rangle + \frac{L_i}{2}\|x_i - x_i^{k-1}\|^2 + r_i(x_i)$

"replace f in (b) by linear approximation"

Remarks

- if $f, \mathcal{X}$ cvx, then all sub problems of (a), (b), (c) are cvx $\Rightarrow$ global minima for all sub-problem
- in general harder to solve (a), (b) than (c) due to the "argmin f"
- (b), (c) need to know L_i^{k-1}

choosing algo.:



Assumptions

(a-1) F is continuous in $\operatorname{dom} F$

(a-2) $\inf_{X \in \operatorname{dom} F} F(x) > -\infty$

(a-3) (1.1) has a Nash-point (useless for matrix factorization problem)

(a-4) Consistent update scheme (a,b,c) for each block

(a-5) For (a), f_i^k is str-cvx : $\ell_i \le L_i^{k-1} \le L_i$

(a-6) For (b), L_i^{k-1} obey $\ell_i \le L_i^{k-1} \le L_i$

(a-7) For (c), ∇f_i^k is Lipschitz

L_i^{k-1} obey $\ell_i \le L_i^{k-1} \le L_i$

$$f_i^k(x_i^k) \le f_i^k(\vec{x}_i^{k-1}) + \langle \vec{g}_i^k, x^k - \vec{x}_i^{k-1} \rangle + \frac{L_i^{k-1}}{2} \|x_i - x_i^{k-1}\|_2^2$$

Conceptual flow of the convergence proof

Step I 1. Show sufficient decrease condition

2. Show square summable

$$\sum_k \|x^k - x^{k-1}\|_2^2 < +\infty$$

Step 2 1. Use Kurdyka-Łojasiewicz inequality to improve $\sum \|\cdot\|^2 \to \sum \|\cdot\|^1$

2. Global convergence + asymptotic rate

Lemma 2.1. Let $a(u)$, $b(u)$ be cvx function on cvx set $\mathcal{U}$.
a is differentiable, b is possibly non-smooth.

Let $A(u) = a(u) + b(u)$. ..... (A)

Let $u^* = \underset{u \in \mathcal{U}}{\operatorname{argmin}} \langle a(v), u - v \rangle + \frac{1}{2}\|u - v\|^2 + b(u)$. ..... (Δ)

If $a(u^*) \le a(v) + \langle \nabla a(v), u^* - v \rangle + \frac{1}{2}\|u^* - v\|_2^2$ ... (#)

Then

$$A(u) - A(u^*) \ge \frac{1}{2}\|u^* - v\|^2 + L\langle v - u^*, u^* - v \rangle \quad \forall u$$

Proof. From (Δ), by first-order optimality / Fermat's rule on u^*,

$$\langle \nabla a(v) + L(u^* - v) + g, u - u^* \rangle \ge 0 \quad , \quad g \in \partial b(u^*) \quad \cdots (*)$$

(why: recall 1. b is possibly non-smooth. so use ∂b
2. $u^* = \operatorname{argmin} \phi \Leftrightarrow \langle \nabla \phi(u^*), u - u^* \rangle \ge 0 \ \forall u$)

Consider $A(u) - A(u^*)$

$$A(u) - A(u^*) \overset{(A)}{=} a(u) - a(u^*) + b(u) - b(u^*)$$

$$\overset{(\#)}{\ge} a(u) - \left(a(v) + \langle \nabla a(v), u^* - v \rangle + \frac{1}{2}\|u^* - v\|_2^2 \right) + b(u) - b(u^*)$$

$$= a(u) - a(v) - \left(\langle \nabla a(v), u^* - v \rangle + \frac{1}{2}\|u^* - v\|_2^2 \right) + b(u) - b(u^*)$$

Trick $-\langle \nabla a(v), u^* - v \rangle = -\langle \nabla a(v), u^* - v + u - u \rangle$

$= -\langle \nabla a(v), u^* - u \rangle - \langle \nabla a(v), u - v \rangle$

$= +\langle \nabla a(v), u - u^* \rangle - \langle \nabla a(v), u - v \rangle$

$$\overset{\text{Trick}}{=} a(u) - a(v) + \langle a(v), u - u^* \rangle - \langle \nabla a(v), u - v \rangle - \frac{1}{2}\|u^* - v\|_2^2 + b(u) - b(u^*)$$

$$\overset{a \text{ is cvx}}{\ge} \langle a(v), u - u^* \rangle - \frac{1}{2}\|u^* - v\|_2^2 + b(u) - b(u^*)$$

Now we have

$$A(u) - A(u^*) \geq \langle \nabla a(v), u-u^* \rangle - \tfrac{L}{2}\|u^*-v\|_2^2 + b(u) - b(u^*)$$

To handle $\langle \nabla a(v), u-u^* \rangle$, use $(*)$

$$(*) \Rightarrow \langle \nabla a(v), u-u^* \rangle \geq -L\langle u^*-v, u-u^* \rangle - \langle g, u-u^* \rangle$$

So $A(u) - A(u^*) \overset{(*)}{\geq} -L\langle u^*-v, u-u^* \rangle - \langle g, u-u^* \rangle$
$- \tfrac{L}{2}\|u^*-v\|_2^2 + b(u) - b(u^*)$

(**) b cvx $\Leftrightarrow$ $b(u) - b(u^*) \geq \langle g, u-u^* \rangle$

$$\geq -L\langle u^*-v, u-u^* \rangle - \tfrac{L}{2}\|u^*-v\|_2^2$$

$$= L\langle u^*-v, u^*-u + v-v \rangle - \tfrac{L}{2}\|u^*-v\|_2^2$$

$$= \tfrac{L}{2}\|u^*-v\|_2^2 + L\langle u^*-v, v-u \rangle \quad //$$

Lemma 2.2

1. Assumption 1.2
2. $\{x^k\}$ generate by algo-1
3. With $0 \leq \omega_i^{k-1} \leq \delta_w \sqrt{\frac{L_i^{k-2}}{L_i^{k-1}}}$ $\forall i \in I_3$, $\delta_w < 1$

Then $\sum_{k=0}^{\infty} \|x^k - x^{k+1}\|_2^2 < \infty$

Proof. For I_3, let $a = f_i^k$, $b = r_i$, $v = \hat{x}_i^{k-1}$, $u = x_i^{k-1}$ $(\Rightarrow u^* = x_i^k)$

$F = f_i + r_i$

$$F_i^k(x_i^{k-1}) - F_i^k(x_i^k)$$

lemma 2.1

$$\geq \frac{L_i^{k-1}}{2}\|x_i^k - \hat{x}_i^{k-1}\|^2 + L_i^{k-1}\langle \hat{x}_i^{k-1} - x_i^{k-1}, x_i^k - \hat{x}_i^{k-1} \rangle$$

$$F_i^k(x_i^{k-1}) - F_i^k(x_i^k) \geq \underbrace{\frac{L_i^{k-1}}{2}\|x_i^k - \hat{x}_i^{k-1}\|^2}_{①} + L_i^{k-1}\underbrace{\langle \hat{x}_i^{k-1} - x_i^{k-1}, x_i^k - \hat{x}_i^{k-1}\rangle}_{②}$$

$$① = \|x_i^k - \hat{x}_i^{k-1} + x_i^{k-1} - x_i^{k-1}\|_2^2$$
$$= \|x_i^k - x_i^{k-1}\|_2^2 + \|x_i^{k-1} - \hat{x}_i^{k-1}\|_2^2 + 2\langle x_i^k - x_i^{k-1}, x_i^{k-1} - \hat{x}_i^{k-1}\rangle$$

(how to think: expand it to get terms to cancel with ②)

$$\frac{L_i^{k-1}}{2}① + L_i^{k-1}② = \frac{L_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|_2^2$$
$$\left.\begin{aligned} &+ \frac{L_i^{k-1}}{2}\|x_i^{k-1} - \hat{x}_i^{k-1}\|_2^2 \\ &- L_i^{k-1}\|x_i^{k-1} - \hat{x}_i^{k-1}\|_2^2 \end{aligned}\right\} -\tfrac{1}{2}\|\cdot\|_2^2$$

Hence: $$F_i^k(x_i^{k-1}) - F_i^k(x_i^k) \geq \frac{L_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|_2^2$$
$$- \frac{L_i^{k-1}}{2}\|x_i^{k-1} - \hat{x}_i^{k-1}\|_2^2$$

Now cancel $\hat{x}_i^{k-1}$ via update $\hat{x}_i^{k-1} = x_i^{k-1} + \omega_i^{k-1}(x_i^{k-1} - x_i^{k-2})$

$$F_i^k(x_i^{k-1}) - F_i^k(x_i^k) \geq \frac{L_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|_2^2$$
$$- \frac{L_i^{k-1}}{2}(\omega_i^{k-1})^2\|x_i^{k-1} - x_i^{k-2}\|_2^2$$

$$\overset{(\star\star)}{\geq} \frac{L_i^{k-1}}{2}\|x_i^k - x_i^{k-1}\|_2^2$$
$$- \frac{L_i^{k-2}}{2}\delta_\omega^2\|x_i^{k-2} - x_i^{k-1}\|_2^2 \qquad \cdots (2.8)$$

$(\star\star)$ $\omega_i^{k-1} \leq \delta_\omega\sqrt{\frac{L_i^{k-2}}{L_i^{k-1}}} \iff -(\omega_i^{k-1})^2 \leq -\delta_\omega^2\frac{L_i^{k-2}}{L_i^{k-1}}$

(2.8) holds for update (c), but also (a), (b) as

$$F_i^k(x_i^{k-1}) - F_i^k(x_i^k) \geq \frac{L_i^{k-1}}{2}\|x_i^{k-1} - x_i^k\|_2^2$$

"Sufficient decrease condition"!!

(2.8) is on block i., now extend to all block

$$F(x^{k-1}) - F(x^k) = \sum_{i=1}^{S} \left(F_i^k(x_i^{k-1}) - F_i^k(x_i^k) \right)$$

$$\overset{(2.8)}{\geq} \sum_{i=1}^{S} \left(\frac{L_i^{k-1}}{2} \|x_i^{k-1} - x_i^k\|_2^2 - \frac{L_i^{k-2} \delta_w^2}{2} \|x_i^{k-1} - x_i^{k-2}\|_2^2 \right)$$

Telescoping sum:

$F(x^0) - F(x^1) = \text{——}$

$F(x^1) - F(x^2) = \text{——}$

$\vdots$

$\oplus \quad F(x^{k-1}) - F(x^k) = \text{——}$

$$F(x^0) - F(x^k) = \sum\sum(\cdot)$$

$$F(x^0) - F(x^k) \geq \sum_{k=1}^{K} \sum_{i=1}^{S} \left(\frac{L_i}{2} \|x_i^{k-1} - x_i^k\|_2^2 - \frac{L_i^{k-2} \delta_w^2}{2} \|x_i^{k-2} - x_i^{k-1}\|_2^2 \right)$$

expand on k

$$= \sum_{i=1}^{S} \Big(\frac{L_i^0}{2} \|x_i^0 - x_i^1\|^2 - \underbrace{\frac{L_i^{-1} \delta_w^2}{2} \|x_i^{-1} - x_i^0\|^2}_{0 \text{ as } x^{-1} = x^0}$$

$$+ \frac{L_i^1}{2} \|x_i^1 - x_i^2\|_2^2 - \frac{L_i^0 \delta_w^2}{2} \|x_i^0 - x_i^1\|^2$$

$$+ \frac{L_i^2}{2} \|x_i^2 - x_i^3\|^2 - \frac{L_i^1 \delta_w^2}{2} \|x_i^1 - x_i^2\|^2$$

$$+ \cdots$$

$$+ \frac{L_i^{k-1}}{2} \|x_i^{k-1} - x_i^k\|_2^2 - \frac{L_i^k \delta_w^2}{2} \|x_i^{k-2} - x_i^{k-1}\|_2^2 \Big)$$

$$= \sum_{i=1}^{S} \Big(\frac{(1-\delta_w^2) L_i^0}{2} \|x_i^{-1} - x_i^0\|^2$$

$$+ \frac{(1-\delta_w^2) L_i^1}{2} \|x_i^0 - x_i^1\|_2^2$$

$$+ \frac{(1-\delta_w^2) L_i^2}{2} \|x_i^1 - x_i^2\|_2^2$$

$$+ \cdots + \frac{(1-\delta_w^2) L_i^{k-1}}{2} \|x_i^{k-2} - x_i^{k-1}\|_2^2 \Big) + \underbrace{\frac{L_i^{k-1}}{2} \|x^{k-1} - x^k\|_2^2}_{\geq 0}$$

$$\geq \sum_{k=1}^{K} \sum_{i=1}^{S} \frac{(1-\delta_w^2) L_i^{k-1}}{2} \|x_i^{k-1} - x_i^k\|_2^2$$

$$\geq \sum_{k=1}^{K} \frac{(1-\delta_w^2) \ell_i}{2} \|x_i^{k-1} - x_i^k\|_2^2 \qquad (\ell_i \leq L_i^{k-1} \ \forall k)$$

Now we have

$$F(x^0) - F(x^k) \geq \underbrace{\frac{1-\delta_0^2}{2}\ell_i}_{c} \sum_{k=1}^{k} \|x^{k-1} - x^k\|^2$$

as F is lower bounded, $-\infty < \inf F \leq F(x^k)$

$\Rightarrow -F(x^k) < +\infty \Rightarrow F(x^0) - F(x^0) < +\infty$

$$\Rightarrow +\infty > F(x^0) - F(x^k) \geq c \sum_{k=1}^{k} \|x^{k-1} - x^k\|^2$$

$$\Rightarrow \sum_{k=1}^{k} \|x^{k-1} - x^k\|_2^2 < +\infty$$

$\Rightarrow \|x^{k-1} - x^k\|_2$ is square-summable!

$\Rightarrow$ the sequence $\{x^k\}$ produced by APG converge. //

Note. Lemma 2.2 said $\{x^k\}$ converge.
It does not say where it converge and how fast it converge.

"where" it converge is by the next lemma.

By Kurdyka-Łojasiewicz inequality, we have

$$\sum_{k=1}^{k} \|x^{k-1} - x^k\|_2^1 < +\infty$$