BREGMAN PROXIMAL ALGORITHMS FOR COMPOSITE AND FINITE-SUM NONCONVEX MINIMIZATION PROBLEMS

Themelis, Andreas
Kyushu University

Latafat, Puya
KU Leuven

Ahookhosh, Masoud
University of Antwerp

Patrinos, Panagiotis
KU Leuven

Bregman proximal algorithms

for composite and finite-sum nonconvex minimization problems

Andreas Themelis

Kyushu University

andreas.themelis@ees.kyushu-u.ac.jp

Puya Latafat

KU Leuven

Masoud Ahookhosh

University of Antwerp

Panagiotis Patrinos

KU Leuven

First-Order Methods with Convergence Analysis @ SIAM OP21

July 22, 2021

Outline

Proximal algorithm(s)

- Proximal point algorithm
- Simple*st* analysis

Bregman proximal algorithm(s)

- Why Bregman

Linesearch B-FBS: the BELLA algorithm

- Newton-type?
- Linesearch?

Block-coordinate B-FBS: Bregman Finito/MISO

Conclusions

Proximal algorithm(s)

Proximal point algorithm

Problem $$\operatorname*{minimize}_{x\in\mathbb{R}^n} \varphi(x) \tag{P}$$

A1. $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper, lsc

A2. $\inf \varphi > -\infty$

Proximal mapping

$$\operatorname{prox}_{\gamma\varphi}(x) = \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{\varphi(w) + \tfrac{1}{2\gamma}\|w - x\|^2\right\} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n \qquad (\gamma > 0)$$

Method **Proximal point algorithm (PPA)**

$$x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$$

$\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ is the extended-real line

lsc: lower semicontinuous

Proximal algorithm(s)

Proximal point algorithm: simple*st* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \leq \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\quad(\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi(x^{k+1})$$

$\hat{\partial}\varphi(z) = \left\{v \mid \liminf_{z\neq w\to z} \frac{\varphi(w)-\varphi(z)-\langle v,w-z\rangle}{\|w-z\|} \geq 0\right\}$ is the (Fréchet) subdifferential of φ at z

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

- P1. $\varphi^\gamma \leq \varphi$
- P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x}-x\|^2$
- P3. φ^γ **continuous**
- P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\quad(\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\leq} \varphi(x^{k+1})$$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \leq \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\quad(\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\leq} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \tag{SD}$$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \le \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \tfrac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\tfrac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\quad(\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\le} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \qquad \text{(SD)}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \leq \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\qquad(\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\leq} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \tag{SD}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star$
- if $(x^k)_{k\in K} \to x^\star$, then $\varphi(x^\star) \overset{lsc}{\leq} \varphi_\star \overset{\text{P3}}{=} \varphi^\gamma(x^\star) \overset{\text{P1}}{\leq} \varphi(x^\star)$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \leq \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\qquad(\bar{x}\in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\leq} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \qquad \text{(SD)}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star = \varphi(x^\star)$
- if $(x^k)_{k\in K} \to x^\star$, then $\varphi(x^\star) \overset{lsc}{\leq} \varphi_\star \overset{\text{P3}}{=} \varphi^\gamma(x^\star) \overset{\text{P1}}{\leq} \varphi(x^\star)$

Proximal algorithm(s)

Proximal point algorithm: simple*st* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \leq \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar x) + \frac{1}{2\gamma}\|\bar x - x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar x}{\gamma} \in \hat\partial\varphi(\bar x)$ $\quad(\bar x \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\leq} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \qquad \text{(SD)}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star = \varphi(x^\star)$
- if $(x^k)_{k\in K} \to x^\star$, then $\varphi(x^\star) \overset{lsc}{\leq} \varphi_\star \overset{\text{P3}}{=} \varphi^\gamma(x^\star) \overset{\text{P1}}{\leq} \varphi(x^\star)$
- (SD) $\Rightarrow \sum_{k\in\mathbb{N}} \frac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \leq \varphi^\gamma(x^0) - \inf\varphi$ $(< \infty)$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \leq \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\quad(\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\leq} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \tag{SD}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star = \varphi(x^\star)$
- if $(x^k)_{k\in K} \to x^\star$, then $\varphi(x^\star) \overset{lsc}{\leq} \varphi_\star \overset{\text{P3}}{=} \varphi^\gamma(x^\star) \overset{\text{P1}}{\leq} \varphi(x^\star)$
- (SD) $\Rightarrow \sum_{k\in\mathbb{N}} \frac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \leq \varphi^\gamma(x^0) - \inf\varphi$ $(< \infty)$

  $\Rightarrow 0 \leftarrow \frac{x^{k-1}-x^k}{\gamma}$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \le \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \tfrac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\qquad (\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\le} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \qquad \text{(SD)}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star = \varphi(x^\star)$
- if $(x^k)_{k\in K} \to x^\star$, then $\varphi(x^\star) \overset{lsc}{\le} \varphi_\star \overset{\text{P3}}{=} \varphi^\gamma(x^\star) \overset{\text{P1}}{\le} \varphi(x^\star)$
- (SD) $\Rightarrow \sum_{k\in\mathbb{N}} \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \le \varphi^\gamma(x^0) - \inf\varphi \; (< \infty)$

  $\Rightarrow 0 \leftarrow \frac{x^{k-1}-x^k}{\gamma} \overset{\text{P4}}{\in} \hat{\partial}\varphi(x^k)$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \leq \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x}-x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar{x}}{\gamma} \in \hat{\partial}\varphi(\bar{x})$ $\quad(\bar{x} \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\leq} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \tag{SD}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star = \varphi(x^\star)$
- if $(x^k)_{k\in K} \to x^\star$, then $\varphi(x^\star) \overset{lsc}{\leq} \varphi_\star \overset{\text{P3}}{=} \varphi^\gamma(x^\star) \overset{\text{P1}}{\leq} \varphi(x^\star)$
- (SD) $\Rightarrow \sum_{k\in\mathbb{N}} \frac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \leq \varphi^\gamma(x^0) - \inf\varphi$ $(< \infty)$

  $\Rightarrow 0 \leftarrow \frac{x^{k-1}-x^k}{\gamma} \overset{\text{P4}}{\in} \hat{\partial}\varphi(x^k)$, $\varphi(x^k) \to \varphi(x^\star)$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

Tool **Moreau envelope**

$$\varphi^\gamma(x) = \min_{w\in\mathbb{R}^n}\left\{\varphi(w) + \tfrac{1}{2\gamma}\|w-x\|^2\right\} : \mathbb{R}^n \to \mathbb{R}$$

P1. $\varphi^\gamma \le \varphi$

P2. $\varphi^\gamma(x) = \varphi(\bar x) + \frac{1}{2\gamma}\|\bar x - x\|^2$

P3. φ^γ **continuous**

P4. $\frac{x-\bar x}{\gamma} \in \hat\partial\varphi(\bar x)$ $\quad(\bar x \in \operatorname{prox}_{\gamma\varphi}(x))$

Convergence of PPA $\quad x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$

$$\varphi^\gamma(x^{k+1}) \overset{\text{P1}}{\le} \varphi(x^{k+1}) \overset{\text{P2}}{=} \varphi^\gamma(x^k) - \tfrac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \qquad \text{(SD)}$$

- (SD) $\Rightarrow \varphi^\gamma(x^k)$ & $\varphi(x^k) \searrow \varphi_\star = \varphi(x^\star)$
- if $(x^k)_{k\in K} \to x^\star$, then $\varphi(x^\star) \overset{lsc}{\le} \varphi_\star \overset{\text{P3}}{=} \varphi^\gamma(x^\star) \overset{\text{P1}}{\le} \varphi(x^\star)$
- (SD) $\Rightarrow \sum_{k\in\mathbb{N}} \frac{1}{2\gamma}\|x^{k+1}-x^k\|^2 \le \varphi^\gamma(x^0) - \inf\varphi$ $(<\infty)$

$$\Rightarrow 0 \leftarrow \frac{x^{k-1}-x^k}{\gamma} \overset{\text{P4}}{\in} \hat\partial\varphi(x^k),\ \varphi(x^k) \to \varphi(x^\star) \quad\Rightarrow\quad 0 \in \partial\varphi(x^\star)$$

$\partial\varphi(z) := \left\{v \mid \exists (z^k, v^k) \to (z,v) \text{ with } v^k \in \hat\partial\varphi(z^k)\ \&\ \varphi(z^k) \to \varphi(z)\right\}$

Proximal algorithm(s)

Proximal point algorithm: simple*st* convergence analysis

PPA

φ proper, lsc, lower bounded

$$x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$$

ω set of limit points of $(x^k)_{k\in\mathbb{N}}$:

1. $\varphi(x^k), \varphi^\gamma(x^k) \searrow \varphi_\star$
2. $\varphi \equiv \varphi^\gamma \equiv \varphi_\star$ on ω
3. $0 \in \partial\varphi(x^\star)$ for any $x^\star \in \omega$
4. φ coercive $\Rightarrow (x^k)_{k\in\mathbb{N}}$ bounded, and $\omega \neq \emptyset$ compact & connected

φ is coercive (AKA level bounded) if $\liminf_{\|x\|\to\infty} \varphi(x) = \infty$

Proximal algorithm(s)

Proximal point algorithm: simpl*est* convergence analysis

PPA

φ proper, lsc, lower bounded

$$x^{k+1} \in \operatorname{prox}_{\gamma\varphi}(x^k)$$

ω set of limit points of $(x^k)_{k\in\mathbb{N}}$:

1. $\varphi(x^k), \varphi^\gamma(x^k) \searrow \varphi_\star$
2. $\varphi \equiv \varphi^\gamma \equiv \varphi_\star$ on ω
3. $0 \in \partial\varphi(x^\star)$ for any $x^\star \in \omega$
4. φ coercive $\Rightarrow$ $(x^k)_{k\in\mathbb{N}}$ bounded, and $\omega \neq \emptyset$ compact & connected

PPA$_\lambda$ (λ-relaxed PPA)

$$\begin{cases} \bar{x}^k \in \operatorname{prox}_{\gamma\varphi}(x^k) \\ x^{k+1} = (1-\lambda)x^k + \lambda\bar{x}^k \end{cases} \overset{\text{same arguments}^\star}{\Longrightarrow} \varphi^\gamma(x^{k+1}) \leq \varphi^\gamma(x^k) - \tfrac{\lambda(2-\lambda)}{2\gamma}\|\bar{x}^k - x^k\|^2$$

All the claims of **PPA** hold for **PPA$_\lambda$** for any $\lambda \in (0,2)$

---

* $\varphi^\gamma(x^+) \leq \varphi(\bar{x}) + \frac{1}{2\gamma}\|\bar{x} - x^+\|^2 = \varphi^\gamma(x) - \frac{1}{2\gamma}\|\bar{x} - x\|^2 + \frac{1}{2\gamma}\|\bar{x} - x^+\|^2 = \varphi^\gamma(x) - \frac{\lambda(2-\lambda)}{2\gamma}\|\bar{x} - x\|^2$

Outline

Bregman proximal algorithm(s)

Why? ~ Structured problems

Proximal Point

☺ extremely simple analysis

☹ not "practical": $\operatorname{prox}_{\gamma\varphi}$ usually hard to compute

Proximal gradient method (AKA Forward-backward splitting)

In many applications, $\varphi =$ "smooth f" $+$ "easy-to-prox g"

$$\begin{aligned} x^{k+1} &\in \operatorname{prox}_{\gamma g}\left(x^k - \gamma \nabla f(x^k)\right) \\ &= \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \left\{ f(x^k) + \langle \nabla f(x^k), w - x^k \rangle + g(w) + \tfrac{1}{2\gamma}\|w - x^k\|^2 \right\} \end{aligned}$$

---

smooth: differentiable with ∇f Lipschitz continuous (Lipschitz constant $:= L_f$)

Bregman proximal algorithm(s)

Why? ~ Structured problems

Proximal Point

☺ extremely simple analysis

☹ not "practical": $\operatorname{prox}_{\gamma\varphi}$ usually hard to compute

Proximal gradient method (AKA **Forward-backward splitting**)

In many applications, $\varphi =$ "smooth f" $+$ "easy-to-prox g"

$$\begin{aligned} x^{k+1} &\in \operatorname{prox}_{\gamma g}\left(x^k - \gamma\nabla f(x^k)\right) \\ &= \operatorname*{arg\,min}_{w\in\mathbb{R}^n}\left\{f(x^k) + \langle\nabla f(x^k), w - x^k\rangle + g(w) + \tfrac{1}{2\gamma}\|w - x^k\|^2\right\} \\ &= \operatorname*{arg\,min}_{w\in\mathbb{R}^n}\left\{\varphi(w) + \mathrm{D}_{\hat h}(w, x^k)\right\} \end{aligned}$$

$\hat h \coloneqq \frac{1}{2\gamma}\|\cdot\|^2 - f$ strongly convex when $\gamma < {}^1\!/_{L_f}$

$$\mathrm{D}_h(z, x) \coloneqq h(z) - h(x) - \langle\nabla h(x), z - x\rangle$$

for a convex $h \in C^1(\mathbb{R}^n)$ is the Bregman distance induced by h

Bregman proximal algorithm(s)

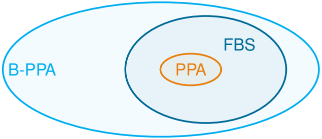
Why? ~ Structured problems

$$\varphi = \text{“}L_f\text{-smooth } f\text{”} + \text{“easy-to-prox } g\text{”}$$

$$\begin{aligned} x^{k+1} &\in \operatorname{prox}_{\gamma g}\left(x^k - \gamma \nabla f(x^k)\right) \\ &= \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \left\{\varphi(w) + \mathrm{D}_{\hat{h}}(w, x^k)\right\} \\ &:= \operatorname{prox}_{\varphi}^{\hat{h}}(x^k) \end{aligned}$$

FBS = Bregman PPA

with dfg $\hat{h} = \frac{1}{2\gamma}\|\cdot\|^2 - f$



dgf: distance-generating function (should comply with *essential smoothness, Legendreness, essential strict convexity*)

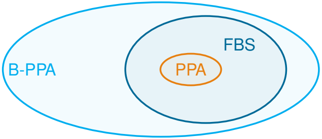
Bregman proximal algorithm(s)

Why? ~ Structured problems

$$\varphi = \text{“}L_f\text{-smooth } f\text{”} + \text{“easy-to-prox } g\text{”}$$

$$\begin{aligned} x^{k+1} &\in \operatorname{prox}_{\gamma g}\left(x^k - \gamma \nabla f(x^k)\right) \\ &= \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \left\{\varphi(w) + \mathrm{D}_{\hat{h}}(w, x^k)\right\} \\ &:= \operatorname{prox}_{\varphi}^{\hat{h}}(x^k) \end{aligned}$$

FBS = Bregman PPA
with dfg $\hat{h} = \frac{1}{2\gamma}\|\cdot\|^2 - f$



✓ Same results of **PPA** $\gamma < 1/L_f$

$$\begin{cases} \frac{1}{2\gamma}\|\cdot - \cdot\|^2 & \to \mathrm{D}_{\hat{h}} \\ \varphi^\gamma & \to \varphi^{\hat{h}} := \min_w \left\{\varphi(w) + \mathrm{D}_{\hat{h}}(w, \cdot)\right\} \end{cases}$$

✓ For **FBS**$_\lambda$: take $0 < \lambda < \frac{2}{1+\gamma L_f}$

$$\varphi^{\hat{h}}(x^+) \leq \varphi^{\hat{h}}(x) - \frac{2-\lambda(1+\gamma L_f)}{\lambda(1-\gamma L_f)} \mathrm{D}_{\hat{h}}(x^+, x)$$

✓ ($\varphi^{\hat{h}}$ is the **forward-backward envelope**)

P. Patrinos and A. Bemporad, *Proximal Newton methods for convex composite optimization,* In: *52th IEEE CDC*, **2013**

AT, L. Stella and P. Patrinos, *Forward-Backward Envelope for the Sum of Two Nonconvex Functions: Further Properties and Nonmonotone Linesearch Algorithms,* SIAM J Opt **28**(3):2274-2303, **2018**

Bregman proximal algorithm(s)

Why? ~ ...

With a suitable h

✓ **Uni-simplify** analysis

$\approx$ M. Teboulle, *A simplified view of first order methods for optimization,* MathProg **170**(1):67–96, Springer, **2018**

Bregman proximal algorithm(s)

Why? ~ ...

With a suitable h

✓ **Uni-simplify** analysis

≈ M. Teboulle, *A simplified view of first order methods for optimization,* MathProg **170**(1):67–96, Springer, **2018**

- Can enforce $x^k \in \operatorname{dom} h$: blend proximal and barrier methods

$$\operatorname*{minimize}_{x\in\mathbb{R}^n} \varphi(x) \quad \text{s.t.} \quad x \in C$$

where C is the closure of $\operatorname{dom} h$ (~~some~~ a lot! caution if φ nonconvex)

$\operatorname{dom} h := \{x \mid h(x) < \infty\}$

Bregman proximal algorithm(s)

Why? ~ ...

With a suitable h

- ✓ **Uni-simplify** analysis

  ≈ M. Teboulle, *A simplified view of first order methods for optimization,* MathProg **170**(1):67–96, Springer, **2018**

- ► Can enforce $x^k \in \operatorname{dom} h$: blend proximal and barrier methods

$$\operatorname*{minimize}_{x \in \mathbb{R}^n} \varphi(x) \quad \text{s.t.} \quad x \in C$$

  where C is the closure of $\operatorname{dom} h$ (~~some~~ a lot! caution if φ nonconvex)

- ✓ Extend applicability
  - ✓ make prox tractable
  - ✓ relax assumptions

f is L_f-**relative smooth** wrt h if $L_f h \pm f$ are convex

---

H. Bauschke, J. Bolte and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications,* Math Op Res **42**(2):330–348, INFORMS, **2016**

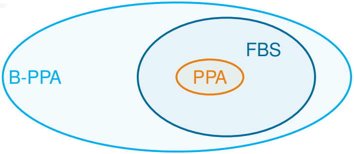
Bregman proximal algorithm(s)

Why? ~ ...

Bregman FBS

f L_f-smooth relative to a dgf h

$$
\begin{aligned}
x^{k+1} &\in \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{ f(x^k) + \langle \nabla f(x^k), w - x^k \rangle + g(w) + \tfrac{1}{\gamma}\, \mathrm{D}_h(w, x^k) \right\} \\
&= \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{ \varphi(w) + \mathrm{D}_{\hat{h}}(w, x^k) \right\} \\
&= \operatorname{prox}_{\varphi}^{\hat{h}}(x^k) \qquad \text{where } \hat{h} := \frac{1}{\gamma} h - f
\end{aligned}
$$



Bregman proximal algorithm(s)

Why? ~ ...

Bregman FBS

f L_f-smooth relative to a dgf h

$$x^{k+1} \in \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \left\{ f(x^k) + \langle \nabla f(x^k), w - x^k \rangle + g(w) + \tfrac{1}{\gamma} \mathrm{D}_h(w, x^k) \right\}$$

$$= \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \left\{ \varphi(w) + \mathrm{D}_{\hat{h}}(w, x^k) \right\}$$

$$= \operatorname{prox}_{\varphi}^{\hat{h}}(x^k) \qquad \text{where } \hat{h} := \frac{1}{\gamma} h - f$$

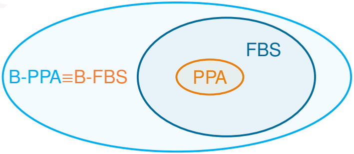
B-BFS $\equiv$ B-PPA

f L_f-smooth relative to dgf h

$\Rightarrow \hat{h} := \frac{1}{\gamma} h - f$ is a dgf

(with same properties)

- $\operatorname{dom} \hat{h} = \operatorname{dom} h$
- h (loc.) str.cvx/smooth $\Rightarrow$ $\hat{h}$ (loc.) str.cvx/smooth
- $\mathrm{D}_h(x^k, y^k) \to 0 \Rightarrow \mathrm{D}_{\hat{h}}(x^k, y^k) \to 0$
- $\ldots$



M. Ahookhosh, AT and P. Patrinos, *A Bregman Forward-Backward Linesearch Algorithm for Nonconvex Composite Optimization: Superlinear Convergence to Nonisolated Local Minima,* SIAM J Opt **31**(1):653-685, **2021**

Outline

Proximal algorithm(s)
- Proximal point algorithm
- Simple*st* analysis

Bregman proximal algorithm(s)
- Why Bregman

Linesearch B-FBS: the BELLA algorithm
- Newton-type?
- Linesearch?

Block-coordinate B-FBS: Bregman Finito/MISO

Conclusions

Linesearch B-FBS: the BELLA algorithm

Problem & objective

Problem $$\operatorname*{minimize}_{x\in\mathbb{R}^n} \varphi(x) \tag{P}$$

A1. $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper, lsc

A2. $\inf \varphi > -\infty$

Toolbox **Bregman proximal mapping**

$$\operatorname{prox}_{\varphi}^{h}(x) = \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \{\varphi(w) + \mathrm{D}_h(w,x)\} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$$

A3. h dgf with $\operatorname{dom} h = \mathbb{R}^n$

Goal

- Globalize **Newton-type methods** $x^+ = x + \tau d$
- by using **only** the **B-PPA** oracle

Linesearch B-FBS: the BELLA algorithm

Newton type?

$$x^+ = x + \tau d$$

But...

☹ φ^h usually nonsmooth

☹ or, anyway, $\nabla\varphi^h = \nabla^2 h\,[\mathrm{id} - \mathrm{prox}_\varphi^h]$

- even if applicable, Armijo-type linesearch would require $\nabla^2 h$
- (this was the original approach of `minFBE`)

Directions

Necessary optimality condition: $0 \in \mathrm{R}_\varphi^h(x^\star) := \left[\mathrm{id} - \mathrm{prox}_\varphi^h\right](x^\star)$

$$\Rightarrow \text{ Get } d \text{ via } \textbf{quasi-Newton} \text{ on } \mathrm{R}_\varphi^h$$

Requires (only)

✓ prox_φ^h

✓ direct linear algebra (scalar products)

L. Stella, AT and P. Patrinos, *Forward-backward quasi-Newton methods for nonsmooth optimization problems*, COAP **67**(3):443–487, Jul **2017**

Linesearch B-FBS: the BELLA algorithm

Linesearch?

$$x^+ = x + \tau d$$

Still...

- ☹ φ^h usually nonsmooth
- ☹ or, anyway, $\nabla\varphi^h = \nabla^2 h\,[\mathrm{id} - \mathrm{prox}_\varphi^h]$
- ☹ is d of *"descent"*?

And yet...

- ☺ φ^h is **continuous**
- ☺ $\varphi^h(\bar{x}) \leq \varphi(\bar{x}) = \varphi^h(x) - \mathrm{D}_h(\bar{x}, x) \qquad (\bar{x} \in \mathrm{prox}_\varphi^h(x))$

Linesearch B-FBS: the BELLA algorithm

Linesearch?

$$\cancel{x^+ = x + \tau d}$$

Still...

- ☹ φ^h usually nonsmooth
- ☹ or, anyway, $\nabla\varphi^h = \nabla^2 h\,[\mathrm{id} - \mathrm{prox}_\varphi^h]$
- ☹ is d of *"descent"*?

And yet...

- ☺ φ^h is **continuous**
- ☺ $\varphi^h(\bar{x}) \leq \varphi(\bar{x}) = \varphi^h(x) - \mathrm{D}_h(\bar{x}, x)$ $\qquad (\bar{x} \in \mathrm{prox}_\varphi^h(x))$

Solution (works for **any** d!)

$$x^+ = (1-\tau)\bar{x} + \tau(x + d)$$

reducing τ until

$$\varphi^h(x^+) \leq \varphi^h(x) - \sigma\,\mathrm{D}_h(\bar{x}, x) \qquad (\sigma \in (0,1) \text{ fixed})$$

Linesearch B-FBS: the BELLA algorithm

The BELLA algorithm

Bregman **E**nve**L**ope **L**inesearch **A**lgorithm

1. $\bar{x} \in \operatorname{prox}^h_\varphi(x)$
2. Choose $d \in \mathbb{R}^n$ and set $\tau = 1$
3. **While** $\varphi^h((1-\tau)\bar{x} + \tau(x+d)) \gneqq \varphi^h(x) - \sigma \, \mathrm{D}_h(\bar{x}, x)$ **do**

   $\tau \leftarrow {}^{\tau}/_{2}$
4. $x^+ = (1-\tau)\bar{x} + \tau(x+d)$

Convergence

$$\varphi^h(x^{k+1}) \leq \varphi^h(x^k) - \sigma \, \mathrm{D}_h(\bar{x}^k, x^k) \tag{SD}$$

Same convergence properties of **(B)-PPA**!

Extras... (under assumptions)

- global/R-linear with KL
- If directions d are "good", $\tau = 1$ eventually always accepted

KL: Kurdyka-Łojasiewicz

M. Ahookhosh, AT and P. Patrinos, *A Bregman Forward-Backward Linesearch Algorithm for Nonconvex Composite Optimization: Superlinear Convergence to Nonisolated Local Minima,* SIAM J Opt **31**(1):653-685, **2021**

Outline

Block coordinate B-FBS

Regularized finite-sum

Problem $$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, \varphi(x) \equiv \frac{1}{N}\sum_{i=1}^{N} f_i(x) + g(x) \tag{P}$$

A1. $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper, lsc

A2. f_i L_{f_i}-smooth relative to a dgf h_i

A3. $\inf \varphi > -\infty$

Consensus formulation

$$\underset{\boldsymbol{x}\in\mathbb{R}^{nN}}{\text{minimize}}\, \Phi(\boldsymbol{x}) \equiv \overbrace{\frac{1}{N}\sum_{i=1}^{N} f_i(x_i)}^{F(\boldsymbol{x})} + \overbrace{\frac{1}{N}\sum_{i=1}^{N} g(x_i) + \delta_\Delta(\boldsymbol{x})}^{G(\boldsymbol{x})} \tag{P$_\Delta$}$$

$\Delta := \left\{\boldsymbol{x} \in \mathbb{R}^{nN} \mid x_1 = x_2 = \cdots = x_N\right\}$ is the **consensus set**

$\delta_E(x) = \begin{cases} 0 & \text{if } x \in E, \\ \infty & \text{otherwise} \end{cases}$ is the indicator function of the set $E \subseteq \mathbb{R}^n$

Block coordinate B-FBS

Bregman Finito/MISO

Block-coordinate B-PPA (or **B-FBS**)

with dgf $\hat{H} \coloneqq \hat{h}_1 \times \ldots \times \hat{h}_N \qquad \hat{h}_i \coloneqq \frac{1}{\gamma_i} h_i - f_i, \qquad \gamma_i \in (0, {}^{N}/_{L_{f_i}})$

1. $\boldsymbol{u} \in \operatorname{prox}_{\Phi}^{\hat{H}}(\boldsymbol{x})$ % B-PPA = B-FBS
2. select/sample $i \in \{1, \ldots, N\}$
3. $x_j^+ = \begin{cases} u_i & \text{if } j = i, \\ x_j & \text{otherwise} \end{cases}$

$f \times g(x, y) \coloneqq (f(x), f(y))$ is the separable stacking of functions f and g

Block coordinate B-FBS

Bregman Finito/MISO

Block-coordinate B-PPA (or **B-FBS**)

with dgf $\hat{H} \coloneqq \hat{h}_1 \times \ldots \times \hat{h}_N \qquad \hat{h}_i \coloneqq \frac{1}{\gamma_i} h_i - f_i, \qquad \gamma_i \in (0, {}^{N}/_{L_{f_i}})$

1. $\boldsymbol{u} \in \operatorname{prox}_{\Phi}^{\hat{H}}(\boldsymbol{x})$ % B-PPA = B-FBS
2. select/sample $i \in \{1, \ldots, N\}$
3. $x_j^+ = \begin{cases} u_i & \text{if } j = i, \\ x_j & \text{otherwise} \end{cases}$

Not wasteful! it is **Bregman MISO/Finito**[+]

1. $z \in \arg\min_w \left\{ g(w) + \sum_i \frac{1}{\gamma_i} h_i(w) - \langle \tilde{s}, w \rangle \right\}$
2. select/sample $i \in \{1, \ldots, N\}$
3. update gradients table $\qquad s_j^+ = \begin{cases} \frac{1}{\gamma_i} \nabla h_i(z) - \frac{1}{N} \nabla f_i(z) & \text{if } j = i, \\ s_j & \text{otherwise} \end{cases}$
4. update aggregated B-FBS step $\tilde{s}^+ = \tilde{s} + (s_i^+ - s_i)$

✓ nonsmooth/nonconvex g
✓ independent stepsizes γ_i
✓ relative-smooth f_i

A. Defazio and J. Domke, *Finito: A faster, permutable incremental gradient method for big data problems,* In: *Int. Conf. Mach. Learn.*, **2014**

J. Mairal, *Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning,* SIAM J. Optim. **25**(2):829–855, **2015**

Block coordinate B-FBS

Convergence results

For ANY sampling

- ✓ Decrease on $\Phi^{\hat{H}}$
- ✓ φ coercive $\Rightarrow$ bounded sequence

Randomized/Shuffled/Cyclic

- ✓ **Sufficient** decrease on $\Phi^{\hat{H}}$
  - ✓ Convergence results of **(B)-PPA** apply!
  - ✓ (including global/R-linear with **KL**)

Extras...

- ✓ Q-linear rates with φ strongly convex (g or f_i can even be nonconvex!)

Analysis agnostic to "our" decomposition of φ!

P. Latafat, AT, M. Ahookhosh and P. Patrinos, *Bregman Finito/MISO for nonconvex regularized finite sum minimization without Lipschitz gradient continuity,* arXiv:2102.10312, **2021**

Outline

Proximal algorithm(s)
- Proximal point algorithm
- Simple*st* analysis

Bregman proximal algorithm(s)
- Why Bregman

Linesearch B-FBS: the BELLA algorithm
- Newton-type?
- Linesearch?

Block-coordinate B-FBS: Bregman Finito/MISO

Conclusions

Conclusions

In this talk...

What we knew

- Proximal point: elegant, simple, and powerful
- Bregman divergence: the key enhancement

What we knew but maybe didn't care

- **B-PPA = B-FBS**
- Separating tasks: $\overbrace{\text{minimize}\,\varphi}^{\text{analysis}}$ VS $\overbrace{\text{minimize}\,f+g\ldots\ \text{ s.t. }\ \ldots}^{\text{algorithm}}$

What maybe we didn't know

- Continuity + sufficient decrease → globalize fast local methods
  - ☺ the **BELLA** algorithm
- Block-coordinate proximal point → incremental proximal algorithms
  - ☺ **Bregman MISO/Finito^{+}**

M. Ahookhosh, AT and P. Patrinos, *A Bregman Forward-Backward Linesearch Algorithm for Nonconvex Composite Optimization: Superlinear Convergence to Nonisolated Local Minima,* SIAM J Opt **31**(1):653-685, **2021**

P. Latafat, AT, M. Ahookhosh and P. Patrinos, *Bregman Finito/MISO for nonconvex regularized finite sum minimization without Lipschitz gradient continuity,* arXiv:2102.10312, **2021**