# Asynchronous Block-Iterative Proximal Processing of Large Data Sets

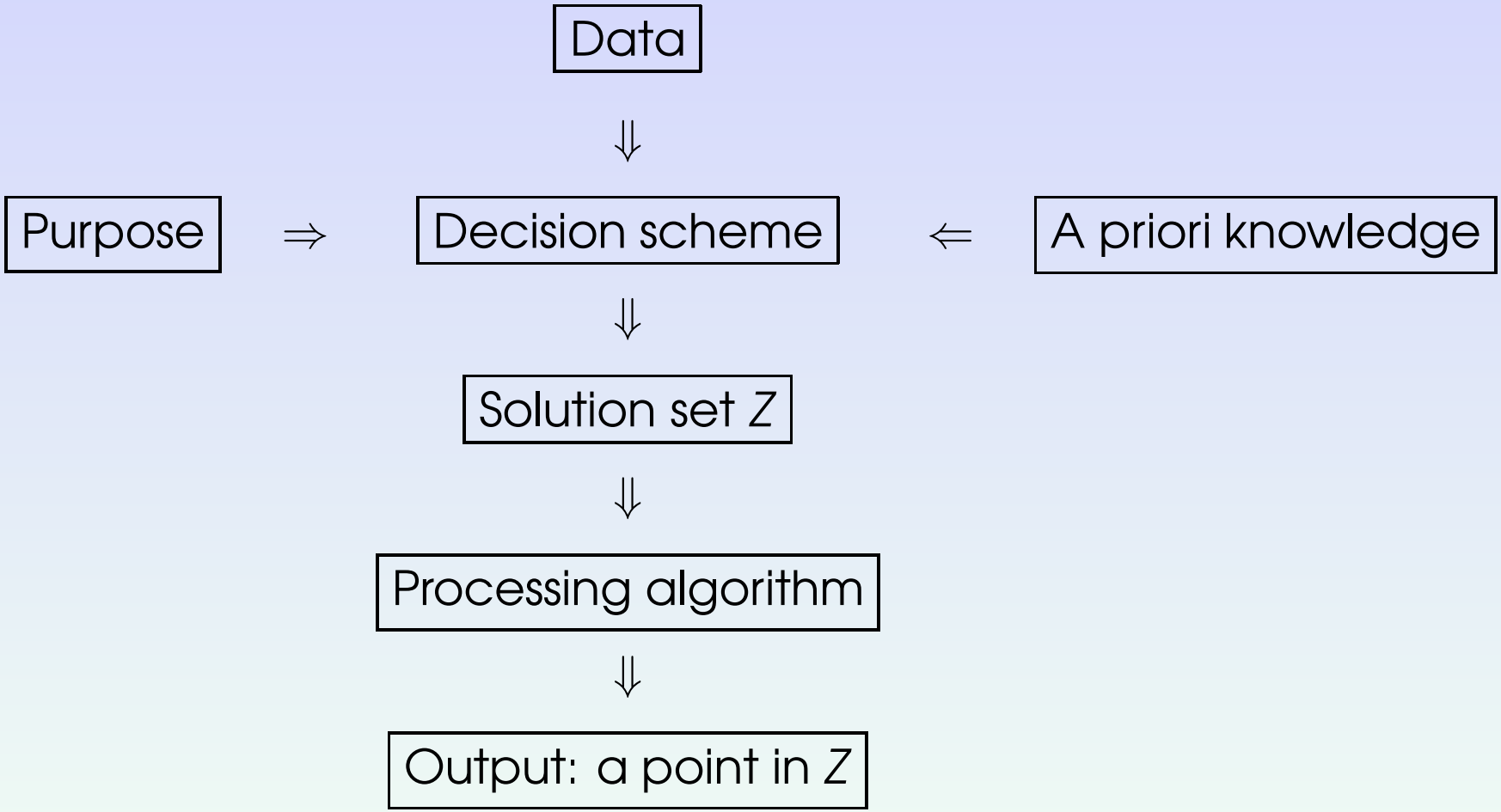
Patrick L. Combettes

Laboratoire Jacques-Louis Lions
Faculté de Mathématiques
Université Pierre et Marie Curie – Paris 6
75005 Paris, France

**Joint work with Jonathan Eckstein, Rutgers University**

Optimization Without Borders, Les Houches, February 9, 2016

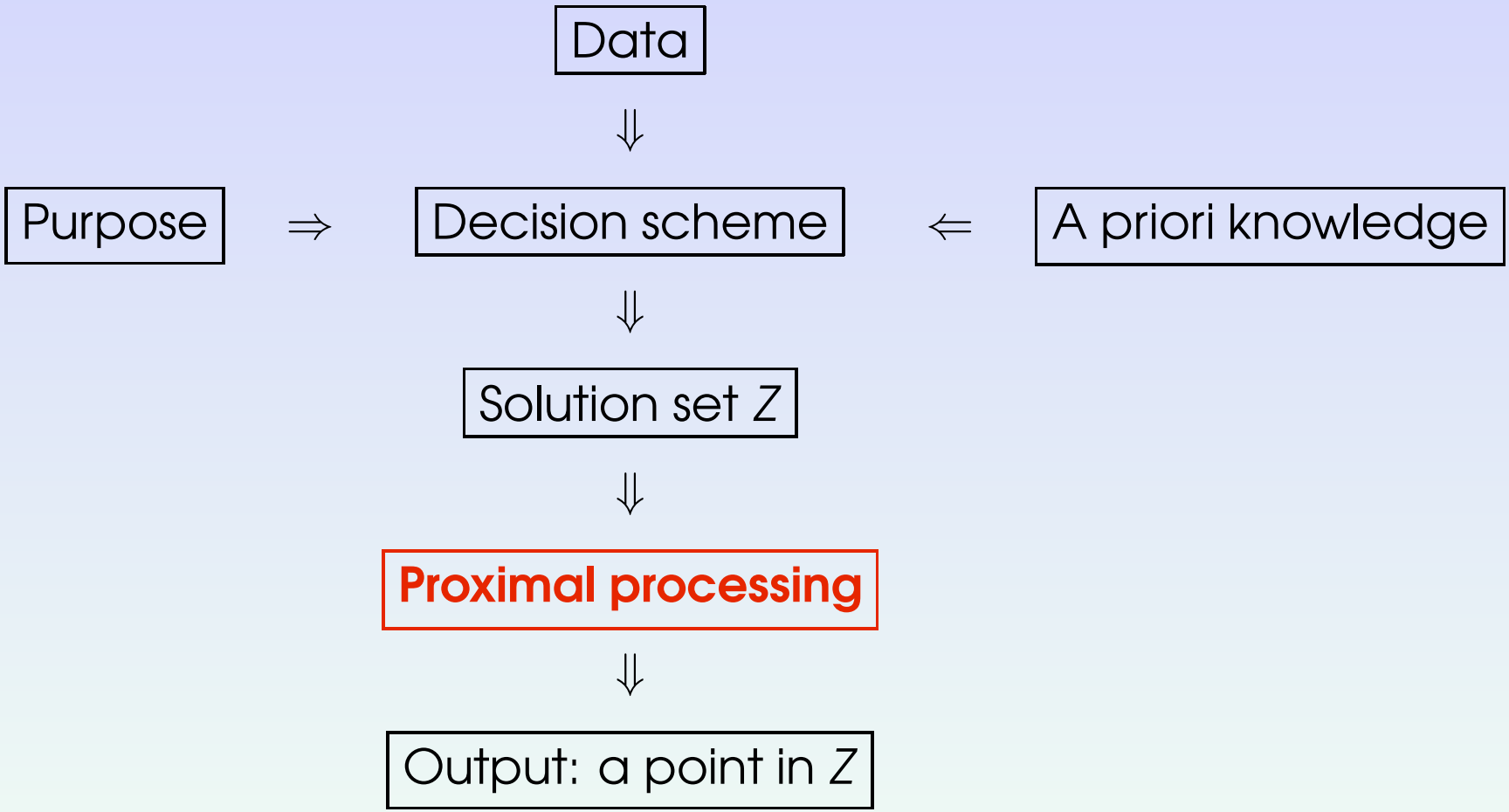
# Extracting information from data



# Optimization in early data science

- Ca. 1750: emergence of the idea of collecting many data sets and aggregating them instead of picking a "best" one.
- Fitting geodesic data by the method of least deviations:
  - R. J. Boscovich, De literaria expeditione per pontificiam ditionem et synopsis... *Bononiensi Scientiarum et Artum Inst. atque Acad. Comment.*, 1757.
  - P. S. Laplace, Sur quelques points du système du monde, *Mémoires Acad. Royale Sci. Paris*, 1789.
- Fitting astronomical data by the method of least squares:
  - A. M. Legendre, *Nouvelles Méthodes pour la Détermination de l'Orbite des Comètes*. Courcier, Paris, 1805.
  - C. F. Gauss, *Theoria Motus Corporum Coelestium*. Perthes and Besser, Hamburg, 1809.
- Gradient method for astronomical data processing:
  - A. Cauchy, Méthode générale pour la résolution des systèmes d'équations simmultanées, *C. R. Acad. Sci. Paris*, 1847.

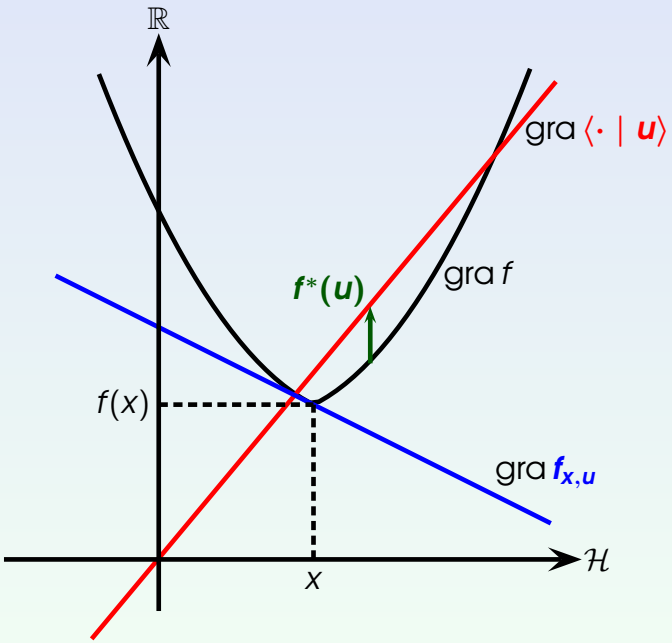
# Proximal data processing



# Basic notation

- $\mathcal{H}$: real Hilbert space
- $\Gamma_0(\mathcal{H})$: proper lower semicontinuous convex functions $f : \mathcal{H} \to ]-\infty, +\infty]$
- $f^*$ : $u \mapsto \sup_{x \in \mathcal{H}} \langle x \mid u \rangle - f(x)$ is the Legendre conjugate of $f$
- The subdifferential of $f$ at $x \in \mathcal{H}$ is

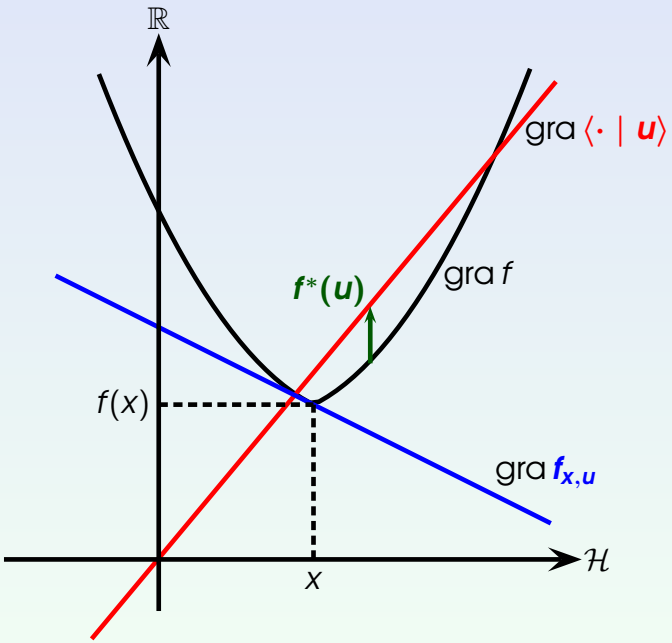
$$\partial f(x) = \big\{ u \in \mathcal{H} \mid (\forall y \in \mathcal{H}) \ \underbrace{\langle y - x \mid u \rangle + f(x)}_{f_{x,u}(y)} \leqslant f(y) \big\}$$



# Basic notation

- $\mathcal{H}$: real Hilbert space
- $\Gamma_0(\mathcal{H})$: proper lower semicontinuous convex functions $f : \mathcal{H} \to ]-\infty, +\infty]$
- $f^* : u \mapsto \sup_{x \in \mathcal{H}} \langle x \mid u \rangle - f(x)$ is the Legendre conjugate of $f$
- The subdifferential of $f$ at $x \in \mathcal{H}$ is

$$\partial f(x) = \big\{ u \in \mathcal{H} \mid (\forall y \in \mathcal{H}) \ \underbrace{\langle y - x \mid u \rangle + f(x)}_{f_{x,u}(y)} \leqslant f(y) \big\}$$




Fermat's rule:
$x$ minimizes $f \Leftrightarrow 0 \in \partial f(x)$

# Moreau's proximity operator

- In 1962, Jean Jacques Moreau (1923–2014) introduced the ***proximity operator*** of a function $f \in \Gamma_0(\mathcal{H})$

$$\operatorname{prox}_f : x \mapsto \operatorname*{argmin}_{y \in \mathcal{H}} f(y) + \frac{1}{2}\|x - y\|^2$$

  to study problems in unilateral mechanics

- Proximity operators turn out to be very effective tools for modeling and solving data-driven problems:
  - PLC, Convexité et signal, *Proc. SMAI Annual Conf.*, Pompadour, France, April 2001
  - PLC and V. R. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.*, vol. 4, 2005
  - PLC and J.-C. Pesquet, Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer, New York, 2011

# Proximity operators

- Many common convex functions in data processing (statistics, machine learning, image recovery, data denoising, support vector machine, signal processing) have explicit proximity operators:
  - $\ell_1$ norm
  - Shatten norm
  - nuclear norm
  - Huber's function
  - Berhu function
  - elastic net regularizer
  - hinge loss
  - Fisher information
  - distance function
  - Vapnik's $\varepsilon$-insensitive loss
  - Burg's entropy
  - $\phi$-divergences
  - etc.

# Proximity operators

- Basic properties:
  - $p = \mathrm{prox}_f x \Leftrightarrow x - p \in \partial f(p)$
  - $\mathrm{prox}_f + \mathrm{prox}_{f^*} = \mathrm{Id}$ (Moreau's decomposition)
    - For $f = \iota_V$, $V$ a closed vector subspace: $P_V + P_{V^\perp} = \mathrm{Id}$
    - $\mathrm{prox}_{\rho|\cdot|} = \mathrm{Id} - \mathrm{prox}_{(\rho|\cdot|)^*} = \mathrm{Id} - P_{[-\rho,\rho]} = \mathrm{soft}_\rho$
  - $(\mathrm{prox}_f x, x - \mathrm{prox}_f x) = (\mathrm{prox}_f x, \mathrm{prox}_{f^*} x) \in \mathrm{gra}\, \partial f$
  - Fix $\mathrm{prox}_f = \mathrm{Argmin}\, f$
  - $\|\mathrm{prox}_f x - \mathrm{prox}_f y\| \leqslant \|x - y\|$
- The last two properties suggest the conceptual algorithm

$$x_{n+1} = \mathrm{prox}_f x_n = x_n - \nabla(f^* \,\square\, \|\cdot\|^2/2)x_n$$

to minimize $f$

# Proximity operators

- Basic properties:
  - $p = \operatorname{prox}_f x \Leftrightarrow x - p \in \partial f(p)$
  - $\operatorname{prox}_f + \operatorname{prox}_{f^*} = \mathrm{Id}$ (Moreau's decomposition)
    - For $f = \iota_V$, $V$ a closed vector subspace: $P_V + P_{V^\perp} = \mathrm{Id}$
    - $\operatorname{prox}_{\rho|\cdot|} = \mathrm{Id} - \operatorname{prox}_{(\rho|\cdot|)^*} = \mathrm{Id} - P_{[-\rho,\rho]} = \operatorname{soft}_\rho$
  - $(\operatorname{prox}_f x, x - \operatorname{prox}_f x) = (\operatorname{prox}_f x, \operatorname{prox}_{f^*} x) \in \operatorname{gra} \partial f$
  - Fix $\operatorname{prox}_f = \operatorname{Argmin} f$
  - $\|\operatorname{prox}_f x - \operatorname{prox}_f y\|^2 \leqslant \|x - y\|^2 - \|\operatorname{prox}_{f^*} x - \operatorname{prox}_{f^*} y\|^2$
- The last two properties suggest the conceptual algorithm

$$x_{n+1} = \operatorname{prox}_f x_n = x_n - \nabla(f^* \,\square\, \|\cdot\|^2/2)x_n$$

to minimize $f$

# Forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$, where $f \in \Gamma_0(\mathcal{H})$, $g\colon \mathcal{H} \to \mathbb{R}$ convex, differentiable, $\nabla g$ is $1/\beta$-Lipschitz-continuous
- The sequence constructed by the algorithm

$$x_{n+1} = \operatorname{prox}_{\gamma f}(x_n - \gamma (\nabla g(x_n)))$$

  - $0 < \gamma < 2\beta$ (Mercier, 1979)

converges weakly to a point in $Z$

# Forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$, where $f \in \Gamma_0(\mathcal{H})$, $g\colon \mathcal{H} \to \mathbb{R}$ convex, differentiable, $\nabla g$ is $1/\beta$-Lipschitz-continuous
- The sequence constructed by the algorithm

$$x_{n+1} = \operatorname{prox}_{\gamma_n f}\big(x_n - \gamma_n (\nabla g(x_n))\big)$$

  - $0 < \gamma < 2\beta$ (Mercier, 1979)
  - $0 < \inf_{n\in\mathbb{N}} \gamma_n \leqslant \sup_{n\in\mathbb{N}} \gamma_n < 2\beta$ (Tseng, 1990)

converges weakly to a point in $Z$

# Forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$, where $f \in \Gamma_0(\mathcal{H})$, $g\colon \mathcal{H} \to \mathbb{R}$ convex, differentiable, $\nabla g$ is $1/\beta$-Lipschitz-continuous
- The sequence constructed by the algorithm

$$x_{n+1} = \operatorname{prox}_{\gamma_n f}\big(x_n - \gamma_n \quad (\nabla g(x_n) + b_n)\big) + a_n$$

  - $0 < \gamma < 2\beta$ (Mercier, 1979)
  - $0 < \inf_{n\in\mathbb{N}} \gamma_n \leqslant \sup_{n\in\mathbb{N}} \gamma_n < 2\beta$ (Tseng, 1990)
  - $\sum_{n\in\mathbb{N}} \|a_n\| < +\infty$, $\sum_{n\in\mathbb{N}} \|b_n\| < +\infty$ (PLC, 2004)

converges weakly to a point in $Z$

# Forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$, where $f \in \Gamma_0(\mathcal{H})$, $g\colon \mathcal{H} \to \mathbb{R}$ convex, differentiable, $\nabla g$ is $1/\beta$-Lipschitz-continuous
- The sequence constructed by the algorithm

$$x_{n+1} = x_n + \lambda_n\big(\operatorname{prox}_{\gamma_n f}\big(x_n - \gamma_n \quad (\nabla g(x_n) + b_n)\big) + a_n - x_n\big)$$

  - $0 < \gamma < 2\beta$ (Mercier, 1979)
  - $0 < \inf_{n\in\mathbb{N}} \gamma_n \leqslant \sup_{n\in\mathbb{N}} \gamma_n < 2\beta$ (Tseng, 1990)
  - $\sum_{n\in\mathbb{N}} \|a_n\| < +\infty$, $\sum_{n\in\mathbb{N}} \|b_n\| < +\infty$ (PLC, 2004)
  - $(\lambda_n)_{n\in\mathbb{N}}$ in $]0, 1]$, $\inf_{n\in\mathbb{N}} \lambda_n > 0$ (PLC, 2004)

converges weakly to a point in $Z$

# Forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$, where $f \in \Gamma_0(\mathcal{H})$, $g\colon \mathcal{H} \to \mathbb{R}$ convex, differentiable, $\nabla g$ is $1/\beta$-Lipschitz-continuous
- The sequence constructed by the algorithm

$$x_{n+1} = x_n + \lambda_n\big(\operatorname{prox}^{U_n}_{\gamma_n f}\big(x_n - \gamma_n U_n^{-1}(\nabla g(x_n) + b_n)\big) + a_n - x_n\big)$$

  - $0 < \gamma < 2\beta$ (Mercier, 1979)
  - $0 < \inf_{n\in\mathbb{N}} \gamma_n \leqslant \sup_{n\in\mathbb{N}} \gamma_n < 2\beta$ (Tseng, 1990)
  - $\sum_{n\in\mathbb{N}} \|a_n\| < +\infty$, $\sum_{n\in\mathbb{N}} \|b_n\| < +\infty$ (PLC, 2004)
  - $(\lambda_n)_{n\in\mathbb{N}}$ in $]0, 1]$, $\inf_{n\in\mathbb{N}} \lambda_n > 0$ (PLC, 2004)
  - $(1+\eta_n)U_{n+1} \succcurlyeq U_n = U_n^* \succcurlyeq \alpha \mathrm{Id}$, $\alpha > 0$, $\eta_n \geqslant 0$, $\sum_{n\in\mathbb{N}} \eta_n < +\infty$ (PLC&Vũ, 2012)

  converges weakly to a point in $Z$

# Forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$, where $f \in \Gamma_0(\mathcal{H})$, $g\colon \mathcal{H} \to \mathbb{R}$ convex, differentiable, $\nabla g$ is $1/\beta$-Lipschitz-continuous
- The sequence constructed by the algorithm

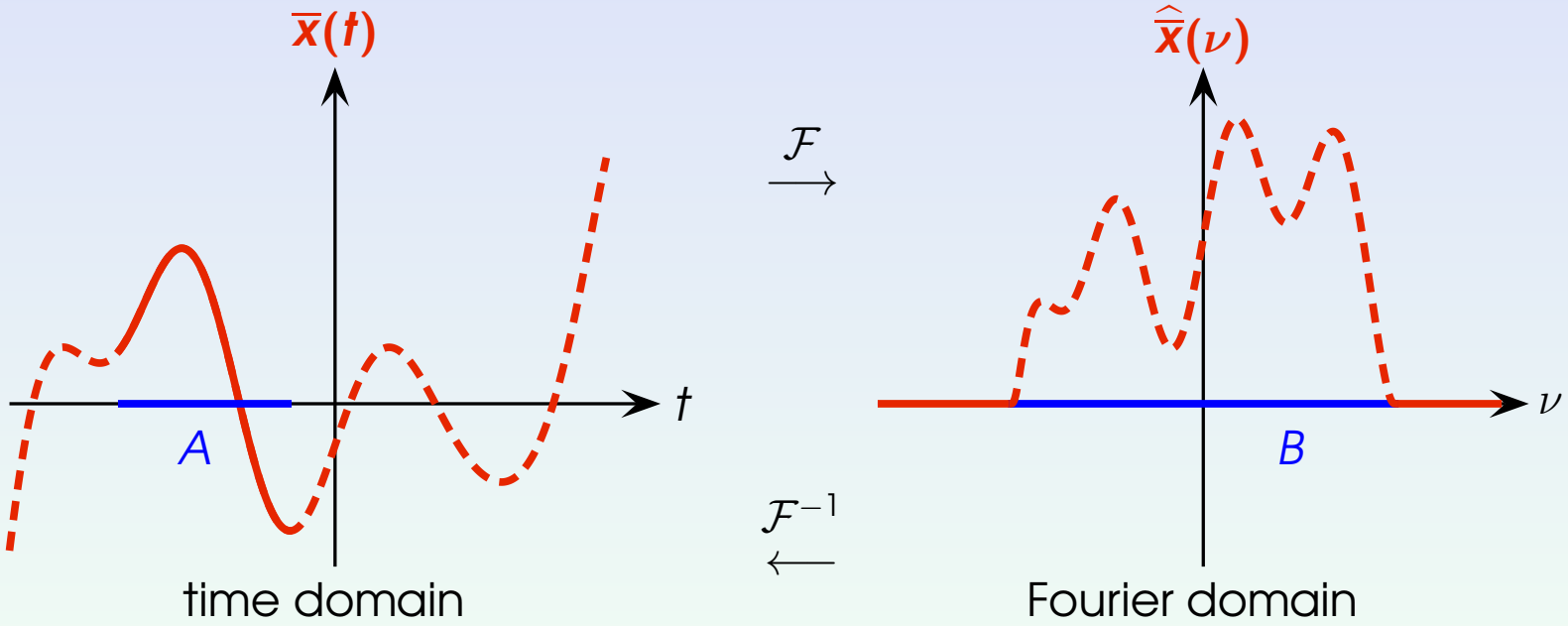
$$x_{n+1} = x_n + \lambda_n\big(\operatorname{prox}^{U_n}_{\gamma_n f}\big(x_n - \gamma_n U_n^{-1}(\nabla g(x_n) + b_n)\big) + a_n - x_n\big)$$

  - $0 < \gamma < 2\beta$ (Mercier, 1979)
  - $0 < \inf_{n\in\mathbb{N}} \gamma_n \leqslant \sup_{n\in\mathbb{N}} \gamma_n < 2\beta$ (Tseng, 1990)
  - $\sum_{n\in\mathbb{N}} \|a_n\| < +\infty$, $\sum_{n\in\mathbb{N}} \|b_n\| < +\infty$ (PLC, 2004)
  - $(\lambda_n)_{n\in\mathbb{N}}$ in $]0, 1]$, $\inf_{n\in\mathbb{N}} \lambda_n > 0$ (PLC, 2004)
  - $(1+\eta_n)U_{n+1} \succcurlyeq U_n = U_n^* \succcurlyeq \alpha \mathrm{Id}$, $\alpha > 0$, $\eta_n \geqslant 0$, $\sum_{n\in\mathbb{N}} \eta_n < +\infty$ (PLC&Vũ, 2012)

  converges weakly to a point in $Z$
- Also: almost surely weakly convergent versions with random block-coordinate sweeping (PLC&Pesquet, SIOPT, July 2015) and/or stochastic approximations (PLC&Pesquet, Pure Appl. Funct. Anal., Jan. 2016)

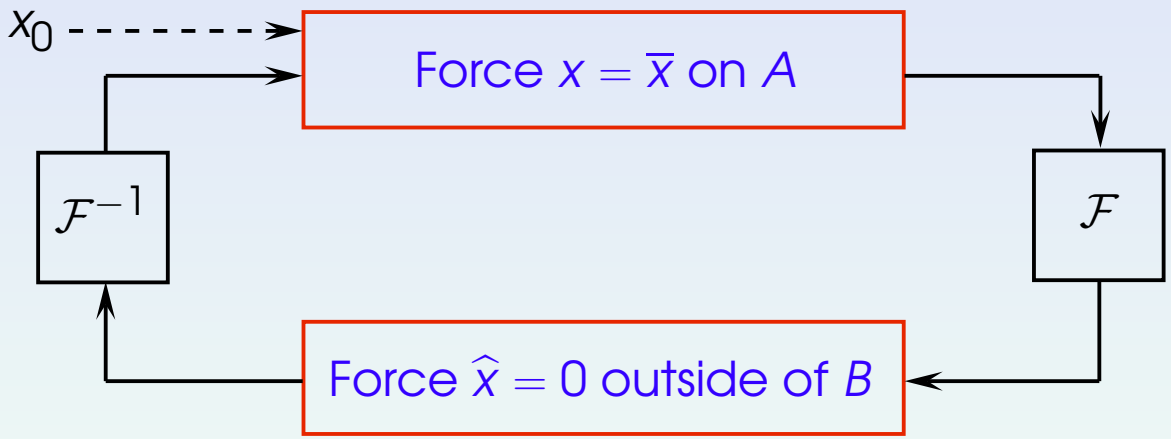
# Example: Band-limited extrapolation (1974-1975)

- **Purpose:** Reconstruct a signal $\overline{x}$
- **Data:** $\overline{x}$ is observed over some region $A$
- **Prior knowledge:** $\overline{x}$ is band-limited (its Fourier transform has compact support $B$)
- **Decision scheme:** Find a signal which is consistent with the above two properties (feasibility)



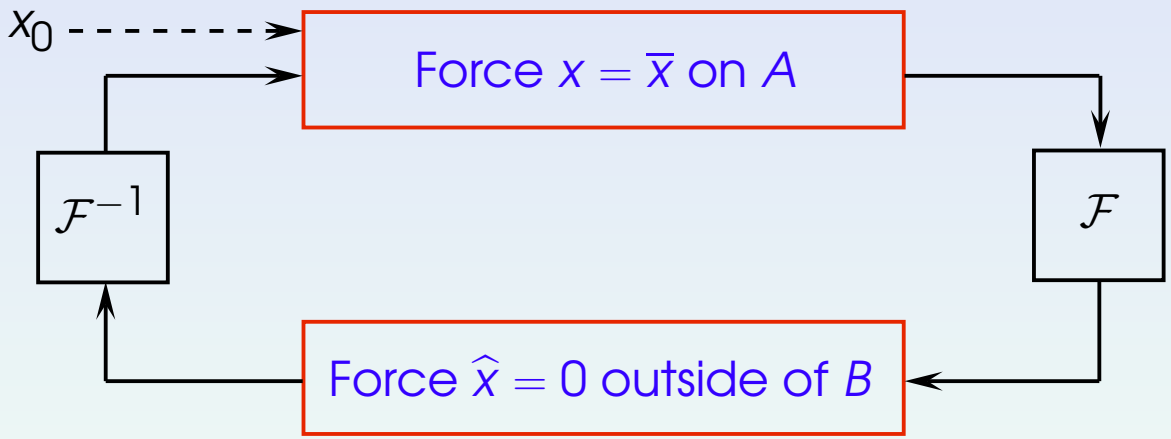
# Example: Band-limited extrapolation (1974-1975)

- **Purpose:** Reconstruct a signal $\overline{x}$
- **Data:** $\overline{x}$ is observed over some region $A$
- **Prior knowledge:** $\overline{x}$ is band-limited (its Fourier transform has compact support $B$)
- **Decision scheme:** Find a signal which is consistent with the above two properties (feasibility)
- **Papoulis' algorithm:**



# Example: Band-limited extrapolation (1974-1975)

- **Purpose:** Reconstruct a signal $\overline{x}$
- **Data:** $\overline{x}$ is observed over some region $A$
- **Prior knowledge:** $\overline{x}$ is band-limited (its Fourier transform has compact support $B$)
- **Decision scheme:** Find a signal which is consistent with the above two properties (feasibility)
- **Papoulis' algorithm:**

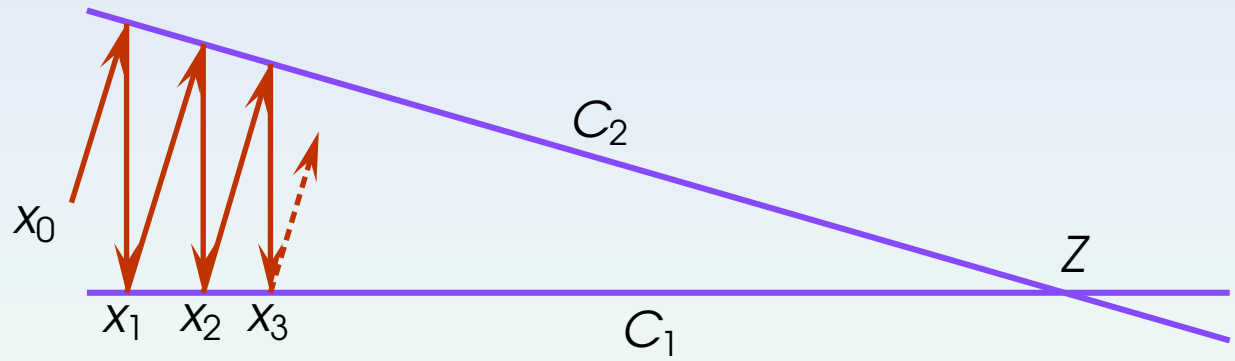



- **Gerchberg's algorithm:** Reconstruct an image with known support from limited diffraction data... at the speed of light!

# Example: Band-limited extrapolation (1974-1975)

- Set $f = \iota_{C_1}$, where $C_1 = \{x \in L^2(\mathbb{R}) \mid \widehat{x}|_{\complement B} = 0\}$
- Set $g = d^2_{C_2}/2$, where $C_2 = \{x \in L^2(\mathbb{R}) \mid x|_A = \overline{x}|_A\}$
- Solution set: $Z = C_1 \cap C_2$
- Data processing algorithm (forward-backward):

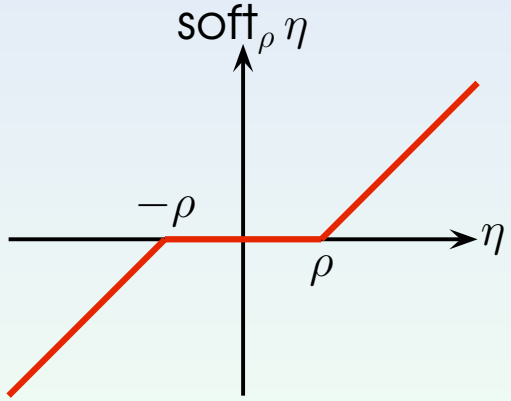
$$x_{n+1} = \operatorname{prox}_f(x_n - \nabla g(x_n)) = P_1 P_2 x_n,$$



# Data processing by iterative soft thresholding

- **Purpose:** Extract a signal/features $\overline{x}$
- **Data:** A noisy, linearly transformed (a blur, Radon transform in tomography, etc.) observation $r = L\overline{x} + w$
- **Prior knowledge:** $\overline{x}$ has a sparse decomposition in some orthonormal basis $(e_k)_{k\in\mathbb{N}}$, $L$ is known, $\|L\| = 1$
- **Decision scheme:** Find a sparse signal which is consistent with the observation
- **Empirical algorithm:** (ca. 2002-03)

$$x_{n+1} = \sum_{k\in\mathbb{N}} \operatorname{soft}_\rho \left( \langle \overbrace{x_n - L^*(Lx_n - r)}^{\text{Landweber step}} \mid e_k \rangle \right) e_k$$

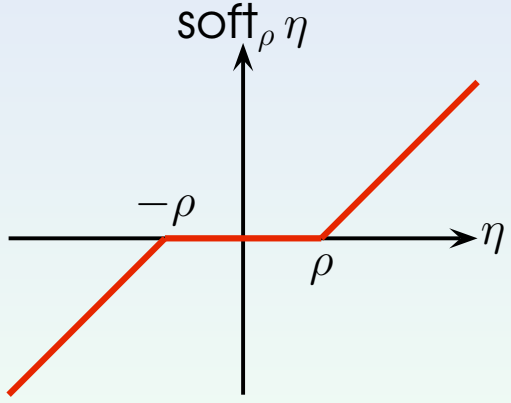


# Data processing by iterative soft thresholding

- **Purpose:** Extract a signal/features $\overline{x}$
- **Data:** A noisy, linearly transformed (a blur, Radon transform in tomography, etc.) observation $r = L\overline{x} + w$
- **Prior knowledge:** $\overline{x}$ has a sparse decomposition in some orthonormal basis $(e_k)_{k\in\mathbb{N}}$, $L$ is known, $\|L\| = 1$
- **Decision scheme:** Find a sparse signal which is consistent with the observation
- **Empirical algorithm:** (ca. 2002-03)

$$x_{n+1} = \sum_{k\in\mathbb{N}} \mathrm{soft}_\rho\left(\langle \overbrace{x_n - L^*(Lx_n - r)}^{\text{Landweber step}} \mid e_k\rangle\right) e_k$$
$$= \mathrm{prox}_f(x_n - \gamma\nabla g(x_n)),$$

where $f = \rho\sum_k |\langle\cdot \mid e_k\rangle|$, $g = \|L\cdot - r\|^2/2$



# Further properties of forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$
- $x_{n+1} = x_n + \lambda_n \Big( \operatorname{prox}_{\gamma_n f} \big( x_n - \gamma_n \nabla g(x_n) \big) - x_n \Big)$, $\varepsilon \leqslant \gamma_n \leqslant (2-\varepsilon)\beta$
- $(\forall n \in \mathbb{N})(\forall z \in Z) \quad \|x_{n+1} - z\| \leqslant \|x_n - z\|$: Fejér monotonicity
- $(\forall n \in \mathbb{N}) \quad (f+g)(x_{n+1}) \leqslant (f+g)(x_n)$
- Convergence is only weak
- Even in the finite dimensional or the linear case, no (upper bound on the worst) rate of convergence of $\|x_n - x_\infty\|$ exists
- $\sum_{n\in\mathbb{N}} \big|(f+g)(x_n) - \inf(f+g)(\mathcal{H})\big|^2 < +\infty$
- If $\sum_{n\in\mathbb{N}}(1-\lambda_n) < +\infty$, $(f+g)(x_n) - \inf(f+g)(\mathcal{H}) = o(1/n)$ (PLC, Salzo, Villa, arxiv, 2015)
- In the case of the projected gradient method, some form of the above results already in:
  - E. S. Levitin and B. T. Polyak, Constrained minimization methods, *Comput. Math. Math. Phys.*, vol. 6, pp. 1–50, 1966

# Birthday gift to Yurii: The 1966 Levitin-Polyak paper



*Theorem 5.1*

Let $Q$ be a bounded closed convex set of Hilbert space $H$, $f(x)$ a functional differentiable on $Q$, where $f'(x)$ satisfies a Lipschitz condition with constant $M$, and $0 < \varepsilon_1 \leqslant \alpha_n \leqslant 2/(M + 2\varepsilon_2)$, $\varepsilon_2 > 0$. Then sequence (5.1) has the following properties:

(1) $f(x^n)$ is monotonically decreasing and $\lim_{n\to\infty} \|x^{n+1} - x^n\| = 0$;

(2) if $f(x)$ is convex, then

$$\lim_{n\to\infty} f(x^n) = f^* = \inf_{x\in Q} f(x),$$

where $f(x^n) - f^* \leqslant c/n$, and a subsequence of $x^n$ exists, weakly convergent to the minimum $x^*$;

(3) if $f(x)$ is strictly convex or $Q$ is strictly convex, while $f'(x) \neq 0$ on $Q$, then $x^n$ is weakly convergent to the (unique) minimum $x^*$;

(4) if $f(x)$ is uniformly convex or $Q$ is uniformly convex, while $f'(x) \neq 0$ on $Q$, then $x^n$ is strongly convergent to $x^*$;

# Further properties of forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$
- $x_{n+1} = x_n + \lambda_n \Big( \operatorname{prox}_{\gamma_n f} \big( x_n - \gamma_n \nabla g(x_n) \big) - x_n \Big)$, $\varepsilon \leqslant \gamma_n \leqslant (2 - \varepsilon)\beta$
- $(\forall n \in \mathbb{N})(\forall z \in Z) \quad \|x_{n+1} - z\| \leqslant \|x_n - z\|$: Fejér monotonicity
- $(\forall n \in \mathbb{N}) \quad (f + g)(x_{n+1}) \leqslant (f + g)(x_n)$
- Convergence is only weak

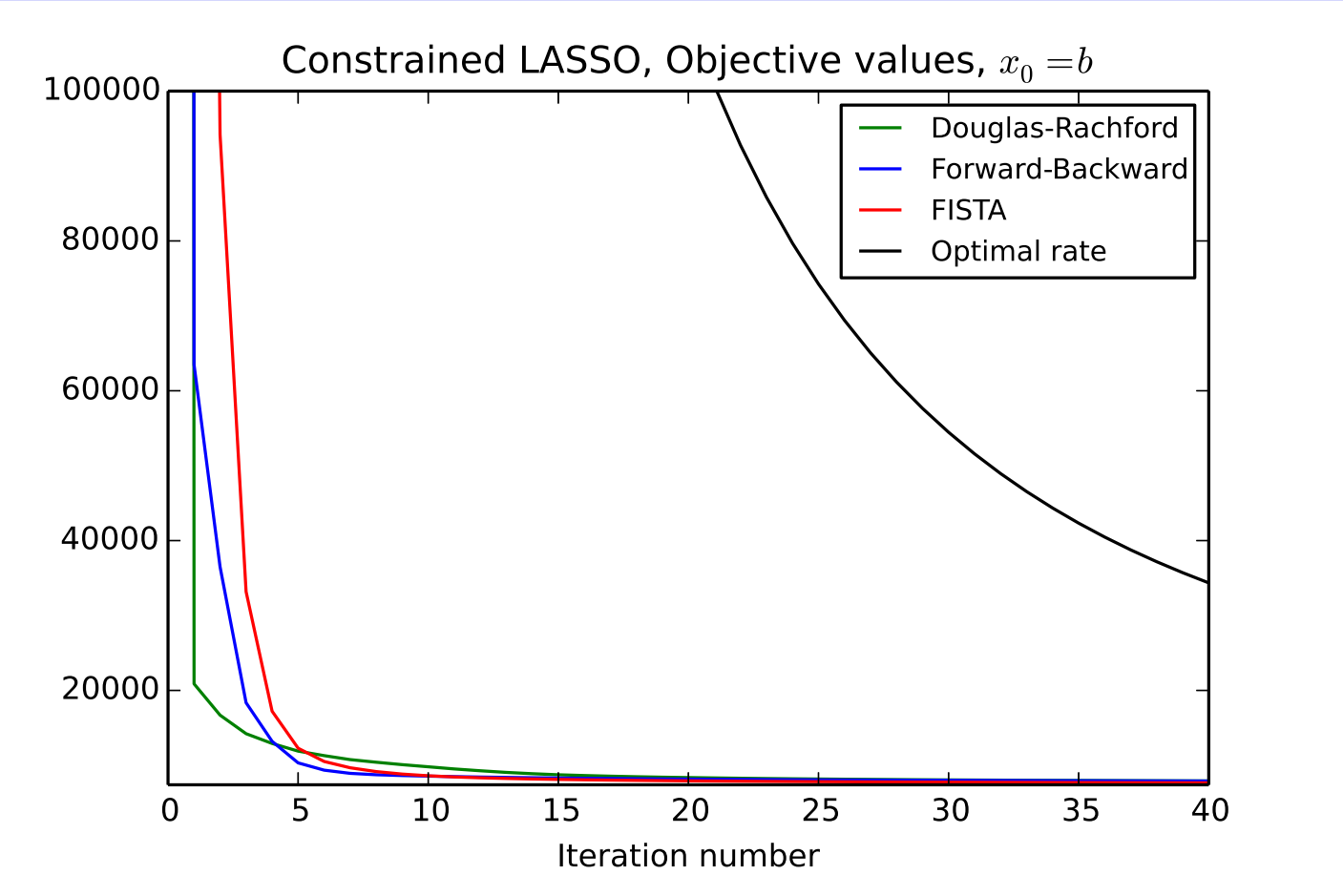
# Further properties of forward-backward splitting

- Solution set: $Z = \operatorname{Argmin} f + g$
- $x_{n+1} = x_n + \lambda_n \Big( \operatorname{prox}_{\gamma_n f}\big(x_n - \gamma_n \nabla g(x_n)\big) - x_n \Big)$, $\varepsilon \leqslant \gamma_n \leqslant (2-\varepsilon)\beta$
- $(\forall n \in \mathbb{N})(\forall z \in Z) \quad \|x_{n+1} - z\| \leqslant \|x_n - z\|$: Fejér monotonicity
- $(\forall n \in \mathbb{N}) \quad (f+g)(x_{n+1}) \leqslant (f+g)(x_n)$
- Convergence is only weak
- Even in the finite dimensional or the linear case, no (upper bound on the worst) rate of convergence of $\|x_n - x_\infty\|$ exists
- $\sum_{n\in\mathbb{N}} \big|(f+g)(x_n) - \inf(f+g)(\mathcal{H})\big|^2 < +\infty$
- If $\sum_{n\in\mathbb{N}}(1-\lambda_n) < +\infty$, $(f+g)(x_n) - \inf(f+g)(\mathcal{H}) = o(1/n)$
  (PLC, Salzo, Villa, arxiv, 2015)
- Worst-case behavior rates should be interpreted with caution and not lead to unrealistic expectations, especially when they concern only function values and not proximity to $Z$...

# On minimizing sequences

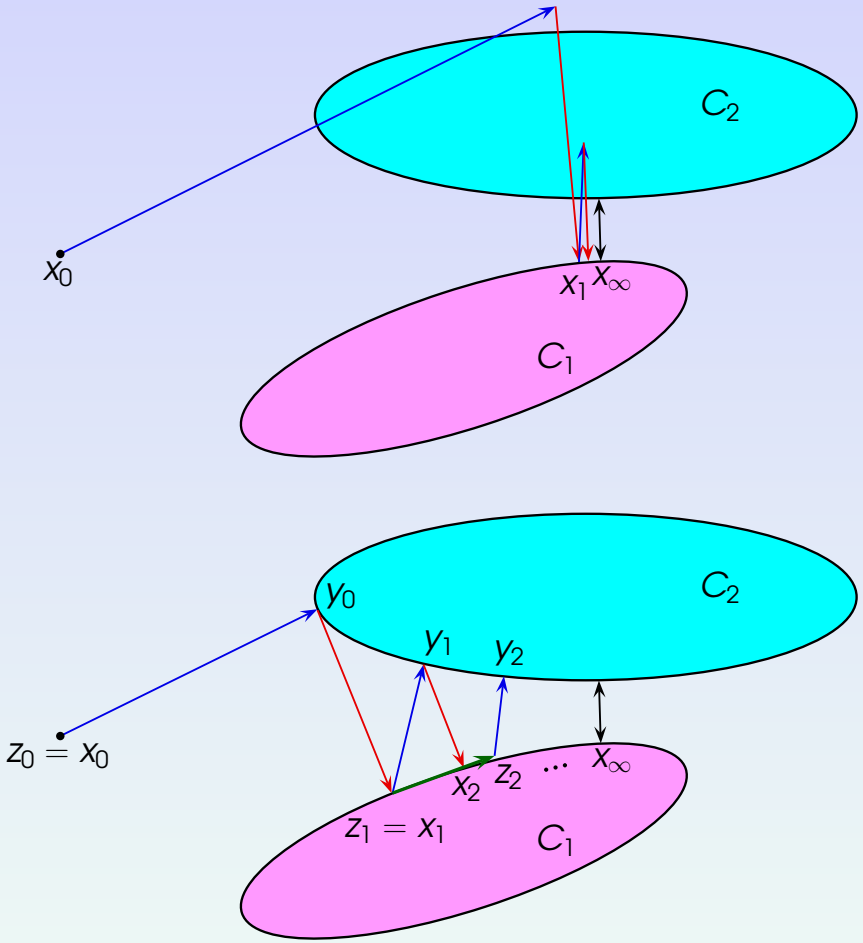
- Let $\Phi \in \Gamma_0(\mathcal{H})$, $Z = \operatorname{Argmin} \Phi \neq \emptyset$ the solution set
- Minimizing sequences may have little to do with actually approaching a point in $Z$ as we can have (even in $\mathbb{R}^2$):
  - $\Phi(x_n) \to \inf \Phi(\mathcal{H})$ and $(\forall z \in Z)\ \|x_n - z\| \geqslant 1$
  - $\Phi(x_n) \to \inf \Phi(\mathcal{H})$ and $(\forall z \in Z)\ \|x_n - z\| \to +\infty$
  - ... and vice versa $\Phi(x_n) \equiv +\infty$ and $x_n \to z \in Z$
- The whole area of metric regularity addresses such issues

# Theoretical rates vs actual behavior




(Image restoration example; PLC&Glaudin)

# Alternating projection method




Top: Forward-backward; Bottom: FISTA (PLC&Pesquet, 2011)

# Proximal splitting methods in convex optimization

- $f \in \Gamma_0(\mathcal{H})$, $\varphi_k \in \Gamma_0(\mathcal{G}_k)$, $\ell_k \in \Gamma_0(\mathcal{G}_k)$ strongly convex, $L_k \colon \mathcal{H} \to \mathcal{G}_k$ linear bounded, $\|L_k\| = 1$, $h \colon \mathcal{H} \to \mathbb{R}$ convex and smooth:

$$\operatorname*{minimize}_{x \in \mathcal{H}} \; f(x) + \sum_{k=1}^{p} (\varphi_k \,\square\, \ell_k)(L_k x - r_k) + h(x)$$

  where: $\varphi_k \,\square\, \ell_k \colon x \mapsto \inf_{y \in \mathcal{H}} \big(\varphi_k(y) + \ell_k(x - y)\big)$

- Example: multiview total variation image recovery from observations $r_k = L_k \overline{x} + w_k$:

$$\operatorname*{minimize}_{x \in \mathcal{H}} \; \sum_{k \in \mathbb{N}} \phi_k(\langle x \mid e_k \rangle) + \sum_{k=1}^{p-1} \alpha_k \underbrace{d_{C_k}}_{\iota_C \,\square\, \|\cdot\|} (L_k x - r_k) + \beta \|\nabla x\|_{1,2}$$

- A splitting algorithm activates each function and each linear operator individually

# A few notions on monotone operators

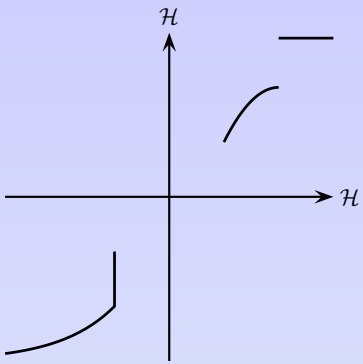
- $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ is maximally monotone: for every $(x, x^*) \in \mathcal{H}^2$,

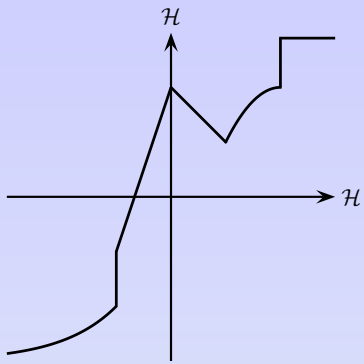
$$(x, x^*) \in \operatorname{gra} A \quad \Leftrightarrow \quad (\forall (y, y^*) \in \operatorname{gra} A) \quad \langle x - y \mid x^* - y^* \rangle \geqslant 0$$

- The resolvent of $A$, $J_A = (\mathrm{Id} + A)^{-1}\colon \mathcal{H} \to \mathcal{H}$, is firmly nonexpansive and $\operatorname{Fix} J_A = \operatorname{zer} A = \big\{x \in \mathcal{H} \mid 0 \in Ax\big\}$
- Minty's parametrization:

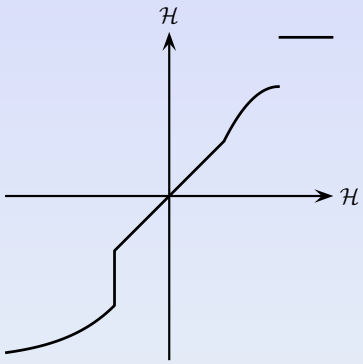
$$(\forall x \in \mathcal{H}) \quad (J_A x, x - J_A x) = (J_A x, J_{A^{-1}} x) \in \operatorname{gra} A$$

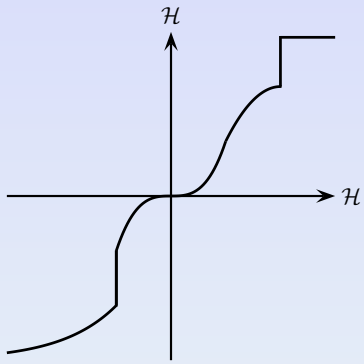
- The problem $0 \in Ax$ covers convex optimization, Nash equilibria, variational inequalities, complementarity problems, saddle point problems, feasibility problems, fixed point problems, PDEs, etc
- Example (Moreau's theorem): $f \in \Gamma_0(\mathcal{H})$, $A = \partial f$. Then $\operatorname{zer} A = \operatorname{Argmin} f$ and $J_A = \operatorname{prox}_f$



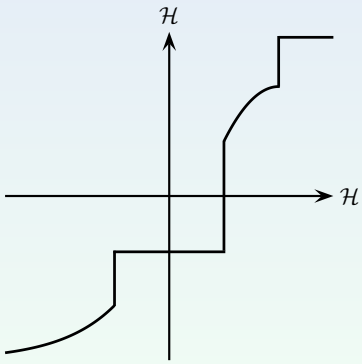



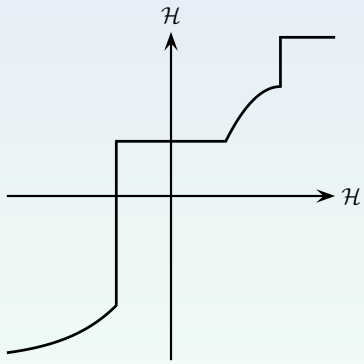

monotone, not monotone




monotone, max. monotone




max. monotone, max. monotone

# Proximal splitting methods in convex optimization

- $A = \partial f$, $C = \nabla h$, $B_k = \partial g_k$, and $D_k = \partial \ell_k$
- $\mathcal{K} = \mathcal{H} \oplus \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_p$
- $\boldsymbol{M} \colon \mathcal{K} \to 2^{\mathcal{K}} \colon (x, v_1, \ldots, v_p) \mapsto$
  $(-z + Ax) \times (r_1 + B_1^{-1} v_1) \times \cdots \times (r_p + B_p^{-1} v_p)$
- $\boldsymbol{Q} \colon \mathcal{K} \to \mathcal{K} \colon (x, v_1, \ldots, v_p) \mapsto$
  $\big(Cx + \sum_{i=1}^{p} L_k^* v_k, -L_1 x + D_1^{-1} v_1, \ldots, -L_p x + D_p^{-1} v_p\big)$
- $\boldsymbol{M}$ and $\boldsymbol{Q}$ are maximally monotone, $\boldsymbol{Q}$ is Lipschitzian, the zeros of $\boldsymbol{M} + \boldsymbol{Q}$ are primal-dual solutions
- Solve $\boldsymbol{0} \in \boldsymbol{M}\boldsymbol{x} + \boldsymbol{Q}\boldsymbol{x}$, where $\boldsymbol{x} = (x, v_1, \ldots, v_p)$ via Tseng's forward-backward-forward splitting algorithm

$$\left\lfloor \begin{aligned} \boldsymbol{y}_n &= \boldsymbol{x}_n - \boldsymbol{Q}\boldsymbol{x}_n \\ \boldsymbol{p}_n &= (\mathrm{Id} + \boldsymbol{M})^{-1} \boldsymbol{y}_n \\ \boldsymbol{q}_n &= \boldsymbol{p}_n - \boldsymbol{Q}\boldsymbol{p}_n \\ \boldsymbol{x}_{n+1} &= \boldsymbol{x}_n - \boldsymbol{y}_n + \boldsymbol{q}_n \end{aligned} \right.$$

in $\mathcal{K}$ to get...

# Proximal splitting methods in convex optimization

- Algorithm:

$$
\begin{array}{l}
\text{for } n = 0, 1, \ldots \\
\quad y_{1,n} = x_n - \big(\nabla h(x_n) + \sum_{k=1}^{m} L_k^* v_{k,n}\big) \\
\quad p_{1,n} = \operatorname{prox}_f y_{1,n} \\
\quad \text{For } \; k = 1, \ldots, p \\
\quad\quad y_{2,k,n} = v_{k,n} + (L_k x_n - \nabla \ell_k^*(v_{k,n})) \\
\quad\quad p_{2,k,n} = \operatorname{prox}_{g_k^*}(y_{2,k,n} - r_k) \\
\quad\quad q_{2,k,n} = p_{2,k,n} + (L_k p_{1,n} - \nabla \ell_k^*(p_{2,k,n})) \\
\quad\quad v_{k,n+1} = v_{k,n} - y_{2,k,n} + q_{2,k,n} \\
\quad q_{1,n} = p_{1,n} - \big(\nabla h(p_{1,n}) + \sum_{k=1}^{m} L_k^* p_{2,k,n}\big) \\
\quad x_{n+1} = x_n - y_{1,n} + q_{1,n}
\end{array}
$$

- $(x_n)_{n\in\mathbb{N}}$ converges weakly to a solution
  - PLC, Systems of structured monotone inclusions: Duality, algorithms, and applications, *SIAM J. Optim.*, vol. 23, 2013

# Proximal splitting methods in convex optimization

- Algorithm:

$$
\begin{array}{l}
\text{for } n = 0, 1, \ldots \\
\quad y_{1,n} = x_n - \left(\nabla h(x_n) + \sum_{k=1}^{m} L_k^* v_{k,n}\right) \\
\quad p_{1,n} = \operatorname{prox}_f y_{1,n} \\
\quad \text{For } k = 1, \ldots, p \\
\quad\quad y_{2,k,n} = v_{k,n} + (L_k x_n - \nabla \ell_k^*(v_{k,n})) \\
\quad\quad p_{2,k,n} = \operatorname{prox}_{g_k^*}(y_{2,k,n} - r_k) \\
\quad\quad q_{2,k,n} = p_{2,k,n} + (L_k p_{1,n} - \nabla \ell_k^*(p_{2,k,n})) \\
\quad\quad v_{k,n+1} = v_{k,n} - y_{2,k,n} + q_{2,k,n} \\
\quad q_{1,n} = p_{1,n} - \left(\nabla h(p_{1,n}) + \sum_{k=1}^{m} L_k^* p_{2,k,n}\right) \\
\quad x_{n+1} = x_n - y_{1,n} + q_{1,n}
\end{array}
$$

- $(x_n)_{n\in\mathbb{N}}$ converges weakly to a solution
  - PLC, Systems of structured monotone inclusions: Duality, algorithms, and applications, *SIAM J. Optim.*, vol. 23, 2013
- This construction, as most advanced recent splitting methods for optimization, involve monotone operators which are not subdifferentials: **monotone operator theory is needed, even in the limited setting of optimization**

# Asynchronous, block-iterative splitting

- For every $i \in I$ (finite), $\mathcal{H}_i$ a Hilbert space, $A_i \colon \mathcal{H}_i \to 2^{\mathcal{H}_i}$ maximally monotone, $z_i^* \in \mathcal{H}_i$
- For every $k \in K$ (finite), $\mathcal{G}_k$ a Hilbert space, $B_k \colon \mathcal{G}_k \to 2^{\mathcal{G}_k}$ maximally monotone, $r_k \in \mathcal{G}_k$, $L_{ki} \colon \mathcal{H}_i \to \mathcal{G}_k$ linear&bounded
- Initial problem: find $(\overline{x}_i)_{i \in I} \in \boldsymbol{\mathcal{H}} = \bigoplus_{i \in I} \mathcal{H}_i$ such that

$$(\forall i \in I) \quad z_i^* \in A_i \overline{x}_i + \sum_{k \in K} L_{ki}^* \Bigg( B_k \Bigg( \sum_{j \in I} L_{kj} \overline{x}_j - r_k \Bigg) \Bigg)$$

# Asynchronous, block-iterative splitting

- For every $i \in I$ (finite), $\mathcal{H}_i$ a Hilbert space, $A_i \colon \mathcal{H}_i \to 2^{\mathcal{H}_i}$ maximally monotone, $z_i^* \in \mathcal{H}_i$
- For every $k \in K$ (finite), $\mathcal{G}_k$ a Hilbert space, $B_k \colon \mathcal{G}_k \to 2^{\mathcal{G}_k}$ maximally monotone, $r_k \in \mathcal{G}_k$, $L_{ki} \colon \mathcal{H}_i \to \mathcal{G}_k$ linear&bounded
- Initial problem: find $(\overline{x}_i)_{i\in I} \in \boldsymbol{\mathcal{H}} = \bigoplus_{i\in I} \mathcal{H}_i$ such that

$$(\forall i \in I) \quad z_i^* \in A_i\overline{x}_i + \sum_{k\in K} L_{ki}^*\left(B_k\left(\sum_{j\in I} L_{kj}\overline{x}_j - r_k\right)\right)$$

- Dual problem: find $(\overline{v}_k^*)_{k\in K} \in \boldsymbol{\mathcal{G}} = \bigoplus_{k\in K} \mathcal{G}_k$ such that

$$(\forall k \in K) \quad -r_k \in -\sum_{i\in I} L_{ki}\left(A_i^{-1}\left(z_i^* - \sum_{l\in K} L_{li}^*\overline{v}_l^*\right)\right) + B_k^{-1}\overline{v}_k^*$$

# Asynchronous, block-iterative splitting

- **Solutions set:** the associated Kuhn-Tucker set

$$\boldsymbol{Z} = \left\{ \left( (\overline{x}_i)_{i\in I}, (\overline{v}_k^*)_{k\in K} \right) \;\middle|\; \overline{x}_i \in \mathcal{H}_i \;\text{ and }\; z_i^* - \sum_{k\in K} L_{ki}^* \overline{v}_k^* \in A_i \overline{x}_i, \right.$$
$$\left. \overline{v}_k^* \in \mathcal{G}_k \;\text{ and }\; \sum_{i\in I} L_{ki}\overline{x}_i - r_k \in B_k^{-1}\overline{v}_k^* \right\}$$

- $\boldsymbol{Z}$ is a closed convex set
- The projection of $\boldsymbol{Z}$ onto $\mathcal{H}$ is the set $\boldsymbol{F}$ of primal solutions
- The projection of $\boldsymbol{Z}$ onto $\mathcal{G}$ is the set $\boldsymbol{F}^*$ of dual solutions

# With proper CQ, this framework includes..

- Let $\boldsymbol{F}$ be the set of solutions to the problem

$$\underset{(x_i)_{i\in I}\in\boldsymbol{\mathcal{H}}}{\text{minimize}} \quad \sum_{i\in I}\big(f_i(x_i)-\langle x_i \mid z_i^*\rangle\big)+\sum_{k\in K} g_k\bigg(\sum_{i\in I} L_{ki}x_i - r_k\bigg)$$

  where $f_i \in \Gamma_0(\mathcal{H}_i)$, $g_k \in \Gamma_0(\mathcal{G}_k)$, $L_{ki}\colon \mathcal{H}_i \to \mathcal{G}_k$ linear&bounded

- Let $\boldsymbol{F}^*$ be the set of solutions to the dual problem

$$\underset{(v_k^*)_{k\in K}\in\bigoplus_{k\in K}\mathcal{G}_k}{\text{minimize}} \quad \sum_{i\in I} f_i^*\bigg(z_i^* - \sum_{k\in K} L_{ki}^* v_k^*\bigg)+\sum_{k\in K}\big(g_k^*(v_k^*)+\langle v_k^* \mid r_k\rangle\big)$$

- Associated Kuhn-Tucker set

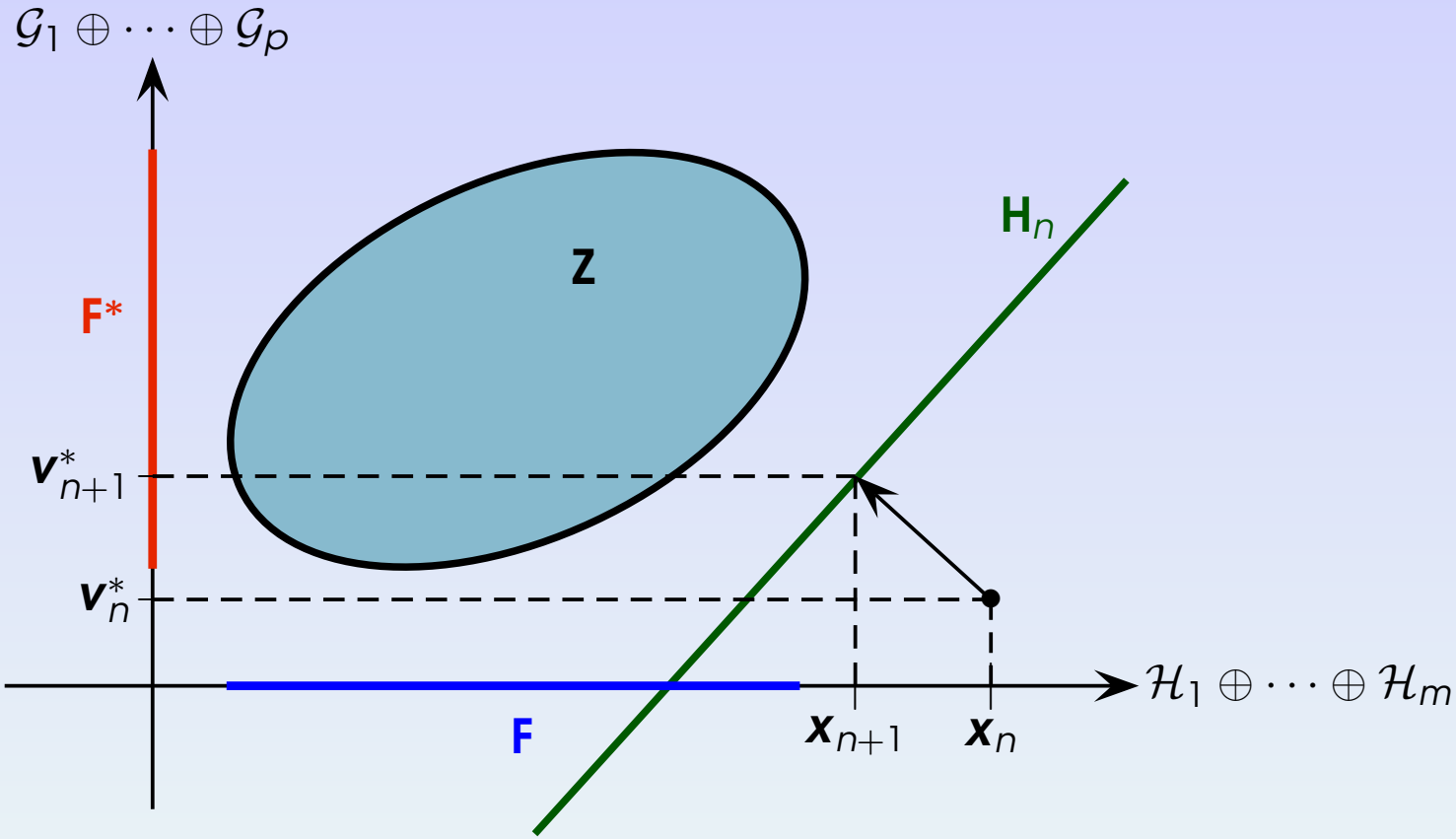
$$\boldsymbol{Z} = \Bigg\{\big((\overline{x}_i)_{i\in I},(\overline{v}_k^*)_{k\in K}\big)\;\Bigg|\; \overline{x}_i\in\mathcal{H}_i \;\text{ and }\; z_i^*-\sum_{k\in K}L_{ki}^*\overline{v}_k^*\in\partial f_i(\overline{x}_i),$$
$$\overline{v}_k^*\in\mathcal{G}_k \;\text{ and }\; \sum_{i\in I}L_{ki}\overline{x}_i - r_k\in\partial g_k^*(\overline{v}_k^*)\Bigg\}$$

# Some limitations of the state of the art

We present a new framework that circumvents simultaneously the limitations of current methods, which require:

- inversions of linear operators or knowledge of bounds on norms of all the $L_{ki}$
- activation of the resolvents of all the monotone operators: impossible in huge-scale problems
- synchronicity: all proximity operator evaluations must be computed and used during the current iteration
- Converge only weakly in general

# Asynchronous block-iterative proximal splitting I




- Choose suitable points in the graphs of $(A_i)_{i\in I}$ and $(B_k)_{k\in K}$ to construct a half-space $\mathbf{H}_n$ containing $\mathbf{Z}$
- Algorithm: $(x_{n+1}, v^*_{n+1}) = P_{\mathbf{H}_n}(x_n, v^*_n) \rightharpoonup (x, v^*) \in \mathbf{Z} \subset \mathbf{F} \times \mathbf{F}^*$

# Main novelties

- **Block iterations:** At iteration $n$, we require calculation of new points in the graphs of only some the operators $(A_i)_{i\in I_n\subset I}$ and $(B_k)_{k\in K_n\subset K}$. The control sequences $(I_n)_{n\in\mathbb{N}}$ and $(K_n)_{n\in\mathbb{N}}$ dictate how frequently the various operators are used.
- **Asynchronicity:** A new point $(a_{i,n}, a^*_{i,n}) \in \operatorname{gra} A_i$ being incorporated into the calculations at iteration $n$ may be based on data $x_{i,c_i(n)}$ and $(v^*_{k,c_i(n)})_{k\in K}$ available at some possibly earlier iteration $c_i(n) \leqslant n$. Therefore, the calculation of $(a_{i,n}, a^*_{i,n})$ could have been initiated at iteration $c_i(n)$, with its results becoming available only at iteration $n$. Likewise, for $(b_{k,n}, b^*_{k,n}) \in \operatorname{gra} B_k$.

Also:

- No knowledge of the $\|L_{ki}\|$s is required
- No linear operator inversion is required
- No bounds required on the proximal parameters

# Asynchronous block-iterative proximal splitting I

for $n = 0, 1, \ldots$

  for every $i \in I_n$

    $l^*_{i,n} = \sum_{k \in K} L^*_{ki} v^*_{k,c_i(n)}$

    $(a_{i,n}, a^*_{i,n}) = \left(J_{\gamma_{i,c_i(n)} A_i}\left(x_{i,c_i(n)} + \gamma_{i,c_i(n)}(z_i - l^*_{i,n})\right), \gamma^{-1}_{i,c_i(n)}(x_{i,c_i(n)} - a_{i,n}) - l^*_{i,n}\right)$

  for every $i \in I \smallsetminus I_n$

    $(a_{i,n}, a^*_{i,n}) = (a_{i,n-1}, a^*_{i,n-1})$

  for every $k \in K_n$

    $l_{k,n} = \sum_{i \in I} L_{ki} x_{i,d_k(n)}$

    $(b_{k,n}, b^*_{k,n}) = \left(r_k + J_{\mu_{k,d_k(n)} B_k}\left(l_{k,n} + \mu_{k,d_k(n)} v^*_{k,d_k(n)} - r_k\right), v^*_{k,d_k(n)} + \mu^{-1}_{k,d_k(n)}(l_{k,n} - b_{k,n})\right)$

  for every $k \in K \smallsetminus K_n$

    $(b_{k,n}, b^*_{k,n}) = (b_{k,n-1}, b^*_{k,n-1})$

  $((t^*_{i,n})_{i \in I}, (t_{k,n})_{k \in K}) = ((a^*_{i,n} + \sum_{k \in K} L^*_{ki} b^*_{k,n})_{i \in I}, (b_{k,n} - \sum_{i \in I} L_{ki} a_{i,n})_{k \in K})$

  $\tau_n = \sum_{i \in I} \|t^*_{i,n}\|^2 + \sum_{k \in K} \|t_{k,n}\|^2$

  if $\tau_n > 0$

    $\theta_n = \dfrac{\lambda_n}{\tau_n} \max\left\{0, \sum_{i \in I}\left(\langle x_{i,n} \mid t^*_{i,n}\rangle - \langle a_{i,n} \mid a^*_{i,n}\rangle\right) + \sum_{k \in K}\left(\langle t_{k,n} \mid v^*_{k,n}\rangle - \langle b_{k,n} \mid b^*_{k,n}\rangle\right)\right\}$

  else $\theta_n = 0$

  for every $i \in I$

    $x_{i,n+1} = x_{i,n} - \theta_n t^*_{i,n}$

  for every $k \in K$

    $v^*_{k,n+1} = v^*_{k,n} - \theta_n t_{k,n}$

# Convergence

- $(I_n)_{n\in\mathbb{N}}$ is a sequence of nonempty subsets of $I$, and $(K_n)_{n\in\mathbb{N}}$ is a sequence of nonempty subsets of $K$ such that $I_0 = I$, $K_0 = K$, and

$$(\forall n \in \mathbb{N}) \quad \left( \bigcup_{j=n}^{n+M-1} I_j = I \quad \text{and} \quad \bigcup_{j=n}^{n+M-1} K_j = K \right). \tag{1}$$

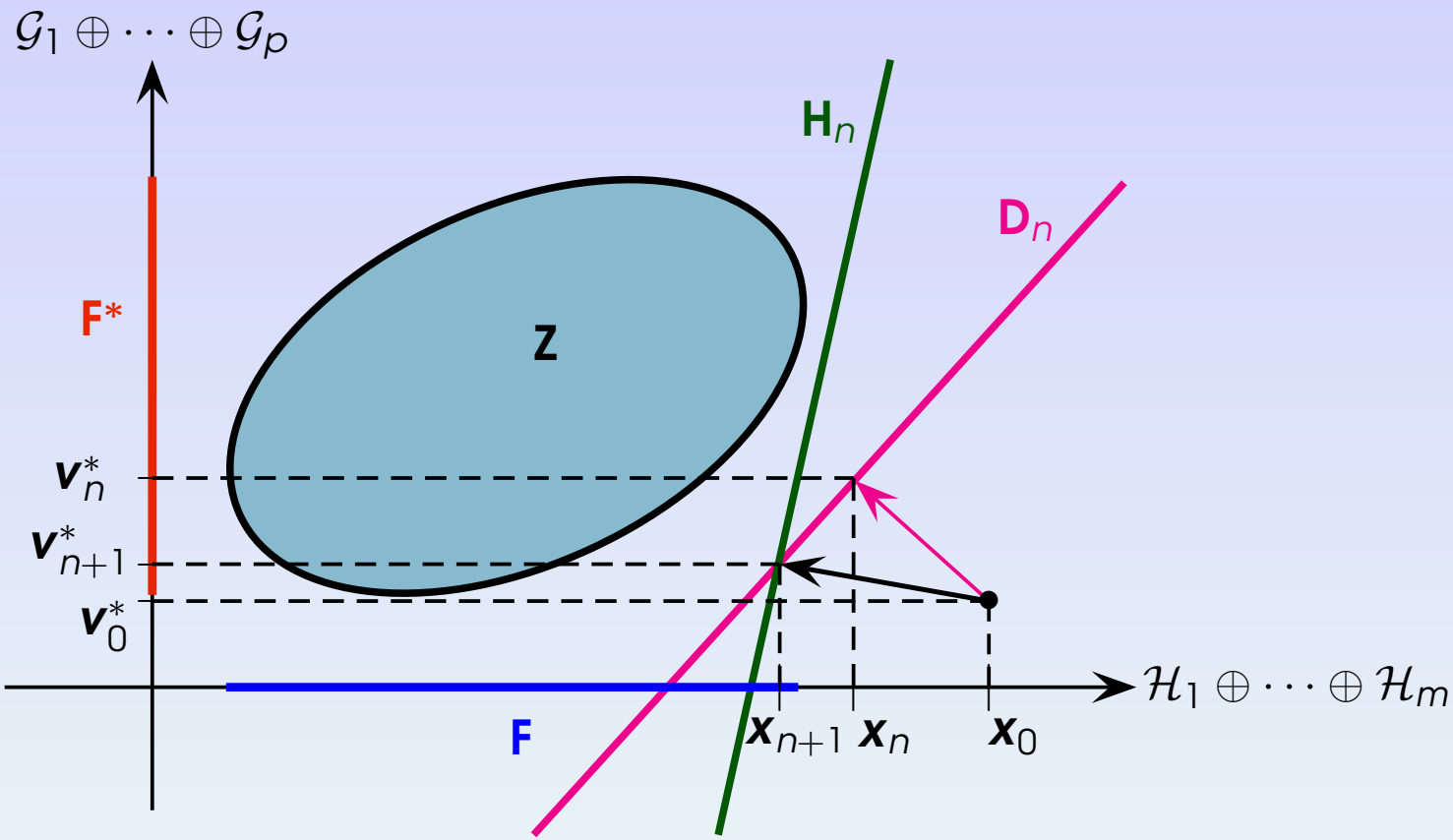
- $(c_i(n))_{n\in\mathbb{N}}$ and $(d_k(n))_{n\in\mathbb{N}}$ are sequences in $\mathbb{N}$ such that

$$(\forall i \in I) \;\; n - D \leqslant c_i(n) \leqslant n \quad \text{and} \quad (\forall k \in K) \;\; n - D \leqslant d_k(n) \leqslant n$$

- $\varepsilon \in ]0, 1[$ and $(\gamma_{i,n})_{n\in\mathbb{N}}$ and $(\mu_{k,n})_{n\in\mathbb{N}}$ are sequences in $[\varepsilon, 1/\varepsilon]$.

Set $x_n = (x_{i,n})_{i\in I}$ and $v_n^* = (v_{k,n}^*)_{k\in K}$. Then $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point $\overline{x} \in \boldsymbol{F}$, $(v_n^*)_{n\in\mathbb{N}}$ converges weakly to a point $\overline{v} \in \boldsymbol{F}^*$, and $(\overline{x}, \overline{v}^*) \in \boldsymbol{Z}$.

# Asynchronous block-iterative proximal splitting II




- Construct $\mathbf{H}_n$ as before
- The half-space $\mathbf{D}_n$ satisfies $(\boldsymbol{x}_n, \boldsymbol{v}_n^*) = P_{\mathbf{D}_n}(\boldsymbol{x}_0, \boldsymbol{v}_0^*)$
- Algorithm: $(\boldsymbol{x}_{n+1}, \boldsymbol{v}_{n+1}^*) = P_{\mathbf{H}_n \cap \mathbf{D}_n}(\boldsymbol{x}_0, \boldsymbol{v}_0^*) \to P_{\mathbf{Z}}(\boldsymbol{x}_0, \boldsymbol{v}_0^*) \in \mathbf{F} \times \mathbf{F}^*$

# References


- PLC and J. Eckstein, Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions, arxiv, 2015
- PLC and J.-C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, *SIAM J. Optim.*, vol. 25, 2015
- H. H. Bauschke and PLC, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011. (2nd ed., Spring 2016)
- H. Attouch, L. M. Briceño-Arias, PLC, A strongly convergent primal-dual method for nonoverlapping domain decomposition, *Numer. Math.*, published online 2015-07-10.