Tight Rates and Equivalence Results of Operator Splitting Schemes

Wotao Yin (UCLA Math)

Workshop on Optimization for Modern Computing

Joint w **Damek Davis** and **Ming Yan**
UCLA CAM 14-51, 14-58, and 14-59

Operator splitting methods

- They are methods for solving problems like

$$\begin{aligned} &\text{find } x \in \mathcal{C}_1 \cap \mathcal{C}_2, \\ &\underset{x}{\text{minimize}}\, f(x) + g(x), \\ &\underset{x,y}{\text{minimize}}\, f(x) + g(y), \quad \text{subject to } Ax + By = b, \end{aligned}$$

  by iteratively performing simple operations.

- **Algorithms**: alternating projection, forward-backward splitting (FBS), Douglas-Rachford splitting (DRS), Peaceman-Rachford splitting (PRS), ADMM, etc.

- Most of them can be written as $x^{k+1} \leftarrow T(x^k)$, where T satisfies
  - $x = T(x) \;\Leftrightarrow\; x$ is a solution.
  - T is nonexpansive. In particular, $\|T(x^k) - x^*\|^2 \leq \|x^k - x^*\|^2$.
  - T is composed of $I - \gamma\nabla h$, $\mathbf{prox}_{\gamma h}$, and $\mathbf{refl}_{\gamma h}$.

This talk

- Reviews some examples of $\mathbf{prox}$ and splitting algorithms.
- Establishes new convergence results, many of which are tight.
- Argues that convergence of DRS, PRS, and ADMM automatically improves upon better regularity properties.
- DRS, PRS, and ADMM are "self-dual" primal-dual algorithms.

Proximal operator

Unlike those with explicit formulas, $\mathbf{prox}$ method is an optimization problem

$$\mathbf{prox}_{\lambda f}(v) := \arg\min_x f(x) + \frac{1}{2\lambda}\|x - v\|^2$$

Examples:

- $f = \iota_{\mathcal{C}} \Rightarrow$ **Euclidean projection** $\mathbf{prox}_f(v) = \mathrm{Proj}_{\mathcal{C}}(v)$
- closed-form formulas for norms and many separable functions

Relation to resolvent:

- $\mathbf{prox}_{\lambda f} = (I + \lambda \partial f)^{-1}$, where f is proper closed convex
- S is maximally monotone $\Rightarrow$ $(I + \lambda S)^{-1}$ is a point-to-point mapping
- proximal-point algorithm (PPA): $x^{k+1} = (I + \lambda S)^{-1}(x^k)$

Properties of $\mathbf{prox}_{\lambda f}$

- **Fixed point is optimal**. $f(x^*) = \min_x f(x) \;\Leftrightarrow\; x^* = \mathbf{prox}_{\gamma f}(x^*)$
- $T = \mathbf{prox}_{\lambda f}$ is **firmly nonexpansive**, i.e.,

$$\|T(x) - T(y)\|^2 \leq \|x - y\|^2 - \|(x - T(x)) - (y - T(y))\|^2$$

  $\Rightarrow$ weak convergence in Hilbert space, and the rate of fixed-point residual

- Interpretation: **backward Euler / implicit gradient**

$$\begin{aligned} x^{k+1} = \mathbf{prox}_{\lambda f}(x^k) \quad &\Leftrightarrow \quad x^{k+1} = (I + \lambda \partial f)^{-1}(x^k) \\ &\Leftrightarrow \quad x^k \in x^{k+1} + \lambda \partial f(x^{k+1}) \\ &\Leftrightarrow \quad \boxed{x^{k+1} = x^k - \lambda \widetilde{\nabla} f(x^{k+1})} \end{aligned}$$

  (We use $\widetilde{\nabla} f$ for the subgradient of f, uniquely determined by $\mathbf{prox}_{\lambda f}$)

- Moreau decomposition:

$$x = \mathbf{prox}_f(x) + \mathbf{prox}_{f^*}(x)$$

  For linear subspace S and $f = \iota_{\mathcal{S}}$, reduces to $x = \mathrm{Proj}_{\mathcal{S}}(x) + \mathrm{Proj}_{\mathcal{S}^\perp}(x)$

Forward-backward splitting (FBS)

$$\underset{x}{\text{minimize}}\ r(x) + f(x)$$

- Suppose $A = \partial r$ and $B = \nabla f$ (f is differentiable). Optimality condition

$$0 \in (\partial r + \nabla f)x^*$$

has the operator form

$$\begin{aligned} 0 \in (A+B)x^* &\iff (I - \gamma B)x^* \in (I + \gamma A)x^* \\ &\iff \underbrace{(I+\gamma A)^{-1}}_{\text{backward}} \underbrace{(I-\gamma B)}_{\text{forward}} x^* = x^*. \end{aligned}$$

- Prox-gradient (prox-linear) iteration

$$x^{k+1} = \mathbf{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k)).$$

- (Sub)gradient form

$$x^{k+1} = x^k - \gamma \widetilde{\nabla} r(x^{k+1}) - \gamma \nabla f(x^k).$$

Reflection operator and averaged operator

- Definition:

$$\mathbf{refl}_f = 2\mathbf{prox}_f - I.$$

- Subgradient form:

$$x_f^k = \mathbf{prox}_f(z^k) = z^k - \widetilde{\nabla} f(x_f^k)$$
$$z^{k+1} = \mathbf{refl}_f(z^k) = z^k - 2\widetilde{\nabla} f(x_f^k).$$

- $\mathbf{refl}_f$ is **nonexpansive**, but **not firmly nonexpansive**.
- **Averaged operator:** weighted average of I and a nonexpansive T.

$$T_\lambda := (1-\lambda)I + \lambda T.$$

So, $\mathbf{prox}_f = (\mathbf{refl}_f)_{1/2}$.

- **Property:** $\lambda \in (0,1]$, $\forall\, x, y$

$$\|T_\lambda(x) - T_\lambda(y)\|^2 \leq \|x-y\|^2 - \frac{1-\lambda}{\lambda}\|(x - T_\lambda(x)) - (y - T_\lambda(y))\|^2$$

Peaceman-Rachford splitting (PRS)

$$\underset{z}{\text{minimize}} \; f(z) + g(z)$$

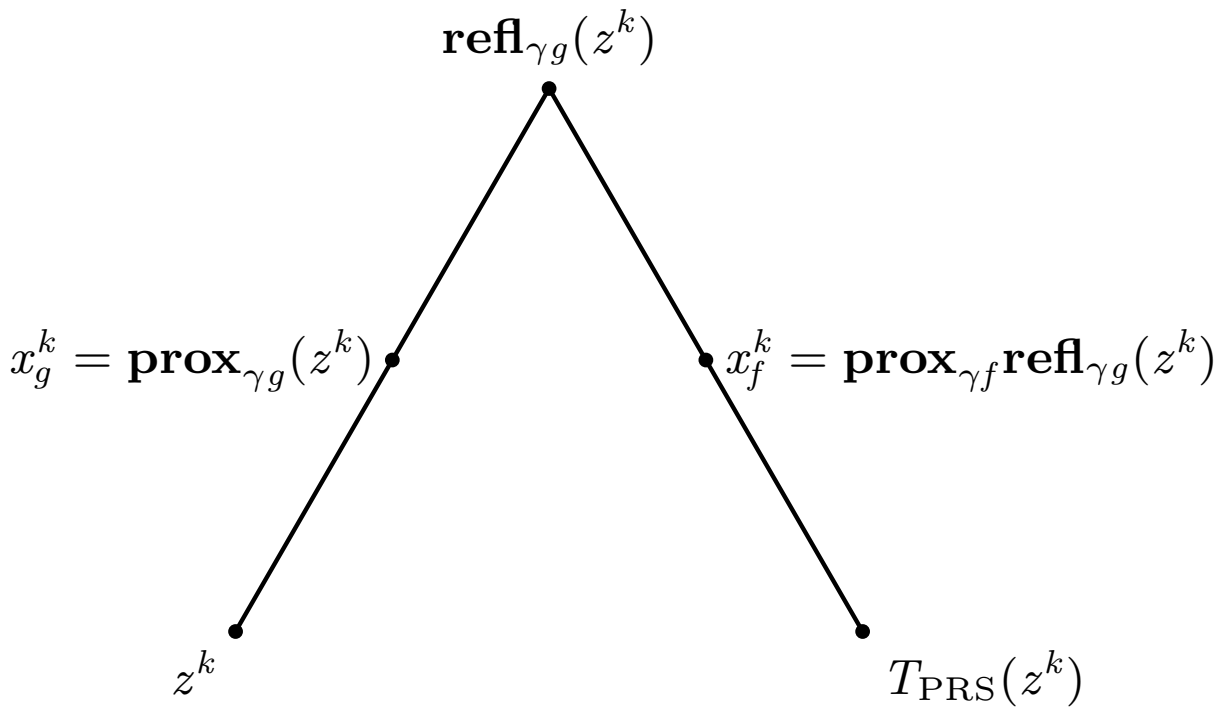
- Iteration:

$$z^{k+1} = T_{\text{PRS}}(z^k) := \mathbf{refl}_{\gamma f}\mathbf{refl}_{\gamma g}(z^k)$$

- Subgradient form:

$$z^{k+1} = z^k - 2\gamma\widetilde{\nabla} f(x_f^k) - 2\gamma\widetilde{\nabla} g(x_g^k).$$

- Diagram:



Peaceman-Rachford splitting (PRS)

$$\underset{z}{\text{minimize}}\ f(z) + g(z)$$

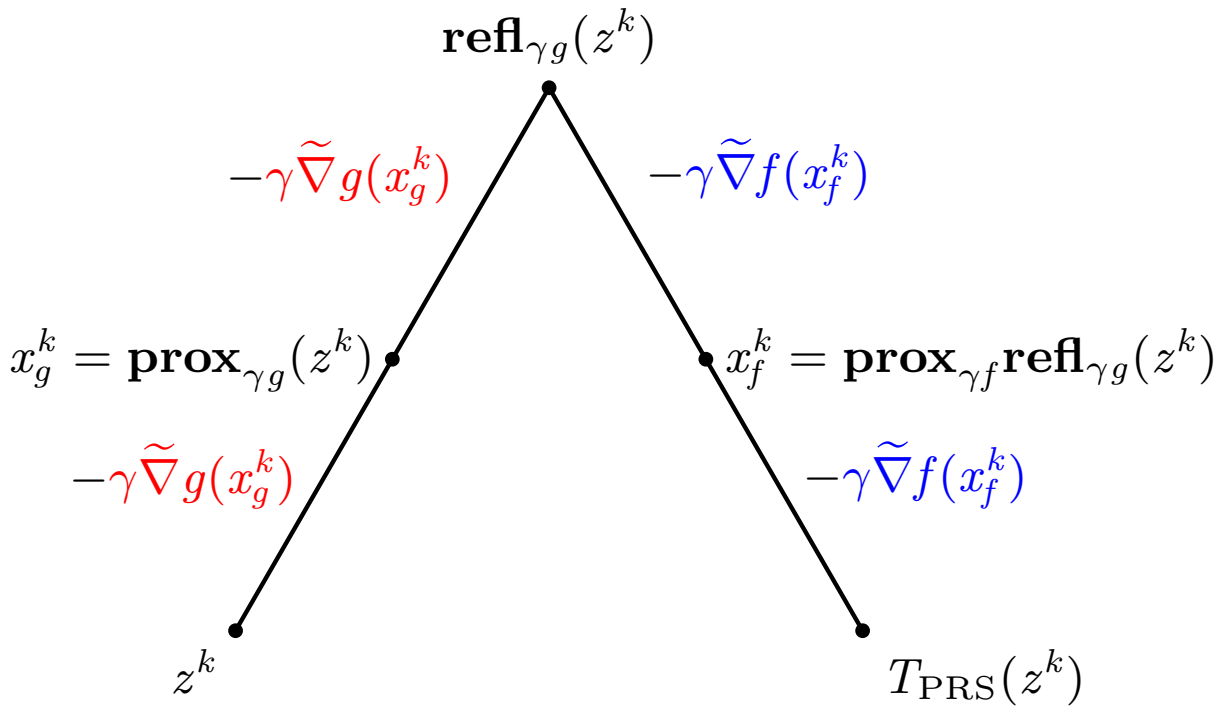
- Iteration:

$$z^{k+1} = T_{\text{PRS}}(z^k) := \mathbf{refl}_{\gamma f}\mathbf{refl}_{\gamma g}(z^k)$$

- Subgradient form:

$$z^{k+1} = z^k - 2\gamma\widetilde{\nabla} f(x_f^k) - 2\gamma\widetilde{\nabla} g(x_g^k).$$

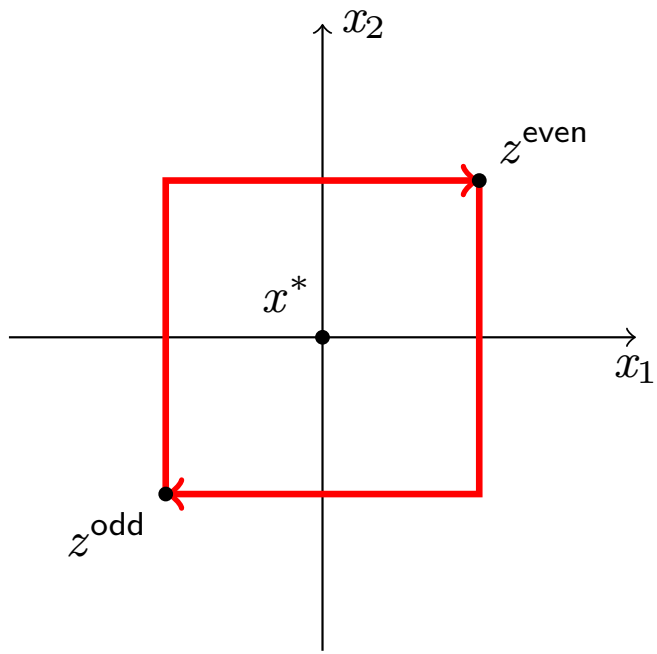
- Diagram:



- PRS iteration may not converge.

  Example: Let $\mathcal{C}_1 = x_1$–axis and $\mathcal{C}_2 = x_2$–axis.

$$\text{minimize} \;\; \iota_{\mathcal{C}_1}(x) + \iota_{\mathcal{C}_2}(x).$$



- Converges if one of the two functions is strongly convex
- Most well-known example of PRS: **method of alternating project**

Douglas-Rachford splitting (DRS) and relaxed PRS

- **Relaxed PRS:** fix z^0, $\gamma > 0$ and *relaxation* parameters $(\lambda_j)_{j\geq 0} \subseteq (0, 1]$

$$z^{k+1} = (T_{\mathrm{PRS}})_{\lambda_k}(z^k).$$

- **DRS:** Corresponds to $\lambda_k \equiv 1/2$.
- Always converges weakly, when a solution exists[1].
- $(T_{\mathrm{PRS}})_{\lambda_k} \equiv$ **"reflect, reflect, λ_k-average."**
- Fixed points $\not\equiv$ minimizers of $f + g$.
- $\mathbf{prox}_{\gamma g}(z^k) \rightharpoonup$ a minimizer (proved in 2011 in Banach space).[2]

[1] Eckstein and Bertsekas, On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators

[2] Svaiter, On weak convergence of the Douglas-Rachford method

First-order algorithms: subgradient forms

$$\underset{x}{\text{minimize}}\, f(x) + g(x)$$

- **(Sub)gradient descent:**

$$z^{k+1} = z^k - \gamma \widetilde{\nabla} f(z^k) - \gamma \widetilde{\nabla} g(z^k).$$

- **Proximal point algorithm (PPA):**

$$z^{k+1} = z^k - \gamma \widetilde{\nabla} f(z^{k+1}) - \gamma \widetilde{\nabla} g(z^{k+1}).$$

- **Forward backward splitting (FBS):**

$$z^{k+1} = z^k - \gamma \widetilde{\nabla} f(z^{k+1}) - \gamma \widetilde{\nabla} g(z^k).$$

- **Relaxed Peaceman-Rachford splitting (PRS):**

$$z^{k+1} = z^k - \lambda \left(\gamma \widetilde{\nabla} f(x_f^k) + \gamma \widetilde{\nabla} g(x_g^k) \right).$$

ADMM

$$\underset{x,y}{\text{minimize}} \;\; f(x) + g(y)$$

$$\text{subject to } Ax + By = b$$

- **ADMM iteration:**
  1. $x^{k+1} = \arg\min_x f(x) + (w^k)^T Ax + \frac{\gamma}{2}\|Ax + By^k - b\|^2$;
  2. $y^{k+1} = \arg\min_y g(y) + (w^k)^T By + \frac{\gamma}{2}\|Ax^{k+1} + By - b\|^2$
  3. $w^{k+1} = w^k + \gamma(Ax^{k+1} + By^{k+1} - b)$.
- Equivalent to **DRS** applied to the dual problem[3]:
  - Lagrangian: $\mathcal{L}(x, y; w) = \underbrace{f(x) + w^T Ax}_{\mathcal{L}_1(x;w)} + \underbrace{g(y) + w^T By - w^T b}_{\mathcal{L}_2(y;w)}$
  - Define:
    $$d_1(w) := -\min_x \mathcal{L}_1(x; w),$$
    $$d_2(w) := -\min_y \mathcal{L}_2(y; w).$$
  - Dual problem:
    $$\underset{w}{\text{minimize}}\; d_1(w) + d_2(w).$$

[3] Gabay 1983

Diagram of ADMM

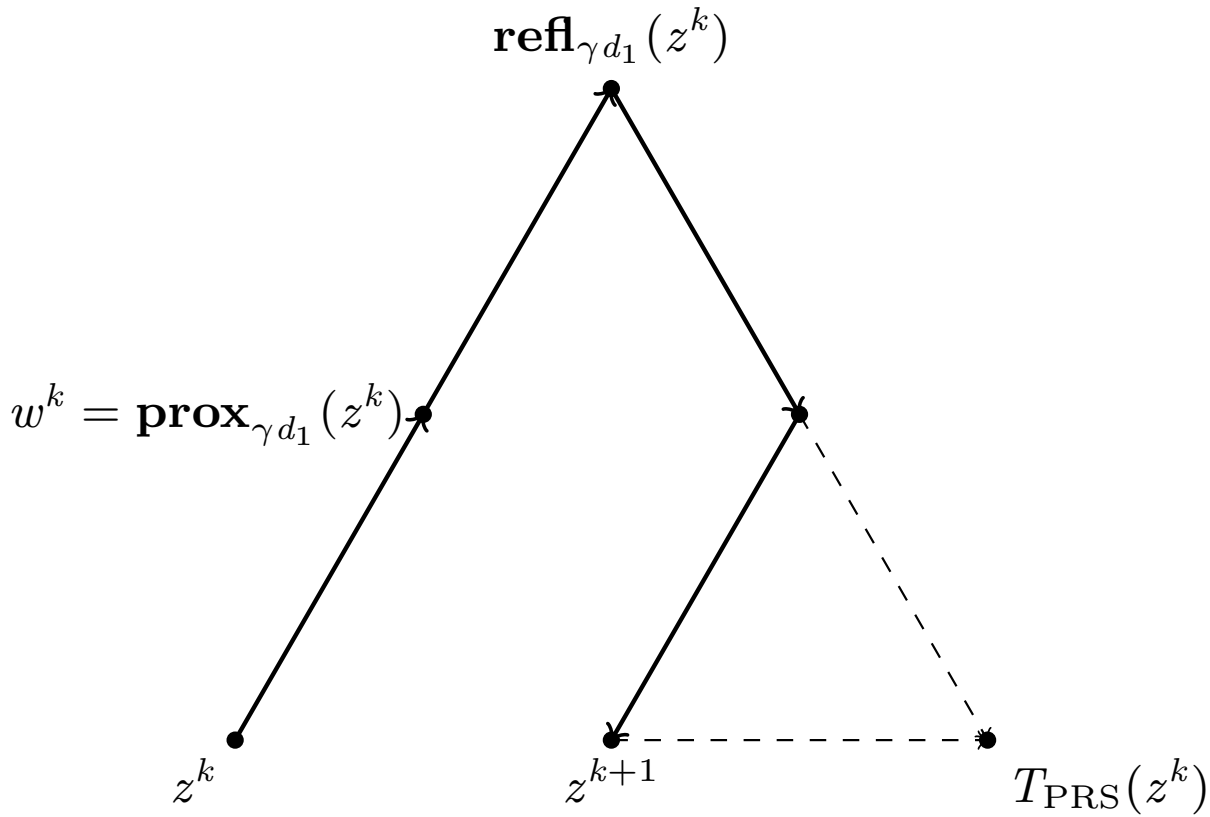


Diagram of ADMM

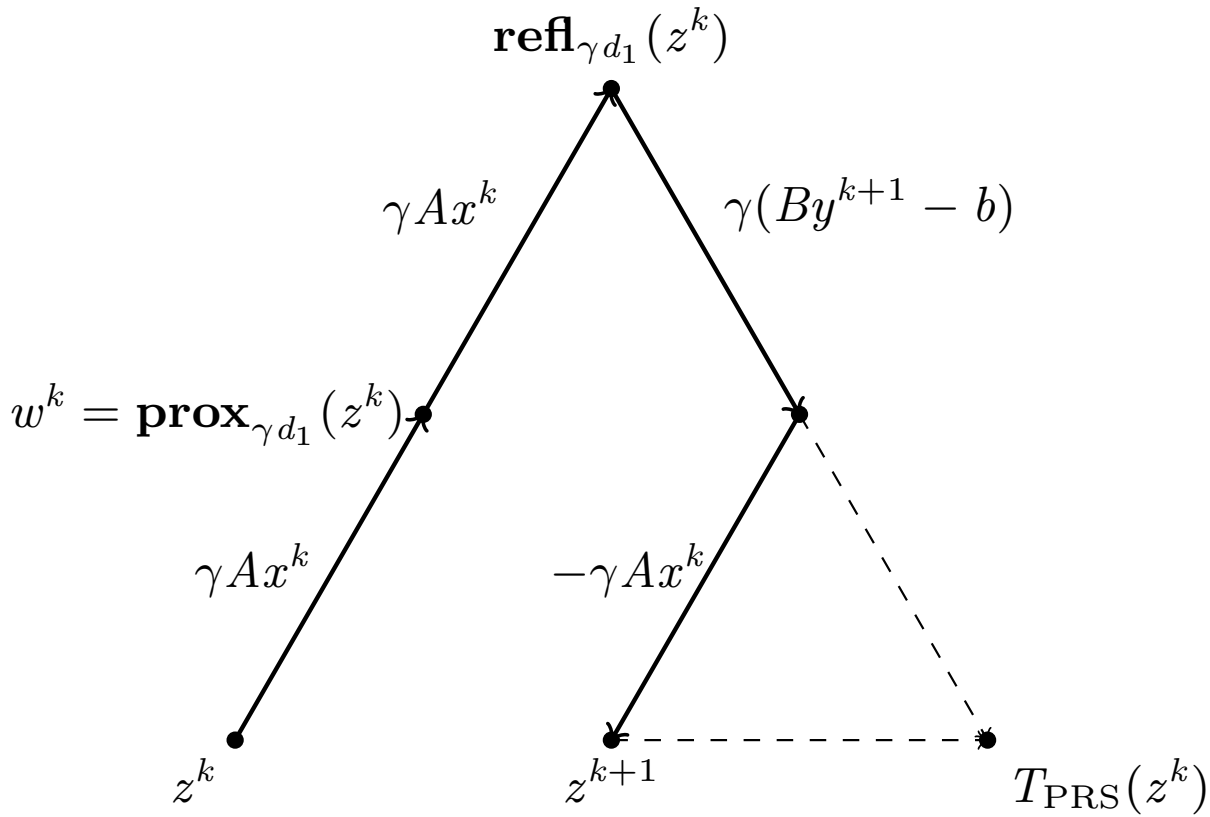


Diagram of ADMM

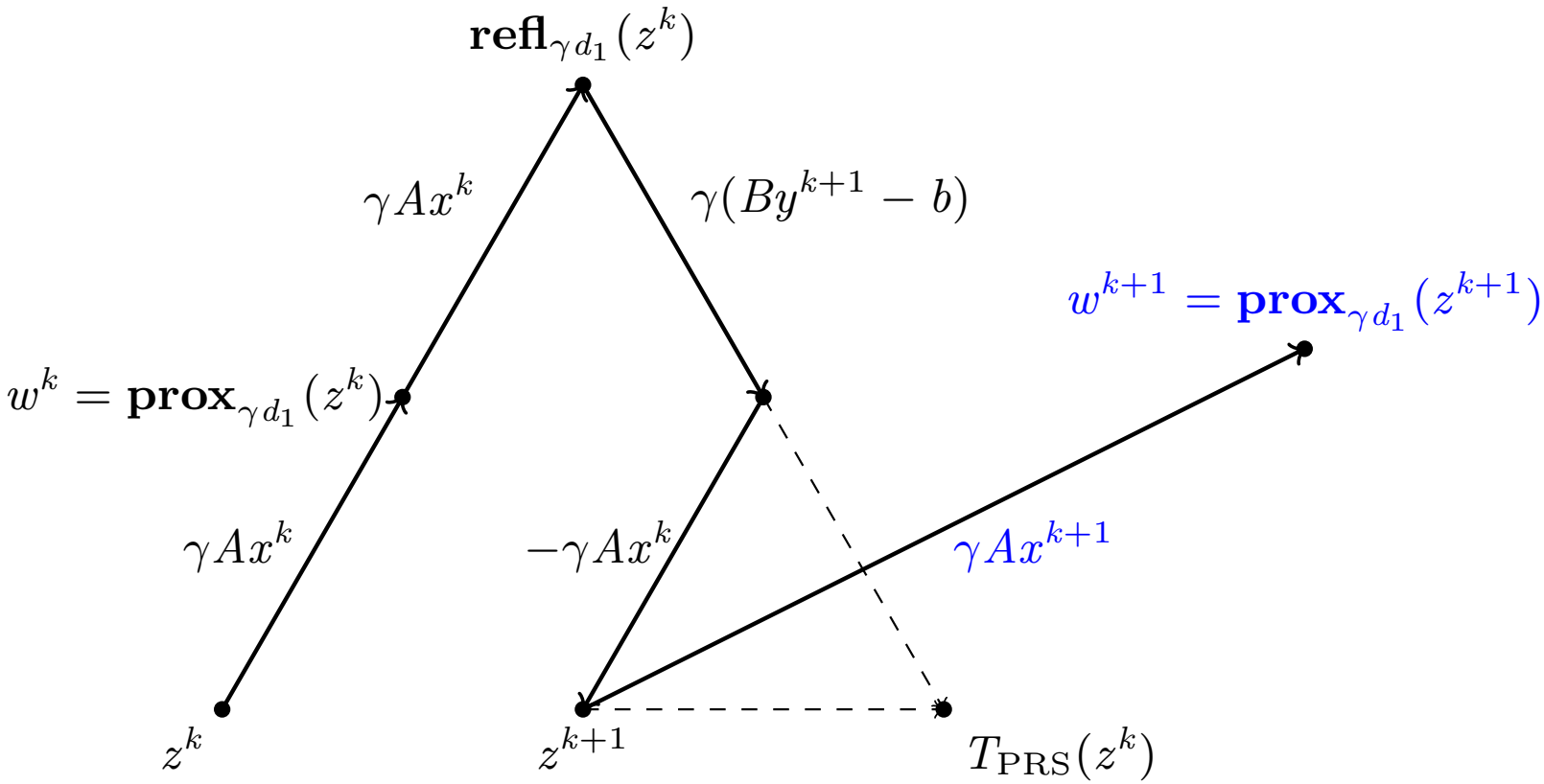
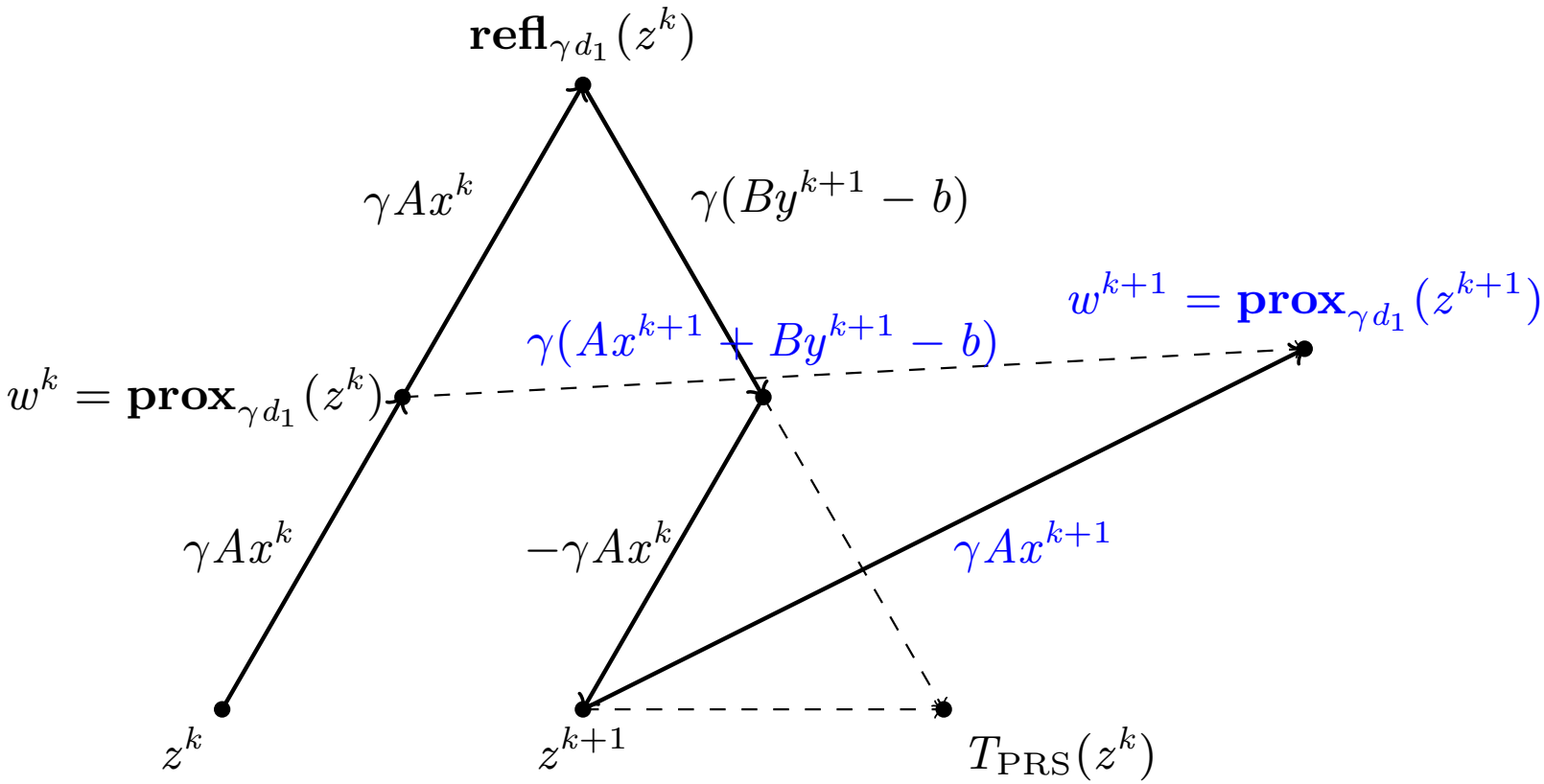


Diagram of ADMM



Generally, Krasnosel'skiĭ-Mann (KM) iteration[4] [5]

- Definitions:
  - $\mathcal{H}$ Hilbert space.
  - $T : \mathcal{H} \to \mathcal{H}$ nonexpansive.
  - **Fixed points:** $z \in \mathcal{H}$ such that $Tz = z$.
- Averaged iteration of T, (aka KM iteration)

$$z^{k+1} = T_{\lambda_k}(z^k) := (1 - \lambda_k)z^k + \lambda_k Tz^k.$$

- **Convergence:**
  - Converges weakly to a fixed point if λ_k bounded away from 0 and 1.
  - If no fixed point and λ_k is bounded away from 0, the sequence $(z^j)_{j\geq 0}$ is unbounded. (Browder-Göhde-Kirk fixed-point theorem.)
- **Special cases:** DRS, PRS, ADMM, FBS, PPA, ...

[4] Krasnosel'skiĭ: Two remarks on the method of successive approximations (1955)

[5] Mann: Mean value methods in iteration (1953)

Part 2: Convergence rates

Fixed-point residual

- The **fixed-point residual** (FPR) of the KM iteration:

$$\|Tz^k - z^k\|^2 = \frac{1}{\lambda_k^2}\|z^{k+1} - z^k\|^2.$$

- $\|Tz - z\| = 0$ often means z is optimal.
- Small FPR implies

$$Tz^k \approx z^k.$$

- The property $\|Tz^k - z^k\| \to 0$ is called *asymptotic regularity*.
- In general, convergence of $\|z^k - z^*\|$ can be arbitrarily slow.
- In optimization $Tz^k - z^k$ is usually some sorts of gradients or subgradients, so it is a dual measure of optimality
- The rate of $\|Tz^k - z^k\|^2$ controls the progress of convergence
- In ADMM:

$$Tz^k - z^k = -2\gamma(Ax^k + By^k - b).$$

History of FPR

- **1978 ($\lambda = 1/2$):** Brèzis and Lions[6] show FPR satisfies

$$\| Tz^k - z^k \|^2 = O\left(\frac{1}{k+1}\right).$$

  If $T = \mathbf{prox}_{\gamma f}$, then

$$\| Tz^k - z^k \|^2 = O\left(\frac{1}{(k+1)^2}\right).$$

- **1996 (General λ):** Baillon and Bruck[7] conjecture $O\left(1/(k+1)\right)$ for nonexpansive maps on Banach spaces.
- **2012 (General λ):** Cominetti, Soto, and Vaisman[8] prove the conjecture of Baillon and Bruck.

[6] Produits infinis de resolvantes

[7] The rate of asymptotic regularity is $O(1/\sqrt{k})$

[8] On the rate of convergence of Krasnosel'skiĭ-Mann iterations and their connection with sums of Bernoullis

Convergence rates

Objective error

- **(Non-ergodic) error:** consider minimizing $h(x)$ and x^* is a minimizer of h:

$$h(x^k) - h(x^*)$$

  - Its convergence to zero does *not* imply strong convergence.
  - Useful as a filter through which we view the distance to the solution.

- **Ergodic error:** Define ergodic iterates:

$$\overline{x}^k = (1/\Lambda_k) \sum_{i=0}^{k} \lambda_i x_g^k, \quad \text{where } \Lambda_k = \sum_{i=0}^{k} \lambda_i,$$

  We measure the quantity

$$h(\overline{x}^k) - h(x^*)$$

History of objective error

- **1967** Polyak proved the subgradient method achieves $O(1/\sqrt{k+1})$.
- **1980s** Nemirovsky and Yudin show lower complexity of $\Omega(1/\sqrt{k+1})$ for general class of subgradient methods.
- **1980s** ? showed gradient descent $O(1/(k+1))$.
- **1983** Nesterov proposed accelerated gradient descent to achieve $O(1/(k+1)^2)$.
- **1991** Güler proved $O(1/(k+1))$ convergence for PPA.
- **2009** Beck and Teboulle proved $O(1/(k+1))$ for FBS, and proposed accelerated variant that achieves $O(1/(k+1)^2)$.
- **2012** Goldstein, O'Donoghue, and Setzer proved $O(1/(k+1))$ for ADMM when objectives both primal objectives are strongly convex.
- **2012** Wei and Ozdaglar showed $O(1/(k+1))$ ergodic convergence of ADMM with specific binary matrix A and B.
- **2013** He and Yuan showed $O(1/(k+1))$ of VI-based optimality violation.
- **Recently**, Bot, Chambolle, Deng, Falidi, Lai, Ma, Monteiro, Peyre, Pock, Svaiter, Zhang, **violation to VI and Lagrangian optimality, duality gap**

Contributions on rates (with Damek Davis)

KM iteration: FPR $o(1/k)$, tight, improved from O to o.

PPA based on $\mathbf{prox}_f$: FPR $o(1/k^2)$, **tight** (by an example in Brezis-Lions'78) improved; objective $o(1/k)$, **tight** (by an infinite-dim example).

FBS based on $I - \nabla g$ and $\mathbf{prox}_f$: same rates as PPA, **tight**.

Relaxed PRS (including DRS and, for some, also PRS): all are new

- FPR: $o(1/k)$, **tight** (by an infinite-dim example)
- Ergodic squared feasibility: $O(1/k^2)$, **tight** (by a 2D example)
- **Lipschitz f or g**: ergodic objective: $o(1/k)$, **tight** (by a 1D example)
  objective: $o(1/\sqrt{k})$, **tight** (by an infinite-dim example)
- **Strongly convex f or g**: strong sequence convergence,
  best sequence error $o(1/k)$
  ergodic error $O(1/k)$
- **Gradient Lipschitz f or g**: best objective $o(1/k)$;
  Limit γ properly: objective $o(1/k)$ and FPR $o(1/k^2)$
- **Strongly convex + gradient Lipschitz (applied to either the same or different functions)**: all rates (FPR, objective, sequence) are linear

ADMM (as dual DRS) by $\partial d_f = A \circ (\partial f^*) \circ A^*$ and $\partial d_g = B \circ (\partial g^*) \circ B^*$.

- f strongly convex $\Rightarrow$ d_f is differentiable and Lipschitz; (same for g)
- f differentiable and AA^* is full-rank $\Rightarrow$ d_f is strongly convex; (same for g)

Translate the results from relaxed PRS to ADMM: **general case:**

- ergodic squared constraint feasibility: $O(1/k^2)$
  squared constraint feasibility: $o(1/k)$
- ergodic objective: $O(1/k)$
  objective: $o(1/\sqrt{k})$
- **strongly convex** f **or** g: squared feasibility $o(1/k^2)$, objective $o(1/k)$
- **strongly convex function + gradient Lipschitz + matrix full-rank**:
  everything linear convergence
  Note: Results in Deng-Yin-2012 cover more cases.

Method of **alternating projection** for finding $x \in C_1 \cap C_2$:

- **linear regularity**: a special case of PRS, all rates linear
- in general: same as relax PRS with gradient Lipschitz objectives
- results extended to $x \in C_1 \cap \cdots \cap C_n$
- when $C_1 \cap C_2 = \emptyset$, converge to the shortest line segment between

DRS (**"reflect, reflect, average"**) for finding $x \in C_1 \cap C_2$:

- in general: a sequence of points in each set
- distance to other set: general $o(1/k)$, ergodic $O(1/k^2)$

Results in a nutshell

- Essentially **tight upper and lower bounds** on **fixed-point residual (FPR)** for KM iterations.
- Relaxed PRS point sequence can converge strongly yet **arbitrarily slowly**.

Objective convergence:

- **On average**, relaxed PRS performs as well as PPA.
- **In the worst case**, relaxed PRS performs nearly as slowly as the subgradient method.
- When ∇g is Lipschitz, DRS performs as well as FBS, yet **no knowledge of Lipschitz constant is needed**.

- Relaxed PRS algorithm converges *linearly* whenever **one of the objectives is strongly convex and one has a Lipschitz derivative**. They can be either the same or different functions.
- For feasibility problems relaxed PRS converges linearly under **regularity assumptions on the intersection**.
- For feasibility problems with no regularity, we can generate a point in each set and bound their distance to each other.
- ADMM produces similar rates for **objective functions** and the **feasibility** separately.

Part 3: Basic lemma for summable and monotonic sequence

Lemma

Suppose that ***nonnegative*** *scalar sequences* $(\lambda_j)_{j\geq 0}$ *and* $(a_j)_{j\geq 0}$ *that are* ***summable*** $\sum_{i=0}^{\infty} \lambda_i a_i < \infty$. *Let* $\Lambda_k := \sum_{i=0}^{k} \lambda_i$ *for* $k \geq 0$.

1. *If* $(a_j)_{j\geq 0}$ *is* ***monotonically nonincreasing****, then*

$$a_k \leq \frac{1}{\Lambda_k}\left(\sum_{i=0}^{\infty} \lambda_i a_i\right) \qquad \textit{and} \qquad a_k = o\left(\frac{1}{\Lambda_k - \Lambda_{\lceil k/2 \rceil}}\right). \tag{1}$$

   1.1 *If* $(\lambda_j)_{j\geq 0}$ *is bounded away from* 0 *and* ∞, *then* $a_k = o(1/(k+1))$;

   1.2 *If* $\lambda_k = (k+1)^p$ *for* $p \geq 0$ *and all* $k \geq 1$, *then* $a_k = o(1/(k+1)^{p+1})$.

2. *Suppose that the nonnegative scalar sequence* $(b_j)_{j\geq 0}$ *is* ***monotonically nonincreasing*** *and satisfies* $b_k \leq \lambda_k a_k - \lambda_{k+1} a_{k+1}$. *Then for all* $k \geq 0$

$$b_k \leq \frac{1}{(k+1)^2}\left(\sum_{i=0}^{\infty} \lambda_i a_k\right) \qquad \text{and} \qquad b_k = o\left(\frac{1}{(k+1)^2}\right). \tag{2}$$

Intuitions

- Every convergence rate follows from this lemma.
- Sequence $(1/(j+1))_{j\geq 0}$ is not summable, so a_k must decrease "*faster*."
- Follows because $\sum_{i+1}^{2i} \lambda_i a_i \to 0$

Extensions:

- Same assumptions except **quasi monotonic**: $a_{k+1} \leq a_k + e_k$. Then, e_k has to converge one-order faster than a_k to preserve its rate.
- Same assumptions but **no monotonicity**:

$$k_{\text{best}} := \arg\min_i \{a_i : i = 0, \dots, k\}.$$

  Then, all the rates hold for $a_{k_{\text{best}}}$ instead of a_k.

The idea of FPR convergence rate

- **In general:**
  - The term $\| Tz^k - z^k \|^2 = \frac{1}{\lambda_k^2} \| z^k - z^{k+1} \|^2$ is **monotonic**.
  - Furthermore $\sum_{i=0}^{\infty} \lambda_k (1 - \lambda_k) \| Tz^k - z^k \|^2 < \infty$.
  - Thus, convergence controlled by $\sum_{i=0}^{k} \lambda_k (1 - \lambda_k)$
- **When ∇g is Lipschitz (PPA, FBS, DRS)**
  - Still have monotonicity, but $\sum_{i=0}^{\infty} (i+1) \| Tz^k - z^k \|^2 < \infty$.
  - **Requires information about the objective functions.**.

Example: PPA convergence rate

$$\underset{x}{\text{minimize}}\ h(x)$$

- PPA iteration: $z^{k+1} = z^k - \gamma \widetilde{\nabla} h(z^{k+1})$
- Minimizer z^*
- Objective error sequence $a_k = h(z^{k+1}) - h(z^*)$
- FPR sequence $b_k = (1/\gamma)\|z^{k+2} - z^{k+1}\|^2$
- For any z,

$$\begin{aligned} h(z^{k+1}) - h(z) &\leq \langle z^{k+1} - z, \widetilde{\nabla} h(z^{k+1}) \rangle \quad (\textbf{(sub)gradient inequality}) \\ &= \frac{1}{\gamma} \langle z^{k+1} - z, z^k - z^{k+1} \rangle \\ &= \frac{1}{2\gamma} \left(\|z^k - z\|^2 - \|z^{k+1} - z\|^2 - \|z^{k+1} - z^k\|^2 \right). \end{aligned}$$

$$h(z^{k+1}) - h(z) \leq \frac{1}{2\gamma}\left(\|z^k - z\|^2 - \|z^{k+1} - z\|^2 - \|z^{k+1} - z^k\|^2\right).$$

- **Nonnegativity:** obvious
- **Summability:** at $z = z^*$:
  $a_k \leq (1/2\gamma)\left(\|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 - \|z^{k+1} - z^k\|^2\right)$
  $\implies \sum_{i=0}^{\infty} a_k \leq (1/2\gamma)\|z^0 - z^*\|.$
- **Monotonicity:** at $z = z^k$

$$0 \leq \boxed{b_k} = (1/\gamma)\|z^{k+2} - z^{k+1}\|^2 \leq h(z^{k+1}) - h(z^{k+2}) = \boxed{a_k - a_{k+1}}.$$

- By the lemma:
$$a_k = o\left(\frac{1}{k+1}\right).$$
- Also, b_k (= FPR) is monotonic $\implies$
$$b_k = o\left(\frac{1}{(k+1)^2}\right).$$

A fundamental inequality

Proposition

If $z^+ = (T_{\mathrm{PRS}})_\lambda(z)$, *then for all* $x \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$

$$
\begin{aligned}
&4\gamma\lambda(f(x_f) + g(x_g) - f(x) - g(x)) \\
&\qquad \leq \|z - x\|^2 - \|z^+ - x\|^2 + \left(1 - \frac{1}{\lambda}\right)\|z^+ - z\|^2 \\
&\qquad = 2\langle z^+ - x, z - z^+\rangle + 2\left(1 - \frac{1}{2\lambda_k}\right)\|z^+ - z\|^2.
\end{aligned}
$$

- **Nonergodic rate:**
  - Use Cauchy-Schwarz on inner product.
  - The objective error involves both x_f and x_g. It can be negative. The inequality also has **the other side, i.e., a lower bound**.
  - Additional regularity properties enable same-point objective error
- **Ergodic rate:** Sum both sides, divide by Λ_j, and use Jensen's inequality.

The other cases

- If ∇f or ∇g is lipschitz then

$$\sum_{i=0}^{\infty} \lambda_k (f(x^k) + g(x^k) - f(x^*) - g(x^*)) < \infty.$$

  $\implies$ "best-point" convergence rates

- When $\lambda_k \equiv 1/2$ and ∇g is Lipschitz, we have to construct **an auxiliary monotonic sequence** that dominates the objective.
- Under strong convexity, $\sum_{i=0}^{\infty} \lambda_k \|x^k - x^*\|^2 < \infty \implies$ "running-best" convergence rates.
- The feasibility problem and the linear convergence result use same fundamental inequality.

Other applications

- More applications in paper:
  - Feasibility;
  - Parallelized model fitting;
  - Linear Programming (linear convergence);
  - Semidefinite programming.

Part 4: Primal-dual equivalence (with Ming Yan)

- **Definition:** apply the same algorithm to both primal and dual problems, with proper initialization and parameters, the iterates of one can be explicitly reconstructed from those of the other.
- Eckstein[9] shows DRS is equivalent to DRS on the dual, for a special case
- Eckstein and Fukushima[10] shows ADMM is equivalent to ADMM on the dual, for a special case $\mathbf{A}\mathbf{A}^T = I$. Rarely mentioned in the literature.
- We extend the result to ADMM and relaxed PRS (including DRS and PRS) for general cases, assuming only convexity and the existence of primal-dual solutions.
- We introduce an equivalent primal-dual algorithm for the saddle-point problem.
- We establish conditions for the equivalence between ADMMs with swapped orders of subproblems.

[9] Eckstien. Splitting methods for monotone operators with applications to parallel optimization, PhD thesis, 89'.

[10] Eckstein and Fukushima. Some reformulations and applications of the alternating direction, 1994.

Remarks

- Different splitting leads to different ADMM iterates.
- Specifically, we consider

$$\begin{aligned} \underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \;\; & f(\mathbf{x}) + g(\mathbf{y}) \\ \text{subject to} \;\; & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b} \end{aligned} \tag{P1}$$

and its dual

$$\underset{\mathbf{v}}{\text{minimize}} \quad f^*(-\mathbf{A}^*\mathbf{v}) + g^*(-\mathbf{B}^*\mathbf{v}) + \langle \mathbf{v}, \mathbf{b} \rangle.$$

ADMM is applied to (P1) the reformulated dual problem:

$$\begin{aligned} \underset{\mathbf{u},\mathbf{v}}{\text{minimize}} \quad & f^*(-\mathbf{A}^*\mathbf{u}) + (g^*(-\mathbf{B}^*\mathbf{v}) + \langle \mathbf{v}, \mathbf{b} \rangle) \\ \text{subject to} \quad & \mathbf{u} - \mathbf{v} = \mathbf{0}. \end{aligned} \tag{D1}$$

- Examples: YALL1 package[11], ℓ_1-ℓ_1 model[12], traffic equilibrium[13], dual alternating projection.

[11] J.Yang and Y.Zhang, Alternating direction algorithms for ℓ_1-problems in compressive sensing, 2011.

[12] Y.Xiao, H.Zhu, S.-Y. Wu. Primal and dual alternating direction algorithms for l1-l1-norm minimization problems in compressive sensing, 2013.

[13] Primal: Fukushima'96; dual: Gabay'83.

Remarks

- Penalty parameter λ in the primal ADMM becomes λ^{-1} in the dual ADMM. It balances primal-dual progress.
- The perfect symmetry between primal and dual ADMMs suggest that ADMM is a primal-dual algorithm to a saddle-point formulation.

Saddle-point formulation and its algorithm

The original problem (P1) is equivalent to

$$\min_{\mathbf{y}} \max_{\mathbf{u}} \; g(\mathbf{y}) + \langle \mathbf{u}, \mathbf{B}\mathbf{y} - \mathbf{b} \rangle - f^*(-\mathbf{A}^*\mathbf{u}).$$

Primal-Dual Algorithm: Initialize $\mathbf{u}^0$, $\mathbf{u}^{-1}$, $\mathbf{y}^0$, $\lambda > 0$, **for** $k = 0, 1, \ldots,$ **do**:

- $\bar{\mathbf{u}}^k = 2\mathbf{u}^k - \mathbf{u}^{k-1}$
- $\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}} \; g(\mathbf{y}) + (2\lambda)^{-1} \|\mathbf{B}\mathbf{y} - \mathbf{B}\mathbf{y}^k + \lambda \bar{\mathbf{u}}^k\|_2^2$
- $\mathbf{u}^{k+1} = \arg\min_{\mathbf{u}} \; f^*(-\mathbf{A}^*\mathbf{u}) + \lambda/2 \|\mathbf{u} - \mathbf{u}^k - \lambda^{-1}(\mathbf{B}\mathbf{y}^{k+1} - \mathbf{b})\|_2^2$

Remarks:

- If $\mathbf{B} = \mathbf{I}$, then it is equivalent to Chambolle-Pock, whose paper also noted the equivalence between it and ADMM.
- ADMM and PD have the same # iterations but different flops per-iteration.

Application: extended monotropic programming

$$\underset{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N}{\text{minimize}} \sum_{i=1}^{N} f_i(\mathbf{x}_i), \quad \text{subject to } \sum_{i=1}^{N} \mathbf{A}_i\mathbf{x}_i = \mathbf{b}.$$

Convert the problem into the following ADMM-ready formulation

$$\begin{cases} \underset{\{\mathbf{x}_i\},\{\mathbf{y}_i\}}{\text{minimize}} & \sum_{i=1}^{N} f_i(\mathbf{x}_i) + \iota_{\{\mathbf{y}:\sum_{i=1}^{N}\mathbf{y}_i=\mathbf{b}\}}(\mathbf{y}) \\ \text{subject to} & \mathbf{A}_i\mathbf{x}_i - \mathbf{y}_i = \mathbf{0}. \end{cases}$$

- **ADMM**: iteratively update $\{\mathbf{x}_i\}, \{\mathbf{y}_i\}, \{\mathbf{u}_i\}$
- **Primal-Dual:** iteratively update $\{\mathbf{y}_i\}, \{\mathbf{u}_i\}$, and at the end recover $\{\mathbf{x}_i\}$

Assumption: $f^*(-\mathbf{A}^*\mathbf{u})$ has an easy form, for example, when

- $f_i(\cdot) = (1/2)\|\cdot\|^2$
- $\mathbf{A}_i \in \mathbb{R}^{m\times n_i}$ and $\mathbf{A}_i\mathbf{A}_i^T = I$.

For each iteration k and block i:

- ADMM: $10m + 2mn_i$
- PD: $10m$ due to the hiding of $\mathbf{x}_i$

Pre/post-processing:

- ADMM has a pre-step of mn_i for each i
- PD has a post-step of mn_i for each i

Distributed computing:

- Same communication for ADMM and PD
- PD has better load balance since its per-iteration flop is independent of n_i

Swap x/y-update order

Two similar ADMM on the same problem:

- ADMM 1 updates $\mathbf{y}$, then $\mathbf{x}$, then dual variable $\mathbf{z}$
- ADMM 2 updates $\mathbf{x}$, then $\mathbf{y}$, then dual variable $\mathbf{z}$

In general, they produce different iterates, but there are exceptions. **Define:**

$$F(\mathbf{s}) := \min_{\mathbf{x}} f(\mathbf{x}) + \iota_{\{\mathbf{x}:\mathbf{A}\mathbf{x}=\mathbf{s}\}}(\mathbf{x}), \tag{3a}$$

$$G(\mathbf{t}) := \min_{\mathbf{y}} g(\mathbf{y}) + \iota_{\{\mathbf{y}:\mathbf{B}\mathbf{y}=\mathbf{b}-\mathbf{t}\}}(\mathbf{y}). \tag{3b}$$

Theorem

1. *Assume* $\mathbf{prox}_G$ *is* ***affine****. Given the iterates of ADMM 2, if* $\mathbf{z}_2^0 \in \partial G(\mathbf{b} - \mathbf{B}\mathbf{y}_2^0)$*, then the iterates of ADMM 1 can be recovered as*

$$\mathbf{x}_1^k = \mathbf{x}_2^{k+1}, \qquad \mathbf{z}_1^k = \mathbf{z}_2^k + \lambda^{-1}(\mathbf{A}\mathbf{x}_2^{k+1} + \mathbf{B}\mathbf{y}_2^k - \mathbf{b}).$$

2. *Assume* $\mathbf{prox}_F$ *is* ***affine****. Given the iterates of ADMM 1, if* $-\mathbf{z}_1^0 \in \partial G(\mathbf{A}\mathbf{x}_1^0)$ *, then the iterates of ADMM 2 can be recovered as*

$$\mathbf{y}_2^k = \mathbf{y}_1^{k+1}, \qquad \mathbf{z}_2^k = \mathbf{z}_1^k + \lambda^{-1}(\mathbf{A}\mathbf{x}_1^{k+1} + \mathbf{B}\mathbf{y}_1^{k+1} - \mathbf{b}).$$

Affine proximal mapping

Definition

A mapping T is affine if, for any $\mathbf{r}_1$ and $\mathbf{r}_2$,

$$T\left(\frac{1}{2}\mathbf{r}_1 + \frac{1}{2}\mathbf{r}_2\right) = \frac{1}{2}T\mathbf{r}_1 + \frac{1}{2}T\mathbf{r}_2.$$

Proposition

Let G be a proper, closed, convex function. The following statements are equivalent:

1. $\mathbf{prox}_{G(\cdot)}$ *is affine;*
2. $\mathbf{prox}_{\lambda G(\cdot)}$ *is affine for $\lambda > 0$;*
3. $a\mathbf{prox}_{G(\cdot)} \circ b\mathbf{I} + c\mathbf{I}$ *is affine for any scalars a, b and c;*
4. $\mathbf{prox}_{G^*(\cdot)}$ *is affine;*
5. *G is **convex quadratic** (or, affine or constant) and has an **affine domain** (either $\mathcal{G}$ or the intersection of hyperplanes in $\mathcal{G}$).*

If function g obeys Part 5, then G defined in (3b) satisfies Part 5, too.

Conclusion

- **Our work**
  - Analyzed relaxed PRS, ADMM, and KM iterations.
  - Provided worst-case non-asymptotic convergence analysis.
  - Provided lower complexity bounds for the basic rates.
  - Showed the limitations of the methods.
  - Established primal-dual equivalence and conditions for order-swapping equivalence

- **Reflections**
  - The methods are essentially **nonexpansive operator splitting** iterations applied to optimality conditions of the original problem
  - When splitting methods use points other than z^k, it **lacks an objective function for monotonic decrease or for acceleration**
  - Splitting methods based on **implicit steps automatically adjust to regularity properties present**. (That's why they are practically fast.)