Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions.

秦雪、黄卫杰、崔金杰

School of Mathematical Sciences
Peking University

December 30, 2014

Damek Davis, Wotao Yin.
Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions.

Optimization Problems

In this paper, we consider two prototype optimization problems:

1. the unconstrained problem (problem 1):

$$\min_{x\in\mathcal{H}} \quad f(x)+g(x) \tag{1}$$

2. the linearly constrained variant (problem 2):

$$\min_{x\in\mathcal{H}_1, y\in\mathcal{H}_2} \quad f(x)+g(y) \tag{2}$$

$$\text{s.t.} \quad Ax+By=b \tag{3}$$

where $b\in\mathcal{G}$, $\mathcal{H}_1,\mathcal{H}_2,\mathcal{G}$ are Hilbert spaces and

$$A:\mathcal{H}_1\to\mathcal{G} \tag{4}$$

$$B:\mathcal{H}_2\to\mathcal{G} \tag{5}$$

are linear operators.

Proximal Operator, Reflection Operator

Definition 1 (Proximal Operator, Reflection Operator)

For any point $x \in \mathcal{H}$ and any scalar $\gamma \in \mathbb{R}_{++}$, we define the proximal operator as

$$\mathbf{prox}_{\gamma f}(x) := \arg\min_{y \in \mathcal{H}} \quad f(y) + \frac{1}{2\gamma}\|y - x\|^2 \tag{6}$$

and reflection operator as

$$\mathbf{refl}_{\gamma f} := 2\mathbf{prox}_{\gamma f} - I_{\mathcal{H}} \tag{7}$$

where $I_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ denote the identity map.

Peaceman-Rachford Splitting(PRS) operator

Definition 2 (PRS operator)

We define the PRS operator:

$$T_{PRS} := \boldsymbol{refl}_{\gamma f} \circ \boldsymbol{refl}_{\gamma g}. \tag{8}$$

Definition 3 (fixed-point residual)

We call the quantity

$$\| T_{PRS} z^k - z^k \|^2 \tag{9}$$

the fixed-point residual (FPR) of the relaxed PRS algorithm.

Subgradient,Subdifferential

Definition 4 (subgradient,subdifferential)

Given a closed, proper, and convex function $f : \mathcal{H} \to (-\infty, \infty]$*, the set* $\partial f(x)$ *denotes its subdifferential at* x *and*

$$\widetilde{\nabla} f(x) \in \partial f(x) \tag{10}$$

denotes a subgradient.

Forward Backward Splitting(FBS) Algorithm

Suppose that the function g in Problem (1) is differentiable and ∇g is $(1/\beta)$-Lipschitz. The FBS algorithm is: given $z^0 \in \mathcal{H}$, for all $k \geq 0$,

$$z^{k+1} = \textbf{prox}_{\gamma f}(z^k - \gamma \nabla g(z^k)). \tag{11}$$

Dauglas-Rachford splitting(DRS) algorithm

Starting from an arbitrary $z^0 \in \mathcal{H}$, repeat

$$x_g^k = \boldsymbol{prox}_{\gamma g}(z^k); \tag{12}$$

$$x_f^k = \boldsymbol{prox}_{\gamma f}(2x_g^k - z^k); \tag{13}$$

$$z^{k+1} = z^k + (x_f^k - x_g^k) \tag{14}$$

where γ is a positive constant (simply scales the objective).

Peaceman-Rachford splitting(PRS) algorithm

Starting from an arbitrary $z^0 \in \mathcal{H}$, repeat

$$z^{k+1} = T_{PRS}(z^k) \tag{15}$$

Lemma 1.1

Let $z \in \mathcal{H}$. Define auxiliary points $x_g := \textbf{prox}_{\gamma g}(z)$ and $x_f := \textbf{prox}_{\gamma f}(\textbf{refl}_{\gamma g}(z))$. Then the identities hold:

$$x_g = z - \gamma\widetilde{\nabla} g(x_g) \tag{16}$$

$$x_f = x_g - \gamma\widetilde{\nabla} g(x_g) - \gamma\widetilde{\nabla} f(x_f) \tag{17}$$

In addition, each relaxed PRS step $z^+ = (T_{PRS})_\lambda(z)$ has the following representation:

$$z^+ - z = 2\lambda(x_f - x_g) = -2\lambda\gamma(\widetilde{\nabla} g(x_g) + \widetilde{\nabla} f(x_f)) \tag{18}$$

Equivalent operator of DRS

The DRS has the equivalent operator-theoretic and subgradient form

$$z^{k+1} = \frac{1}{2}(I_{\mathcal{H}} + T_{PRS})(z^k) = z^k - \gamma(\widetilde{\nabla} f(x_f^k) + \widetilde{\nabla} g(x_g^k)), \quad k = 0, 1 \cdots .$$

where $\widetilde{\nabla} f(x_f^k) \in \partial f(x_f^k)$ and $\widetilde{\nabla} g(x_g^k) \in \partial g(x_g^k)$.

Relaxed PRS

In the DRS algorithm, we can replace the (1/2)-average of $I_{\mathcal{H}}$ and T_{PRS} with any other weight and this results the **relaxed PRS** algorithm:

$$z^{k+1} = (1-\lambda_k)z^k + \lambda_k \boldsymbol{refl}_{\gamma f} \circ \boldsymbol{refl}_{\gamma g}(z^k) \tag{19}$$

The special cases $\lambda_k \equiv 1/2$ and $\lambda_k \equiv 1$ are called the DRS and PRS algorithms, respectively.
The relaxed PRS algorithm can be applied to problem (2).

ADMM and Relaxed ADMM

ADMM is equivalent to DRS applied to the Lagrange dual of Problem 2.
If we let

$$\mathcal{L}(x, y, w) := f(x) + g(y) - < w, Ax + By - b > \quad (20)$$

$$d_f(w) := f^*(A^* w) \quad (21)$$

$$d_g(w) := g^*(B^* w) - < w, b > \quad (22)$$

Relax ADMM is equivalent to relaxed PRS applied to the following problem:

$$\min_{w \in \mathcal{G}} \quad d_f(w) + d_g(w) \quad (23)$$

ADMM and Relaxed ADMM

Applying the relaxed PRS algorithm to (23) according to Lemma (1.1)

$$w_{d_g}^k = \textbf{prox}_{\gamma d_g}(z^k); \tag{24}$$

$$w_{d_f}^k = \textbf{prox}_{\gamma d_f}(2w_{d_g}^k - z^k); \tag{25}$$

$$z^{k+1} = z^k + 2\lambda_k(w_{d_f}^k - w_{d_g}^k). \tag{26}$$

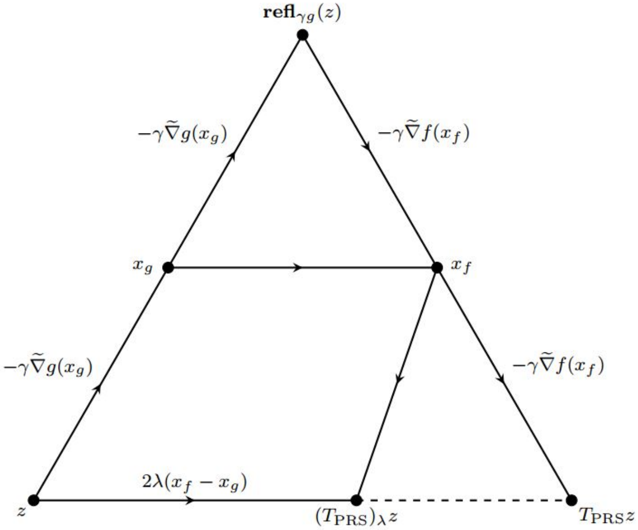

Fig. 1 A single relaxed PRS iteration starting from z.

$$S_f(x, y) = \max\left\{\frac{\mu_f}{2}\|x - y\|^2, \frac{\beta_f}{2}\|\widetilde{\nabla} f(x) - \widetilde{\nabla} f(y)\|^2\right\} \tag{27}$$

Strong convexity

Assume: one of the functions is strong convex and the sequence $(\lambda_j)_{j\geq 0} \subset (0, 1]$ is bounded away from zero.

Theorem 1 (Auxiliary term bound)

Suppose that $(z^j)_{j\geq 0}$ is generated by Algorithm 1. Then for all $k \geq 0$,

$$8\gamma\lambda_k(S_f(x_f^k, x^*) + S_g(x_g^k, x^*)) \leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \left(1 - \frac{1}{\lambda_k}\right)\|z^{k+1} - z^k\|^2. \tag{28}$$

Therefor, $8\gamma \sum_{i=0}^{\infty} \lambda_k(S_f(x_f^i, x^*) + S_g(x_g^i, x^*)) < \|z^0 - z^*\|^2$*, and*

1. ***Best iterate convergence:*** *If* $\underline{\lambda} := \inf_{j \geq 0} \lambda_j > 0$, *then*

$$S_f(x_f^{best}, x^*) + S_g(x_g^{best}, x^*) \leq \frac{\|z^0 - z^*\|^2}{8\gamma\underline{\lambda}(k+1)}, \tag{29}$$

and thus

$$S_f(x_f^{best}, x^*) = o\left(\frac{1}{k+1}\right) \quad \text{and} \quad S_g(x_g^{best}, x^*) = o\left(\frac{1}{k+1}\right) \tag{30}$$

2. ***Ergodic convergence:*** *Let* $\overline{x}_f^k = (1/\Lambda_k)\sum_{i=0}^k \lambda_i x_f^i$ *and* $\overline{x}_g^k = (1/\Lambda_k)\sum_{i=0}^k \lambda_i x_g^i$. *Then*

$$\overline{S}_f(x_f^k, x^*) + \overline{S}_g(x_g^k, x^*) \leq \frac{\|z^0 - z^*\|^2}{8\gamma\Lambda_k} \tag{31}$$

where

$$\overline{S}_f(x_f^k, x^*) := \max\left\{\frac{\mu_f}{2}\|\overline{x}_f^k - x^*\|^2, \frac{\beta_f}{2}\left\|\frac{1}{\Lambda_k}\sum_{i=0}^k \widetilde{\nabla} f(x_f^k) - \widetilde{\nabla} f(x^*)\right\|^2\right\} \tag{32}$$

3. ***Nonergodic convergence:*** *If* $\underline{\tau} = \inf_{j\geq 0} \lambda_j(1-\lambda_j) > 0$*, then*

$$S_f(x_f^k, x^*) + S_g(x_g^k, x^*) \leq \frac{\|z^0 - z^*\|^2}{4\gamma\sqrt{\underline{\tau}(k+1)}}, \qquad (33)$$

and thus

$$S_f(x_f^k, x^*) + S_g(x_g^k, x^*) = o\left(\frac{1}{\sqrt{k+1}}\right). \qquad (34)$$

It is not clear whether the "best iterate" convergence results of Theorem can be improved to a convergence rate for the entire sequence because the value $S_f(x_f^k, x)$ and $S_g(x_g^k, x)$ are not necessarily monotonic.

Lipschitz derivatives

Assumption 4 The gradient of at least one of the functions f and g is Lipschitz.

- In general, we can only deduce the summability and not the monotonicity of the objective errors in Problem 1, we can only show that the smallest objective error after k iterations is of order $o(1/(k+1))$.
- If $\lambda_k \equiv 1/2$, the implicit stepsize parameter γ is small enough, and the gradient of g is $(1/\beta)$-Lipschitz, we show that a sequence that dominates the objective error is monotonic and summable, and deduce a convergence rate for the entire sequence.

Theorem 2 (Best iterate convergence under Lipschitz assumption)

Let $z \in \mathcal{H}$, let $z^+ = (T_{PRS})_\lambda z$, let z^* be a fixed point of T_{PRS}, and let $x^* = prox_{\gamma g}(z^*)$. Suppose that $\underline{\tau} = \inf_{j \geq 0} \lambda_j(1 - \lambda_j) > 0$, and let $\underline{\lambda} = \inf_{j \geq 0} \lambda_j$. If ∇f (respectively ∇g) is $(1/\beta)$-Lipschitz, and $x^k = x_g^k$ (respectively $x^k = x_f^k$), then

$$f(x^{k_{best}}) + g(x^{k_{best}}) - f(x^*) - g(x^*) = o\left(\frac{1}{k+1}\right). \qquad (35)$$

The main conclusion of Theorem is that as long as $\underline{\tau} > 0$, the "best" relaxed PRS iterate convergence with rate $o(1/(k+1))$ for any input parameters. This result is in stark contrast to the FBS algorithm, which may fail to converge if γ is too large.

Assumption 5: The function g is differentiable on dom $(f) \cap$ dom (g), the gradient ∇g is $(1.\beta)$-Lipschitz, and the sequence of relaxation parameters $(\lambda_j)_{j\geq 0}$ is constant and equal to $1/2$.
With this assumptions, we will show that for a special choice of θ^*(Lemma 5) and for γ small enough, the following sequence is monotonic and summable(Proposition 7 and 9):

$$\left(2\gamma\Big(f(x_f^j)+g(x_f^j)-f(x)-g(x)\Big)+\theta^*\gamma^2\|\nabla g(x_g^{j+1})-\nabla g(x_g^j)\|^2+\frac{(1-\theta^*)\gamma}{\beta^2} \tag{36}$$

We then use Lemma 1 to deduce
$f(x_f^j)+g(x_f^j)-f(x)-g(x)=o(1/(k+1))$.

Lemma 3 (Extra contraction of derivative operaots)

Suppose that ∇g is $(1/\beta)$-Lipschitz, and let $x, y \in \mathcal{H}$. If $x^+ = prox_{\gamma g}(x)$ and $y^+ = prox_{\gamma f}(y)$, then

$$\|\nabla g(x^+) - \nabla g(y^+)\|^2 \leq \frac{1}{\gamma^2 + \beta^2}\|x - y\|^2. \qquad (37)$$

Corollary 4 (Joint descent theorem)

If g is differentiable and ∇g is $(1/\beta)$-Lipschitz, then for all pairs $x, y \in dom\,(g) \cap dom\,(f)$. points $z \in dom\,(g)$, and subgradients $\widetilde{\nabla} f \in \partial f(x)$, we have

$$f(x)+g(x) \leq f(y)+g(y)+ < x-y, \nabla g(z)+\widetilde{\nabla} f(x) > + \frac{1}{2\beta}\|z-x\|^2. \qquad (38)$$

Lemma 5 (maximizing γ range)

Let $\beta > 0$, and let

$$\kappa := \sup \left\{ \frac{\gamma}{\beta} | \gamma > 0, \theta \in [0,1], \theta\gamma^2 \leq \left(2\gamma\beta - \frac{\gamma^3}{\beta}\right), \frac{(1-\theta)\gamma^2}{\beta^2} \leq 1 \right\}. \quad (39)$$

Then κ is the positive root of $x^3 + x^2 - 2x - 1$. Therefore, $(\gamma^*, \beta^*) = (\kappa\beta, 1 - 1/\kappa^2)$.

Lemma 6 (Gradient sum bounded)

For all $\gamma > 0$

$$\sum_{i=0}^{\infty} \|\nabla g(x_g^i) - \nabla g(x_g^{i+1})\|^2 \leq \frac{1}{\gamma^2 + \beta^2} \|z^0 - z^8\|^2. \quad (40)$$

Lemma 7 (Summability)

If $\gamma < \kappa\beta$, choose $\theta = \theta^*$ as in Lemma 5; otherwise, set $\theta = 1$. Then

$$\sum_{i=0}^{\infty} \Big(2\gamma(f(x_f^k) + g(x_f^k) - f(x^*) - g(x^*)) + \theta\gamma^2 \|\nabla g(x_g^{k+1}) - \nabla g(x_g^k)\|^2 + \frac{(1-\theta)\gamma^2}{\beta^2}\|x_g^{k+1} - x_g^k\|^2 \Big) \le \tag{41}$$

Theorem 8 (Differentiable function convergence rate)

Let $\rho \approx 2.2056$ *be the positive real root of* $x^3 - 2x^2 - 1$*. Then*

$$f(x_f^{k_{best}}) + g(x_f^{k_{best}}) - f(x^*) - g(x^*)$$
$$\leq \frac{1}{2\gamma(k+1)} \begin{cases} \|x_g^0 - x^*\|^2, & \text{if} \gamma < \rho\beta; \\ |x_g^0 - x^*\|^2 + \frac{1}{\beta^2+\gamma^2}\left(\frac{\gamma^3}{\beta} - 2\gamma\beta - \beta^2\right)\|z^0 - z^*\|^2, & \text{otherwise}. \end{cases} \quad (42)$$

and

$$f(x_f^{k_{best}}) + g(x_f^{k_{best}}) - f(x^*) - g(x^*) = o\left(\frac{1}{k+1}\right). \quad (43)$$

Furthermore, if $\gamma < \kappa\beta$*, then*

$$f(x_f^k) + g(x_f^k) - f(x^*) - g(x^*) \leq \frac{\|x_g^0 - x^*\|^2}{2\gamma(k+1)} \quad (44)$$

and

$$f(x_f^k) + g(x_f^k) - f(x^*) - g(x^*) = o\left(\frac{1}{\ }\right) \quad (45)$$

Theorem 9 (Differentiable function FPR rate)

Suppose that $\gamma < \kappa\beta$. *Then for all* $k \geq 1$, *we have*

$$\|z^k - z^{k+1}\|^2 \leq \frac{\beta^2 \|x_g^0 - x^*\|^2}{k^2(1 + \gamma/\beta^2)(\beta^2 - \gamma^2/\kappa^2)} \tag{46}$$

and

$$\|z^k - z^{k+1}\|^2 = o\left(\frac{1}{k^2}\right). \tag{47}$$

Linear convergence

Assumption The gradient of at least one of the functions f and g is Lipschitz, and at least one of the functions f and g . In symbols: $(\mu_f + \mu_g)(\beta_f + \beta_g) > 0$. Linear convergence of relaxed PRS is expected whenever Assumption is true. In addition, by the strong convexity of $f + g$, the minimizer of Problem 1 is unique.

Theorem 10 (Consequences of linear converge)

Let $(C_j)_{j\geq 0} \subset [0,1]$ *be a positive scalar sequence, and suppose that for all* $k \geq 0$*,*

$$\|z^{k+1} - z^*\| \leq C_k \|z^k - z^*\|. \tag{48}$$

Fix $k \geq 1$*. Then*

$$\|x_g^k - x^*\|^2 + \gamma^2 \|\widetilde{\nabla} g(x_g^k) - \widetilde{\nabla} g(x^*)\|^2 \leq \|z^0 - z^*\|^2 \prod_{i=0}^{k-1} C_i^2; \tag{49}$$

$$\|x_f^k - x^*\|^2 + \gamma^2 \|\widetilde{\nabla} f(x_f^k) - \widetilde{\nabla} f(x^*)\|^2 \leq \|z^0 - z^*\|^2 \prod_{i=0}^{k-1} C_i^2. \tag{50}$$

If $\lambda < 1$, then the FPR rate holds:

$$\|(T_{PRS})_\lambda z^k - z^k\| \le \sqrt{\frac{\lambda}{1-\lambda}\|z^0 -^* \|^2 \prod_{i=0}^{k-1} C_i^2}. \tag{51}$$

Consequently, if the gradient ∇f (respectively ∇g), is $(1/\beta)$-Lipschitz and $x^k = x_g^k$ (respectively $x^k = x_f^k$), then

$$\begin{aligned} & f(x^k) + g(x^k) - f(x^*) - g(x^*) \\ \le & \frac{\|z^0 - z^*\|^2}{\gamma} \prod_{i=0}^{k-1} C_i^2 \times \begin{cases} 1, & \text{if} \gamma \le \beta; \\ 1 + \frac{\gamma-\beta}{2\beta}, & \text{otherwise}. \end{cases} \end{aligned} \tag{52}$$

At least one of the functions f and g will carry both regularity properties. In symbols: $\mu_f\beta_f + \mu_g\beta_g > 0$.

Theorem 11 (Linear convergence with regularity of g)

Let z^* ne a fixed point of T_{PRS}, let $x^* = prox_{\gamma g}(z^*)$, and suppose that $\mu_g\beta_g > 0$. For all $\lambda \in [0,1]$, let $C(\lambda) := \left(1 - 4\gamma\lambda\mu_g/(1+\gamma/\beta_g)^2\right)^{1/2}$. Then for all $k \geq 0$,

$$\|z^{k+1} - z^*\| \leq C(\lambda_k)\|z^k - z^*\|. \tag{53}$$

Remark

- For all $\lambda \in [0, 1]$, the constant $C(\lambda)$ is minimal when $\gamma = \beta_g$, i.e. $C(\lambda) = (1 - \lambda_k \mu_g \beta_g)^{1/2}$.
- Furthermore, for any choice of γ, we have the bound $C(1) \leq C(\lambda)$.

The following theorem deduces linear convergence of relaxed PRS whenever f carries both regularity properties. Note that linear convergence of the PRS algorithm ($\lambda_k \equiv 1$) does not follow.

Theorem 12 (Linear convergence with regularity of f)

Let z^* be a fixed point of T_{PRS}, let $x^* = prox_{\gamma g}(z^*)$, and suppose that $\nu_f \beta_f > 0$. For all $\lambda \in [0, 1]$, let

$$C(\lambda) := \left(1 - (\lambda/2)\min\left\{4\gamma\mu_f/(1+\gamma/\beta_f)^2, (1-\lambda)\right\}\right)^{1/2}. \quad (54)$$

Then for all $k \geq 0$,

$$\|z^{k+1} - z^*\| \leq C(\lambda_k)\|z^k - z^*\|. \quad (55)$$

Theorem 13 (Linear convergence: mixed case)

Let z^* be a fixed point of T_{PRS}, let $x^* = prox_{\gamma g}(z^*)$, and suppose that ∇g, (respectively ∇f), is $(1/\beta)$-Lipschitz and f, (respectively g), is μ-strongly convex. For all $\lambda \in [0,1]$, let $C(\lambda) := (1 - (4\lambda/3)\min\{\gamma, \mu, \beta/\gamma, (1-\lambda)\})^{1/2}$. Then for all $k \geq 0$,

$$\|z^{k+1} - z^*\|. \tag{56}$$