A brief report on the paper - Linear Convergence and Metric Selection for Douglas-Rachford Splitting and ADMM

Nishchal Hoysal G

April 20, 2022

1 Introduction

The paper at hand [1] considers optimization problems of the form

$$\min_{x \in \mathbb{R}^n} f(x) + g(x) \tag{1}$$

where, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ are closed, proper and convex functions. $f()$ is assumed to be differentiable.

2 Definitions

Suppose an operator $e : \mathbb{R}^n \rightarrow \mathbb{R}^n$.

1. e(.) is said to be monotone if

$$\forall x, y \in \mathbb{R}^n : \big(e(x) - e(y)\big)^T (x - y) \geq 0$$

2. e(.) is said to be m-strongly monotone if

$$\forall x, y \in \mathbb{R}^n : \big(e(x) - e(y)\big)^T (x - y) \geq m||x - y||_2^2$$

3. e(.) is said to be maximal monotone on dom(e), if it is monotone and is not a proper subset of any other monotone operator on dom(e).
4. e(.) is non-expansive if

$$\forall x, y \in \mathbb{R}^n : ||e(x) - e(y)||_2 \leq ||x - y||_2$$

5. e(.) is firmly non-expansive if it is 1-strongly monotone, i.e.,

$$\forall x, y \in \mathbb{R}^n : \big(e(x) - e(y)\big)^T (x - y) \geq ||x - y||_2^2$$

6. $$J_{ce} := (I + ce)^{-1}$$

   is called the c-resolvent of e(.), where I is the identity operator. 1-resolvent of e(.) is just called its resolvent.

3 Properties of monotone operators [2]

1. Let $C \subseteq \mathbb{R}^n$ be closed, bounded and convex. Let $U : C \to C$ be a non-expansive map. Then $T := I - U$ is a monotone operator on C

2. If F and G are monotone, then,

   (a) $F + G$ is monotone

   (b) αF is monotone $(\alpha > 0)$

   (c) F^{-1} is monotone

   (d) $T \in \mathbb{R}^{nxm}$, $T^T \circ F \circ T$ is monotone (on $\mathbb{R}^m$)

3. A monotone operator F is maximal monotone on $\mathbb{R}^n$ iff $Im(I + F) = \mathbb{R}^n$

4. F is maximal monotone iff it is a connected curve with no end points, with non-negative (may be infinite) slope

4 Properties of non-expansive operators [3]

1. All firmly non-expansive operators are non-expansive

2. Operator J is firmly non-expansive iff $2J - I$ is non-expansive

3. An operator is firmly non-expansive iff it is of the form $\frac{1}{2}\big(I + C\big)$ for some non-expansive operator C

4. J is firmly non-expansive iff $I - J$ is firmly non-expansive

5. J is non-expansive $\implies$ $-J$ is non-expansive

5 Results connecting monotonicity, non-expansiveness and c-resolvent [3]

1. Let $c > 0$. An operator T on $\mathbb{R}^n$ is monotone iff its c-resolvent, J_{cT} is firmly non-expansive. Furthermore, T is maximal monotone iff J_{cT} is firmly non-expansive and $dom(J_{cT}) = \mathbb{R}^n$.

2. An operator K is firmly non-expansive iff $K^{-1} - I$ is monotone. Furthermore, K is firmly non-expansive with full domain iff $K^{-1} - I$ is maximal monotone.

3. Given any maximal monotone operator T, $c > 0$ and $x \in \mathbb{R}^n$, we have $0 \in T(x)$ iff $J_{cT}(x) = x$

Theorem 1. *[4]Proposition 25.1 Let $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, $B : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ and let $\gamma > 0$. If A and B are monotone, then*

$$zer(A + B) = J_{\gamma B}(FixR_{\gamma A}R_{\gamma B})$$

where, $J_{\gamma B} = (I + \gamma B)^{-1}$ and $R_{\gamma A} = 2J_{\gamma A} - I$

Theorem 2. *[3] Let T be a maximal monotone operator on $\mathbb{R}^n$ and let $\{z_k\}_{k\geq 0}$ be such that*

$$z_{k+1} = (1 - \rho_k)z_k + \rho_k w_k, \ \forall k \geq 0$$

where

$$||w_k - (I + c_kT)^{-1}(z_k)|| \leq \epsilon_k \ \forall k \geq 0.$$

$$\{\epsilon_k\}_{k\geq 0}, \{\rho_k\}_{k\geq 0} \text{ and } \{c_k\}_{k\geq 0} \subseteq [0, \infty)$$

are sequences such that,

$$\sum_{k=0}^{\infty} \epsilon_k < \infty, \ \inf_{k\geq 0} \rho_k > 0, \ \sup_{k\geq 0} \rho_k < 2 \text{ and } \inf_{k\geq 0} c_k > 0.$$

Such $\{z_k\}_{k\geq 0}$ is said to conform to the generalized proximal point algorithm. Then if T possesses any zero, $\{z_k\}_{k\geq 0}$ converges to a zero of T. If T has no zero, the $\{z_k\}_{k\geq 0}$ is unbounded.

Theorems 1 and 2 pave way to the proximal algorithms to get zeros of operators.

6 Proximal operator and splitting algorithms

Given a function $f : D \to (-\infty, \infty]$, the proximal mapping of f is the operator given by

*

$$P_f(x) := \underset{u\in D}{argmin}\{f(u) + \frac{1}{2}||u - x||^2\} \ \forall x \in D \tag{2a}$$

For a closed, proper and convex function $f(.)$, since the objective function is strongly convex, $\forall x \in D$, $P_f(x)$ is a singleton.

It can be shown that the proximal operator of a function f, $P_f()$, is firmly non-expansive. Hence the operator $R_f := (2P_f - I)$ is non-expansive. R_f is called the reflexive operator of f.

Many optimization problems can be converted to the form

$$\min_{x\in\mathbb{R}^n} m(x)(:= f(x) + g(x)) \tag{3}$$

where, $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are closed, proper and convex functions and $f()$ is differentiable.

Since $f()$ and $g()$ are closed, proper and convex, we can write

$$\partial_m(x) = \partial_f(x) + \partial_g(x) \ \forall x \in dom(m).$$

Note that sub-gradient operators, $\partial_f(x)$ and $\partial_g(x)$ are maximal monotone. Also, for closed, proper and convex $f()$, it can be shown that $P_f(z) = (I + \partial_f)^{-1}(z)$, i.e, $P_f()$ is a resolvent of sub-gradient operator of $f()$.

Constructing Cayley operators $R_{\gamma g} := 2P_{\gamma g} - I$ and $R_{\gamma f} := 2P_{\gamma f} - I$ ($\gamma > 0$, as a parameter) and using theorem 1, x^* is a minimizer of 3 iff,

$$z = R_{\gamma g} R_{\gamma f} z \text{ and } x^* = P_{\gamma f}(z) \tag{4}$$

Douglas-Rachford (DR) algorithm, i.e.,

$$z_{k+1} = ((1-\alpha)I + \alpha R_{\gamma g} R_{\gamma f}) z_k, \ z_0 \in \mathbb{R}^n \tag{5}$$

can be shown to converge to $zer(f() + g())$ using theorem 2 for $\alpha \in (0, 1)$.

This algorithm 5 can be split and represented as

$$x_k = P_{\gamma f}(z_k) \tag{6a}$$

$$y_k = P_{\gamma g}(2x_k - z_k) \tag{6b}$$

$$z_{k+1} = z_k + 2\alpha(y_k - x_k) \tag{6c}$$

$$z_0 \in \mathbb{R}^n \tag{6d}$$

The paper [1] establishes linear convergence of DR-algorithm on the problem 3 under the assumptions that,

1. f and g are closed, proper and convex
2. f is σ-strongly convex and β-smooth

Theorem 2 of [1] gives the upper bound rate of convergence as $|1 - \alpha| + \alpha\delta$, when $\alpha \in (0, (1/1 + \delta))$ where,

$$\delta = max \left\{ \frac{\gamma\beta - 1}{\gamma\beta + 1}, \frac{1 - \gamma\sigma}{1 + \gamma\sigma} \right\}$$

with $\gamma \in (0, \infty)$, a lemma in appendix A of [1] shows that $\delta \in (0, 1)$

It is also shown that the optimal values of α and γ are 1 and $1/\sqrt{\beta\sigma}$ respectively. The optimal convergence rate is derived as $\frac{\kappa - 1}{\kappa + 1}$ where $\kappa = \sqrt{\frac{\beta}{\sigma}}$.

An example presented in section 3(B) of [1] shows that the derived convergence rate is tight.

7 Alternating Direction Method of Multipliers (ADMM)

ADMM is generally used to solve the optimization problems of the form

$$\min_{Ax+By=c} f(x) + g(y) \tag{7}$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ and A is a surjective map.

The ideas probably stem from dual gradient methods with augmented lagrangian [5]. The ADMM algorithm with scaled-variable and augmented lagrangian is as given below.

$$x_{k+1} = argmin_x\{f(x) + 2\gamma||Ax + By_k - c + u_k||^2\} \tag{8a}$$

$$x^A_{k+1} = 2\alpha Ax_{k+1} - (1 - 2\alpha)(By_k - c) \tag{8b}$$

$$y_{k+1} = argmin_y\{g(y) + \frac{\gamma}{2}||x^A_{k+1} + By - c + u_k||^2\} \tag{8c}$$

$$u_{k+1} = u_k + x^A_{k+1} + By_k - c \tag{8d}$$

$$y_0 \text{ and } z_0 \text{ random initializers} \tag{8e}$$

Remark 1. *It has been shown that solving the primal problem 7 using ADMM is equivalent to solving its dual problem using DR-algorithm. Hence the analysis of ADMM in [1], the authors consider DR-algorithm applied on the dual problem.*

7.1 Dual problem

Consider the problem

$$\min_{Ax=b} f(x) \tag{9}$$

Formulating the dual of this problem, lagrangian is $\mathcal{L}(x, u) = f(x)+u^T(Ax-b)$. The dual problem is

$$\max_u \min_x f(x) + u^T(Ax - b) \tag{10}$$

Clearly, $f^*(-A^T u) = min_x f(x) + u^T Ax$ where f^* is the convex conjugate of f.

Therefore, the dual problem of 7 can be written as,

$$\min_u d_1(u) + d_2(u) \tag{11}$$

where, $d_1(u) = f^*(-A^T u) + c^T u$ and $d_2(u) = g^*(-Bu)$.

Remark 2. *Let f be a closed, proper and convex function. Then its convex conjugate f^*, is also closed, proper and convex with $f = (f^*)^*$. Moreover, if f is σ-strongly convex, f^* is $1/\sigma$-smooth and if f is β-smooth, then f^* is $1/\beta$-strongly convex.*

By the results in 1 and 2, the linear convergence results in DR-algorithm can be translated to ADMM-algorithm as follows

For $\gamma \in (0, \infty)$ and $\alpha \in (0, 1/(1+\hat{\delta}))$, the ADMM algorithm in 8e converges at least with rate $|1-\alpha| + \alpha\hat{\delta}$ where,

$$\hat{\delta} = max\left(\frac{\gamma\hat{\beta}-1}{1+\gamma\hat{\beta}}, \frac{1-\gamma\hat{\sigma}}{1+\gamma\hat{\sigma}}\right)$$

$\hat{\beta} = \left(\frac{||A^T||^2}{\sigma}\right)$ and $\hat{\sigma} = \left(\frac{\theta^2}{\beta}\right)$ for some $\theta > 0$ such that $||A^T u|| \geq \theta ||u||\ \forall u$.

Similar to DR-algorithm, the optimal values of α and γ here are 1 and $\frac{1}{\sqrt{\hat{\beta}\hat{\sigma}}}$ respectively. Also, the optimal convergence rate is derived as $\frac{\hat{\kappa}-1}{\hat{\kappa}+1}$ where $\hat{\kappa} = \sqrt{\frac{\hat{\beta}}{\hat{\sigma}}}$

A similar example has also been presented in the paper to show that the convergence rate bound derived is tight.

8 Metric selection

From previous discussions, it is clear that the metric-space in which the minimization problem is defined has an effect on the rate of convergence for ADMM and DR-algorithms. In this regard, section VI of [1] discusses the optimal metric selection. As discussed in the previous section, the optimal convergence rate depends on the smoothness and stong convexity parameters of function f in problem 7.

In general, assuming that the function f is 1-strongly convex with an inner product metric defined by H and 1-smooth with an inner product metric L, then an implication of proposition 5 in [1] is that an inner product metric defined by E maximizes the rate of convergence if E is the minimizer of the problem

$$\min_{E \in \mathbb{S}^n_{++}} \frac{\lambda_{max}(EAH^{-1}A^TE^T)}{\lambda_{max}(EAL^{-1}A^TE^T)} \tag{12}$$

Section 6 of [6] discusses an SDP formulation which solves the problem 12 when $H = L$ as described below.

Theorem 3. *Suppose that $Q \in \mathbb{S}^m_{++}$. Then a matrix $E \in \mathbb{S}^m_{++}$ that minimizes the ratio $\frac{\lambda_{max}(EQE^T)}{\lambda_{min}(EQE^T)}$ can be computed by solving the convex SDP*

*

$$minimize\ t \tag{13a}$$

$$subject\ to\ tQ \geq L \tag{13b}$$

$$Q \leq L \tag{13c}$$

$$L \in \mathbb{S}^m_{++} \tag{13d}$$

where, $L = (E^TE)^{-1}$.

The paper restricts to diagonal matrices E.

With such an inner product metric E, the problem 7 can be solved with an maximum convergence rate by solving the preconditioned following problem.

$$\min_{E(Ax+By)=Ec} f(x) + g(y) \tag{14}$$

9 Simulations and results

To check the validity of results (on optimal value of γ parameter) in the paper by simulations, the following problem was considered.

(Note: The problem here is slightly different from the problem considered in the paper. This was done to ensure the knowledge of closed form solutions to the intermediate steps of ADMM so that the computations can be done with less time.)

$$\min \frac{1}{2}||Ax - b||_2^2 + w||x||_1 \tag{15a}$$

where $x \in \mathbb{R}^{20}$, $A \in \mathbb{R}^{300x200}$ sparse with average 10 non-zero entries per row and $b \in \mathbb{R}^{300}$. Each non-zero element in A and b is drawn from normal distribution (mean 0, variance 1). w was drawn from uniform distribution on $[0, 1]$.

The value of γ was varied between $[10^{-3}\gamma^*, 10^2\gamma^*]$ where, γ^* is the theoretical optimal for γ.

It can be shown that the objective function in 15a is $||A||$-smooth and $\lambda_{min}(A^TA)$-strongly convex. Hence, the optimal metric E is obtained by solving problem 13d for diagonal L.

ADMM method was used on problem 15a to analyse the results in the paper.

Figure **??** shows the plot of γ vs number of iterations to achieve a relative accuracy of 10^{-4}. From the figure, it can be seen that preconditioning works better.

References

[1] P. Giselsson and S. Boyd, "Linear convergence and metric selection for douglas-rachford splitting and admm," 2014. [Online]. Available: https://arxiv.org/abs/1410.8479

[2] S. Boyd, "Lecture slides from stanford university on monotone operators," https://web.stanford.edu/class/ee364b/lectures/monotone_slides.pdf, accessed: 2022-04-20.

[3] J. Eckstein and D. Bertsekas, "On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 04 1992.

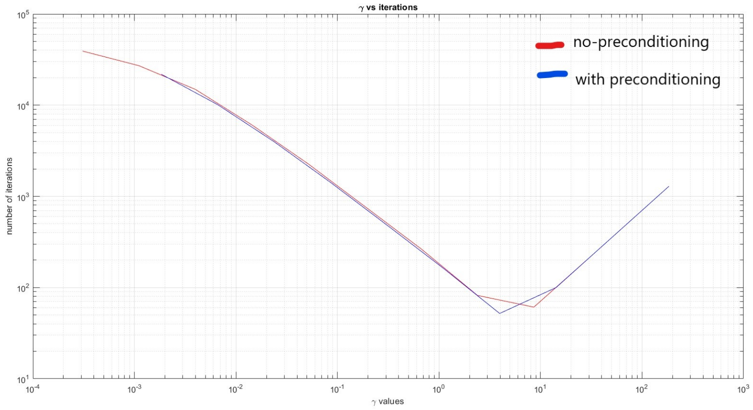

Figure 1: The variation of number of iterations taken by ADMM to attain a relative accuracy of 10^{-4} against different values of γ with and without preconditioning.

[4] H. H. Bauschke, P. L. Combettes *et al.*, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011, vol. 408.

[5] B. Poczos and R. Tibshirani, “Lecture slides from cmu on convex optimization,” https://www.stat.cmu.edu/~ryantibs/convexopt-F13/lectures/23-dual-meth.pdf, accessed: 2022-04-20.

[6] P. Giselsson and S. Boyd, “Metric selection in fast dual forward–backward splitting,” *Automatica*, vol. 62, pp. 1–10, 12 2015.