Anderson Accelerated Douglas–Rachford Splitting

by Anqi Fu, Junzi Zhang and Stephen Boyd

CHEN Li

Friday Seminar
Dec 11, 2020

Outline

Problem Setting

- Problem of prox-affine form:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & Ax = b \end{array} \tag{1}$$

where

$$x = [x_1; x_2; \cdots; x_N] \in \mathbb{R}^{n_1+n_2+\cdots n_N}$$

$$f(x) = \sum_{i=1}^{N} f_i(x_i) \text{ is proper, closed and convex}$$

$$A = [A_1, A_2, \cdots, A_N] \in \mathbb{R}^{m\times(n_1+n_2+\cdots n_N)}$$

Problem Setting

- Problem of prox-affine form:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & Ax = b \end{array} \tag{1}$$

where

$$x = [x_1; x_2; \cdots; x_N] \in \mathbb{R}^{n_1+n_2+\cdots n_N}$$
$$f(x) = \sum_{i=1}^{N} f_i(x_i) \text{ is proper, closed and convex}$$
$$A = [A_1, A_2, \cdots, A_N] \in \mathbb{R}^{m\times(n_1+n_2+\cdots n_N)}$$

- Assumption: only accessible proximal map oracle

$$\mathbf{prox}_{tf}(\cdot) \text{ and } \Pi_{\{x:Ax=b\}}(\cdot)$$

Basics and Notations

- Problem of prox-affine form:

$$\begin{array}{ll} \min\limits_x & f(x) \\ \text{s.t.} & Ax = b \end{array}$$

- Optimality condition:

$$\begin{aligned} & Ax = b \\ & \frac{v - x}{t} + A^T \lambda = 0 \\ & x_i = \mathbf{prox}_{tf_i}(v_i), \ \forall i = 1, ..., N \end{aligned}$$

  are sufficient. They are necessary if $\mathbf{relint\ dom}\, f \cap \{x : Ax = b\} \neq \emptyset$ (Slater's condition).

- Residual $r = (r_{\text{prim}}, r_{\text{dual}})$

$$\begin{aligned} r_{\text{prim}} &= Ax - b \\ r_{\text{dual}} &= \tfrac{v-x}{t} + A^T \lambda \end{aligned} \tag{2}$$

Douglas-Rachford Splitting (DRS)

For convex composite problem

$$\min_x f(x) + g(x) \tag{3}$$

- In iteration $k = 1, 2, ...,$ DRS runs
  - $x_{k+1} = \mathbf{prox}_{tf}(y_k)$
  - $y_{k+1} = y_k + \mathbf{prox}_{tg}(2x_{k+1} - y_k) - x_{k+1}$
- Essentially, $y_{k+1} = F(y_k)$ is a fixed point iteration (FPI) where

$$F = \frac{(2\mathbf{prox}_{tg} - I) \circ (2\mathbf{prox}_{tf} - I) + I}{2} \tag{4}$$

  is firmly nonexpansive[1].
- y is a fixed point of F if and only if $x = \mathbf{prox}_{tf}(y)$ satisfies $0 \in \partial f(x) + \partial g(x)$.

[1] Jonathan Eckstein and Dimitri P Bertsekas. "On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators". In: *Mathematical Programming* 55.1-3 (1992), pp. 293–318.

Vanilla DRS Algorithm for Problem (1)

Algorithm 2.1 Douglas–Rachford Splitting (DRS)

Input: initial point v^0, penalty coefficient $t > 0$.
for $k = 1, 2, \dots$ **do**
    $x^{k+1/2} = \mathbf{prox}_{tf}(v^k)$
    $v^{k+1/2} = 2x^{k+1/2} - v^k$
    $x^{k+1} = \Pi(v^{k+1/2})$
    $v^{k+1} = v^k + x^{k+1} - x^{k+1/2}$
end for

- $\Pi(v^{k+1/2}) = v^{k+1/2} - A^T(AA^T)^\dagger(Av^{k+1/2} - b)$ is the projection on affine space $\{x : Ax = b\}$
- $F_{DRS} = I + \Pi \circ (2\mathbf{prox}_{tf} - I) - \mathbf{prox}_{tf}$
- Residuals: $r^k_{\text{prim}} = Ax^{k+1/2} - b$ and $r^k_{\text{dual}} = \frac{v^k - x^{k+1/2}}{t} + A^T\lambda^k$
- Dual solution λ^k can be chosen to minimize $\|r^k_{\text{dual}}\|_2$
- v^k converges globally and sublinearly to a fixed point of F

Andersen Acceleration (AA)

- Focus on original type-II AA[2]

---

[2] Donald G Anderson. "Iterative procedures for nonlinear integral equations". In: *Journal of the ACM (JACM)* 12.4 (1965), pp. 547–560.

Andersen Acceleration (AA)

- Let $G(v) = v - F(v)$ be the residual and $M^k = \min(M_{max}, k)$ be the memory size.
- At iteration k, type-II AA stores the most recent $M^k + 1$ iterates $[v^k, ..., v^{k-M^k}]$ and replace $v^{k+1} = F(v^k)$ by

$$v^{k+1} = \sum_{j=0}^{M^k} \alpha_j^k F(v^{k-M^k+j}) \tag{5}$$

where α_j^k is determined by solving

$$\begin{aligned} \min_{\alpha^k} \quad & \| \textstyle\sum_{j=0}^{M^k} \alpha_j^k G(v^{k-M^k+j}) \|_2^2 \\ \text{s.t.} \quad & \textstyle\sum_{j=0}^{M^k} \alpha_j^k = 1 \end{aligned} \tag{6}$$

- Type-II AA can be regarded as an **extrapolation**

[2] **Anderson**, "Iterative procedures for nonlinear integral equations".

Outline

Adaptive Regularization

- Challenge: Type-II AA suffers from instability
- Simple idea: Add regularization on α^k.
- Solution:
  - Define

$$g^k = G(v^k), \quad y^k = g^{k+1} - g^k, \quad s^k = v^{k+1} - v^k$$
$$Y_k = [y^{k-M^k}, ..., y^{k-1}], \quad S_k = [s^{k-M^k}, ..., s^{k-1}]$$

Then the problem (6) can be rewritten as

$$\min_{\gamma^k} \|g^k - Y_k \gamma^k\|_2^2 \tag{7}$$

where $\gamma^k = [\gamma_0^k; ...; \gamma_{M^k-1}^k]$ and

$$\alpha_0^k = \gamma_0^k, \quad \alpha_i^k = \gamma_i^k - \gamma_{i-1}^k \text{ for } i = 1, ..., M^k - 1, \quad \alpha_{M_k}^k = 1 - \gamma_{M^k-1}^k$$

  - Add ℓ_2-regularization term scaled by norm of S_k and Y_k.

$$\min_{\gamma^k} \|g^k - Y_k \gamma^k\|_2^2 + \eta \left(\|S_k\|_F^2 + \|Y_k\|_F^2\right) \|\gamma^k\|_2^2 \tag{8}$$

  - Larger norm (less stable) $\Longrightarrow$ less AA

Safeguard Step

- To achieve global convergence, a safeguard step is needed
- Check whether the current residual norm is small enough,
  - If true, then takes AA update in the next $R - 1$ iterations
  - If false, then takes vanilla FPI
- $R \in \mathbb{Z}_{++}$ is used to control the degree of safeguarding: Larger R means more aggressive AA

Safeguard Step

- To achieve global convergence, a safeguard step is needed
- Check whether the current residual norm is small enough,
  - If true, then takes AA update in the next $R-1$ iterations
  - If false, then takes vanilla FPI
- $R \in \mathbb{Z}_{++}$ is used to control the degree of safeguarding: Larger R means more aggressive AA
- Safeguard condition:

$$\|g^k\|_2 = \|G_{DRS}(v^k)\|_2 \leq D\|g^0\|_2(n_{AA}/R+1)^{-(1+\epsilon)}$$

where $D > 0$, $\epsilon > 0$ and n_{AA} is the cumulative number of AA updates.

A2DR Algorithm

Algorithm 1: Anderson Accelerated Douglas-Rachford (A2DR)

Input: initial v^0, penalty $t > 0$, regularization $\eta > 0$, safeguarding $D > 0$, $\epsilon > 0$, $R \in \mathbb{Z}_{++}$, max memory $M_{max} \in \mathbb{Z}_+$

Initialize $n_{AA} = 0$, $R_{AA} = 0$, $I_{safe} = \texttt{True}$;

for $k = 1,2,\ldots$ **do**

    `# Memory Update`;

    Set $M^k = \min(M_{max}, k)$, compute $v_{DRS}^{k+1} = F_{DRS}(v^k)$, $g^k = v^k - v_{DRS}^{k+1}$, update Y_k and S_k;

    `# Adaptive Regularization`;

    Solve problem (8) and compute α^k;

    Compute AA candidate $v_{AA}^{k+1} = \sum_{j=0}^{M^k} \alpha_j^k v_{DRS}^{k-M^k+j}$;

    `# Safeguard`;

    if $I_{safe} = \mathit{True}$ *or* $R_{AA} \geq R$ **then**

        if $\|g^k\|_2 = \|G_{DRS}(v^k)\|_2 \leq D\|g^0\|_2(n_{AA}/R + 1)^{-(1+\epsilon)}$ **then** $v^{k+1} = v_{AA}^{k+1}$, $n_{AA} = n_{AA} + 1$, $I_{safe} = \texttt{False}$, $R_{AA} = 1$;

        else $v^{k+1} = v_{DRS}^{k+1}$, $R_{AA} = 1$;

    else

        $v^{k+1} = v_{AA}^{k+1}$, $n_{AA} = n_{AA} + 1$, $R_{AA} = R_{AA} + 1$

    end

    Terminate and output $x^{k+1/2}$ if $\|r^k\|_2 \leq \epsilon_{tol} = \epsilon_{abs} + \epsilon_{rel}\|r^o\|_2$

end

Outline

Infeasibility and Unboundedness

Proposition 1

Let f^* be the conjugate function of f, then

(i) *If* $\mathbf{dist}\left(\mathbf{dom}\, f, \{x : Ax = b\}\right) > 0$, *then the problem* (1) *is infeasible.*

(ii) *If* $\mathbf{dist}\left(\mathbf{dom}\, f^*, \mathbf{range}\left(A^T\right)\right) > 0$, *then the problem* (1) *is unbounded.*

- We call the problem (primal) strongly infeasible if (i) holds and dual strongly infeasible if (ii) holds. In either case, we call it pathological, otherwise it is called solvable.[3]
- Proof. Fenchel duality and Lemma 1 in[4]

[3] Yanli Liu, Ernest K Ryu, and Wotao Yin. "A new use of Douglas–Rachford splitting for identifying infeasible, unbounded, and pathological conic programs". In: *Mathematical Programming* 177.1-2 (2019), pp. 225–253.

[4] Ernest K Ryu, Yanli Liu, and Wotao Yin. "Douglas–Rachford splitting and ADMM for pathological convex optimization". In: *Computational Optimization and Applications* 74.3 (2019), pp. 747–778.

Convergence Results

Theorem 1 (Solvable case)

Suppose the problem (1) *is solvable. Then for any initialization* v^0, *hyperparameters* $\eta > 0, D > 0, \epsilon > 0, R \in \mathbb{Z}_{++}, M_{max} \in \mathbb{Z}_+$, *we have*

$$\liminf_{k\to\infty} \|r^k\|_2 = 0 \tag{9}$$

and the AA candidates are adopted infinitely often. Additionally, if F_{DRS} *has a fixed point,* v^k *converges to a fixed-point of* F_{DRS} *and* $x^{k+1/2}$ *converges to a solution of problem* (1) *as* $k \to \infty$.

Theorem 2 (Pathological case)

Suppose the problem (1) *is pathological. Then for any initialization* v^0, *hyperparameters* $\eta > 0, D > 0, \epsilon > 0, R \in \mathbb{Z}_{++}, M_{max} \in \mathbb{Z}_+$, *the difference* $v^k - v^{k+1}$ *converges to some nonzero vector* $\delta v \in \mathbb{R}^n$. *Furthermore, if* $\lim_{k\to\infty} Ax^{k+1/2} = b$, *then problem* (1) *is unbounded and* $\|\delta v\|_2 = t\mathbf{dist}\left(\mathbf{dom}\, f^*, \mathbf{range}\left(A^T\right)\right)$. *Otherwise, problem* (1) *is infeasible and* $\|\delta v\|_2 \geq \mathbf{dist}\left(\mathbf{dom}\, f, \{x : Ax = b\}\right)$ *with equality when the dual problem is feasible.*

A Useful Lemma for Stability Characterization

Note the solution to adaptive regularization problem (8) is

$$\gamma^k = (Y_k^T Y_k + \eta(\|S_k\|_F^2 + \|Y_k\|_F^2) I)^{-1} Y_k^T g^k$$

Hence the AA candidate is

$$v^{k+1} = v^k - H_k g^k$$

where

$$H_k = I + (S_k - Y_k)(Y_k^T Y_k + \eta(\|S_k\|_F^2 + \|Y_k\|_F^2) I)^{-1} Y_k^T$$

Lemma 1

The matrix H_k, $k \geq 1$ satisfy $\|H_k\|_2 \leq 1 + 2/\eta$.

Proof.

$$\begin{aligned} \|H_k\|_2 \quad & \leq 1 + \tfrac{\|S_k - Y_k\|_2 \|Y_k\|_2}{\eta(\|S_k\|_F^2 + \|Y_k\|_F^2)} \leq 1 + \tfrac{\|S_k - Y_k\|_F \|Y_k\|_F}{\eta(\|S_k\|_F^2 + \|Y_k\|_F^2)} \\ & \leq 1 + \tfrac{\|S_k\|_F \|Y_k\|_F + \|Y_k\|_F^2}{\eta(\|S_k\|_F^2 + \|Y_k\|_F^2)} \leq 1 + \tfrac{2}{\eta} \end{aligned}$$

A Lemma Connecting FPI to Residuals in DRS

Lemma 2

Suppose that $\liminf_{j\to\infty} \|v^j - F_{DRS}(v^j)\|_2 \le \epsilon$ *for any* $\epsilon > 0$, *then*

$$\liminf_{j\to\infty} \|r^j_{\mathrm{prim}}\|_2 \le \|A\|_2 \epsilon, \quad \liminf_{j\to\infty} \|r^j_{\mathrm{dual}}\|_2 \le \epsilon / t$$

Proof. We have the following fact from DRS iterations.

$$\begin{aligned}
x^{j+1/2} &= \mathbf{prox}_{tf}\left(v^j\right) & &\Longrightarrow \frac{v^j - x^{j+1/2}}{t} = g^j \in \partial f(x^{j+1/2}) \\
v^{j+1/2} &= 2x^{j+1/2} - v^j & & \\
x^{j+1} &= \Pi(v^{j+1/2}) & &\Longrightarrow x^{j+1} = v^{j+1/2} - A^T \tilde{\lambda}_j \\
& & & \tilde{\lambda}_j = (AA^T)^\dagger (Av^{j+1/2} - b) \\
& & & Ax^{j+1} = b \\
v^{j+1} &= v^j + x^{j+1} - x^{j+1/2} & &
\end{aligned}$$

Hence

$$r^j_{\mathrm{prim}} = Ax^{j+1/2} - b = A(x^{j+1/2} - x^{j+1}) = A(v^j - v^{j+1})$$

and $\|r^j_{\mathrm{prim}}\|_2 \le \|A\|_2 \|v^j - F_{DRS}(v^j)\|_2$

Proof of Lemma 2

Recall the facts:

$$\begin{aligned}
&x^{j+1/2} = \mathbf{prox}_{tf}\left(v^j\right) && \Longrightarrow \tfrac{v^j - x^{j+1/2}}{t} = g^j \in \partial f(x^{j+1/2}) \\
&v^{j+1/2} = 2x^{j+1/2} - v^j && \\
&x^{j+1} = \Pi(v^{j+1/2}) && \Longrightarrow x^{j+1} = v^{j+1/2} - A^T\tilde{\lambda}_j \\
& && \quad\ \tilde{\lambda}_j = (AA^T)^\dagger(Av^{j+1/2} - b) \\
& && \quad\ Ax^{j+1} = b \\
&v^{j+1} = v^j + x^{j+1} - x^{j+1/2} &&
\end{aligned}$$

- $r^j_{\text{dual}} = g^j + A^T\lambda^j$ where $\lambda^j \in \arg\min_\lambda \|g^j + A^T\lambda\|_2$.
- Note that

$$\begin{aligned}
g^j &= \tfrac{v^j - x^{j+1/2}}{t} = \tfrac{v^{j+1} - x^{j+1}}{t} \\
&= \tfrac{1}{t}(v^{j+1} - v^j + v^j - x^{j+1}) \\
&= \tfrac{1}{t}(v^{j+1} - v^j + v^j - x^{j+1/2} + x^{j+1/2} - v^{j+1/2} + v^{j+1/2} - x^{j+1}) \\
&= \tfrac{1}{t}(F_{DRS}(v^j) - v^j) + g^j + g^j + A^T\tilde{\lambda}_j/t
\end{aligned}$$

Hence

$$\|r^j_{\text{dual}}\|_2 \leq \|g^j + A^T\tilde{\lambda}^j/t\|_2 = \|\frac{1}{t}(v_j - F_{DRS}(v^j))\|_2$$

Proof of Theorem 1 (Solvable Case)

- Fact: The problem (1) is solvable $\iff \delta v^* = 0$[5] where δv^* is the infimal displacement vector of F_{DRS} defined as

$$\delta v^* = \Pi_{\overline{\mathbf{range}(I - F_{DRS})}}(0).$$

By definition we have $\|\delta v^*\|_2 = \inf_{v \in \mathbb{R}^n} \|v - F_{DRS}(v)\|_2$

- From Lemma 2, it is sufficient to show $\liminf_{k\to\infty} \|g^k\|_2 = 0$.

- For convenience, we define the following notations:
  - k_i: the initial iteration counts for accepting AA updates
  - l_i: the iteration counts for accepting DRS candidates

  Hence for each iteration k, either $k = k_i + K$ for some i and $0 \le K \le R-1$, or $k = l_i$ for some i.

[5] Heinz H Bauschke, Warren L Hare, and Walaa M Moursi. "On the range of the Douglas–Rachford operator". In: *Mathematics of Operations Research* 41.3 (2016), pp. 884–897.

Proof of Theorem 1 (Solvable Case cont.)

Recall the goal is to show $\liminf_{k\to\infty} \|g^k\|_2 = 0$.

- If the set of k_i is infinite, i.e., the AA candidate is adopted infinitely often, then

$$0 \le \liminf_{k\to\infty} \|g^k\|_2 \le \liminf_{i\to\infty} \|g^{k_i}\|_2 \le D\|g^0\|_2 \lim_{i\to\infty} (i+1)^{-(1+\epsilon)} = 0$$

  from the safeguard condition and $n_{AA}/R = i$ in iteration k_i.
- Otherwise, the AA candidate is never adopted after finite iterations and the algorithm becomes vanilla DRS. By Theorem 2 in[6], we have

$$\lim_{k\to\infty} g^k = \lim_{k\to\infty} v^k - v^{k+1} = \delta v^* = 0.$$

- Thus we always have $\liminf_{k\to\infty} \|g^k\|_2 = 0$.
- Finite k_i's cannot happen. Otherwise, $\lim_{k\to\infty} g^k = 0$ and n_{AA} is upper bounded imply the safeguard condition must be satisfied eventually, a contradiction.

[6] A Pazy. "**Asymptotic behavior of contractions in Hilbert space**". In: *Israel Journal of Mathematics* 9.2 (1971), pp. 235–240.

Proof of Theorem 1 (Solvable Case cont.)

Now suppose F_{DRS} has a fixed point v^*. We need to prove v^k converges to a fixed point of F_{DRS} and $x^{k+1/2}$ converges to a solution of problem (1).

Step 1 Prove $\|v^k - v^*\|_2$ is bounded.

Step 2 Prove $\lim_{k\to\infty} \|g^k\|_2 = 0$

Step 3 Prove $\|v^k - v^*\|_2$ is quasi-Fejérian.

- Finally, given results in Step 2 and Step 3, we can use Theorem 3.8 in[7] to conclude that $\lim_{k\to\infty} \|v^k - v^*\|_2$ exists and v^k converges to some fixed point of F_{DRS} (not necessarily v^*).
- The convergence of $x^{k+1/2}$ to a solution of problem (1) follows from continuity of proximal operators.

[7] **Patrick L Combettes**. "**Quasi-Fejérian analysis of some optimization algorithms**". In: *Studies in Computational Mathematics*. Vol. 8. Elsevier, 2001, pp. 115–152.

Proof of Theorem 1 (Solvable Case cont.)

For AA updates,

- we have

$$\begin{aligned}\|g^{k+1}\|_2 &= \|G_{DRS}(v^{k+1})\|_2 \\ &\leq \|G_{DRS}(v^{k+1}) - G_{DRS}(v^k)\|_2 + \|G_{DRS}(v^k)\|_2 \\ &\leq \|v^{k+1} - v^k\|_2 + \|G_{DRS}(v^k)\|_2 \\ &\leq \|H_k\|_2\|g^k\|_2 + \|g^k\|_2 \leq (2 + 2/\eta)\|g^k\|_2\end{aligned}$$

  since G_{DRS} is non-expansive and $\|H_k\|_2 \leq 1 + 2/\eta$.

Proof of Theorem 1 (Solvable Case cont.)

For AA updates,

- we have

$$\begin{aligned} \|g^{k+1}\|_2 &= \|G_{DRS}(v^{k+1})\|_2 \\ &\leq \|G_{DRS}(v^{k+1}) - G_{DRS}(v^k)\|_2 + \|G_{DRS}(v^k)\|_2 \\ &\leq \|v^{k+1} - v^k\|_2 + \|G_{DRS}(v^k)\|_2 \\ &\leq \|H_k\|_2\|g^k\|_2 + \|g^k\|_2 \leq (2 + 2/\eta)\|g^k\|_2 \end{aligned}$$

  since G_{DRS} is non-expansive and $\|H_k\|_2 \leq 1 + 2/\eta$.
- Hence for any $0 \leq K \leq R - 1$, we have

$$\|g^{k_i+K}\|_2 \leq (2 + 2/\eta)^K \|g^{k_i}\|_2 \leq D\|g^0\|_2(2 + 2/\eta)^K (i + 1)^{-(1+\epsilon)}$$

  and $\lim_{i\to\infty} \|g^{k_i+K}\|_2 = 0$.

Proof of Theorem 1 (Solvable Case cont.)

For AA updates,

- we have
$$\begin{aligned}\|g^{k+1}\|_2 &= \|G_{DRS}(v^{k+1})\|_2 \\ &\leq \|G_{DRS}(v^{k+1}) - G_{DRS}(v^k)\|_2 + \|G_{DRS}(v^k)\|_2 \\ &\leq \|v^{k+1} - v^k\|_2 + \|G_{DRS}(v^k)\|_2 \\ &\leq \|H_k\|_2\|g^k\|_2 + \|g^k\|_2 \leq (2+2/\eta)\|g^k\|_2\end{aligned}$$
since G_{DRS} is non-expansive and $\|H_k\|_2 \leq 1 + 2/\eta$.
- Hence for any $0 \leq K \leq R-1$, we have
$$\|g^{k_i+K}\|_2 \leq (2+2/\eta)^K\|g^{k_i}\|_2 \leq D\|g^0\|_2(2+2/\eta)^K(i+1)^{-(1+\epsilon)}$$
and $\lim_{i\to\infty}\|g^{k_i+K}\|_2 = 0$.
- Similarly, for any $w \in \mathbb{R}^n$, we have
$$\begin{aligned}\|v^{k_i+K+1} - w\|_2 &\leq \|v^{k_i+K} - w\|_2 + (1+2/\eta)\|g^k\|_2 \\ &\leq \cdots \leq \|v^{k_i} - w\|_2 + (1+2/\eta)\textstyle\sum_{j=1}^K \|g^{k_i+j}\|_2 \\ &\leq \|v^{k_i} - w\|_2 + (1+2/\eta)\textstyle\sum_{j=1}^K \|g^{k_i}\|_2(2+2/\eta)^j \\ &\leq \|v^{k_i} - w\|_2 + (1+2/\eta)D\|g^0\|_2(i+1)^{-(1+\epsilon)}C_R\end{aligned} \tag{10}$$
where $C_R = \sum_{j=1}^{R-1}\|g^{k_i}\|_2(2+2/\eta)^j$ is a constant.

Proof of Theorem 1 (Solvable Case cont.)

For DRS updates,

- Since F_{DRS} is firmly non-expansive, we have

$$\|v^{l_i+1} - v^*\|_2^2 \leq \|v^{l_i} - v^*\|_2^2 - \|g^{l_i}\|_2^2 \leq \|v^{l_i} - v^*\|_2^2 \tag{11}$$

  for any $i \geq 0$.

Hence for any $k \geq 0$,

$$\|v^k - v^*\|_2 \leq \|v^0 - v^*\|_2 + (1 + 2/\eta)D\|g^0\|_2 C_R \sum_{i=0}^{\infty}(i+1)^{-(1+\epsilon)} \triangleq E$$

implies that $\|v^k - v^*\|_2$ is bounded.
Thus Step 1 is completed.

Proof of Theorem 1 (Solvable Case cont.)

By squaring equation (10) and combined with equation (11), we have

$$\sum_{i=0}^{\infty} \|g^{l_i}\|_2^2 \leq \|v^0 - v^*\|_2^2 + \text{const}$$

where

$$\begin{aligned} \text{const} \quad &= ((1+2/\eta)C_R D\|g^0\|_2)^2 \textstyle\sum_{i=0}^{\infty}(i+1)^{-(2+2\epsilon)} \\ &+ (2+4/\eta)C_R DE\|g^0\|_2 \textstyle\sum_{i=0}^{\infty}(i+1)^{-(1+\epsilon)} \end{aligned}$$

Hence $\lim_{i\to\infty}\|g^{l_i}\|_2 = 0$. Together with $\lim_{i\to\infty}\|g^{k_i+K}\|_2 = 0$ for any

$0 \leq K \leq R-1$, we obtain $\lim_{k\to\infty}\|g^k\|_2 = 0$.
Hence Step 2 is completed.

Proof of Theorem 1 (Solvable Case cont.)

By squaring equation (10) and combined with equation (11), we have

$$\|v^{k+1} - v^*\|_2 \leq \|v^k - v^*\|_2 + \epsilon_k$$

where $\epsilon_{l_i} = 0$ and

$$\begin{aligned} \epsilon_{k_i+K} &= ((1+2/\eta)D\|g^0\|_2)^2(2+2/\eta)^{2K}(i+1)^{-(2+2\epsilon)} \\ &\quad +(2+4/\eta)DE\|g^0\|_2(2+2/\eta)^K(i+1)^{-(1+\epsilon)} \end{aligned}$$

for $0 \leq K \leq R-1$.

Hence

$$\epsilon_k \geq 0 \text{ and } \sum_{k=0}^{\infty} \epsilon_k < \infty,$$

i.e., $\|v^k - v^*\|_2$ is quasi-Fejérian.
Step 3 is completed.

Proof of Theorem 2 (Pathological Case)

- Suppose problem (1) is pathological, then $\delta v^* \neq 0$, hence $\|\delta v^*\|_2 > 0$.
- The safeguard will always be invoked for sufficiently large k due to large residuals. Hence the algorithm reduces to vanilla DRS in the end.
- The remaining results follow from previous studies in[8],[9].

[8] Ernest K Ryu, Yanli Liu, and Wotao Yin. "Douglas–Rachford splitting and ADMM for pathological convex optimization". In: *Computational Optimization and Applications* 74.3 (2019), pp. 747–778.

[9] Yanli Liu, Ernest K Ryu, and Wotao Yin. "A new use of Douglas–Rachford splitting for identifying infeasible, unbounded, and pathological conic programs". In: *Mathematical Programming* 177.1-2 (2019), pp. 225–253.

Outline

Presolve

- Solve $Ax = b$ as a least squares problem first to detect infeasibility
- Preconditioning: scale x and linear constraints
  - Goal: Reduce condition number of coefficient matrix
  - Method: regularized Sinkhorn-Knopp method
  - Choose $D = \mathbf{diag}\,(d_1, ..., d_m)$ and $E = \mathbf{diag}\,(e_1 I_{n_1}, ..., e_N I_{n_N})$ with $d_i > 0, e_j > 0$ for all $i = 1, .., m, j = 1, ..., N$ so the scaled problem is

$$\begin{array}{ll} \min\limits_{\hat{x}} & \sum_{i=1}^{N} \hat{f}_i(\hat{x}_i) \\ \text{s.t.} & \sum_{i=1}^{N} \hat{A}_i \hat{x}_i = \hat{b} \end{array} \tag{12}$$

where $\hat{f}_i(\hat{x}_i) = f_i(e_i \hat{x}_i)$, $\hat{A} = D[A_1, A_2, ..., A_N]E$, $\hat{b} = Db$. One can recover $x^* = E\hat{x}^*$.

Presolve

- Solve $Ax = b$ as a least squares problem first to detect infeasibility
- Preconditioning: scale x and linear constraints
  - Goal: Reduce condition number of coefficient matrix
  - Method: regularized Sinkhorn-Knopp method
  - Choose $D = \mathbf{diag}\,(d_1, ..., d_m)$ and $E = \mathbf{diag}\,(e_1 I_{n_1}, ..., e_N I_{n_N})$ with $d_i > 0, e_j > 0$ for all $i = 1, .., m, j = 1, ..., N$ so the scaled problem is

$$\begin{array}{ll} \min\limits_{\hat{x}} & \sum_{i=1}^N \hat{f}_i(\hat{x}_i) \\ \text{s.t.} & \sum_{i=1}^N \hat{A}_i \hat{x}_i = \hat{b} \end{array} \tag{12}$$

    where $\hat{f}_i(\hat{x}_i) = f_i(e_i \hat{x}_i)$, $\hat{A} = D[A_1, A_2, ..., A_N]E$, $\hat{b} = Db$. One can recover $x^* = E\hat{x}^*$.
  - The scaling parameters are determined by solving

$$\min_{u,v} \sum_{i=1}^m \sum_{j=1}^N B_{ij} e^{u_i + u_j} - N\mathbf{1}^T u - m\mathbf{1}^T v + \gamma \left(N \sum_{i=1}^m e^{u_i} + m \sum_{j=1}^N e^{v_j} \right) \tag{13}$$

    where $\gamma > 0$ and $B_{ij} = \sum_{l=n_1+\cdots+n_{j-1}+1}^{n_1+\cdots+n_j} A_{il}^2$ and setting $d^i = e^{u_i/2}$ and $e_j = e^{v_j/2}$

Presovle cont.

- Preconditioning: regularized Sinkhorn-Knopp method
  - The problem

    $$\min_{u,v} \sum_{i=1}^{m} \sum_{j=1}^{N} B_{ij} e^{u_i + u_j} - N\mathbf{1}^T u - m\mathbf{1}^T v + \gamma \left(N \sum_{i=1}^{m} e^{u_i} + m \sum_{j=1}^{N} e^{v_j} \right)$$

    is strictly convex.
  - The parameter $\gamma = \frac{m+N}{mN}\sqrt{\epsilon^{mp}}$ where ϵ^{mp} is the machine precision.
  - Note when $\gamma = 0$ and the problem has a solution, the resulting $\hat{A}$ is equilibrated exactly, i.e., all rows have the same ℓ_2-norm and the columns have the same ℓ_2-norm in the blockwise sense.
  - Coordinate descent is used to solve the above problem
  - They further scale u and v to make the arithmetic mean of u and v equal and $\|DAE\|_F = \sqrt{\min(m, N)}$.

Parameter setting

- Proximal step parameter: $t = \frac{1}{10}\left(\prod_{j=1}^{N} e_j\right)^{-2/N}$
  - It is chosen to minimize $\sum_{i=1}^{N}(\log t - \log(ce_i^{-2}))^2$ where $c = 1/10$
    Since $\hat{x}_i = \mathbf{prox}_{t\hat{f}_i}(\hat{v}_i) = \frac{1}{e_i}\mathbf{prox}_{e_i^2 t f_i}(e\hat{v}_i)$
- Memory parameter $M_{max} = 10$
- Regularization coefficient $\eta = 10^{-8}$
- Safeguarding parameters
  - $D = 10^6$
  - $\epsilon = 10^{-6}$
  - $R = 10$
- Stopping criteria parameters: $\epsilon_{abs} = 10^{-6}$ and $\epsilon_{rel} = 10^{-8}$.
- Initialization: $v^0 = 0$

Implementation

- Least squares : `scipy.sparse.linalg.lsqr`
  - Evaluate projection $\Pi(\cdot)$
  - Compute dual variable by minimizing $\|r_{\mathrm{dual}}^k\|_2$
  - Solve adaptive regularization (8): use SVD-based solver `numpy.linalg.lstsq` as default since Y_k is tall-and-thin
- Solver interface: `result = a2dr(p_list, A_list, b)`
- Multiprocessing: `multiprocessing`
- Pathological case detection: Theoretically one can use successive difference $v^k - v^{k+1}$ as certificates of infeasibility or unboundedness. But the current software does not implement such certificates.

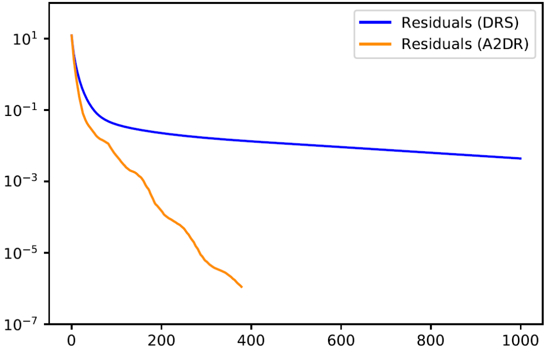
Outline

Example 1: Nonnegative Least Squares

Let $F \in \mathbb{R}^{p \times q}$ and $g \in \mathbb{R}^p$,

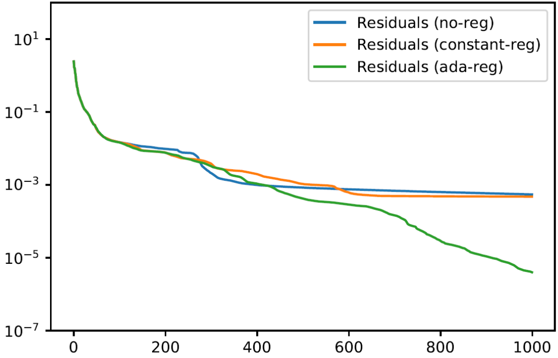
$$\begin{array}{ll} \min & \|Fz - g\|_2^2 \\ \text{s.t.} & z \geq 0 \end{array} \iff \begin{array}{ll} \min & \|Fx_1 - g\|_2^2 + \mathcal{I}_{\mathbb{R}_+^q}(x_2) \\ \text{s.t.} & x_1 - x_2 = 0 \end{array}$$

- Set $p = 10000$, $q = 8000$. F is sparse with 0.1% nonzero items drawn from i.i.d. $\mathcal{N}(0,1)$ and $g \in \mathcal{N}(0, I)$.
- The proximal operator is evaluated using LSQR.
- A2DR(55s, 10^{-10}) beats OSQP(349s, 10^{-6}) and SCS(327s, 10^{-6})



Example 1: Nonnegative Least Squares (cont.)

- Examine the effect of adaptive regularization.
- Set $p = 300$, $q = 500$.

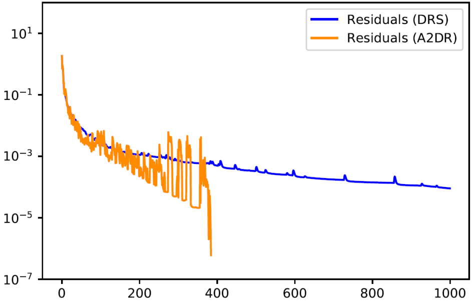


Example 2: Sparse Inverse Covariance Estimation

Let $Q = \frac{1}{p}\sum_{l=1}^{p} z_l z_l^T$ where $z_l \in \mathbb{R}^q$ and $\alpha > 0$. The problem is

$$\begin{array}{ll} \min & -\log\det(S_1) + \mathrm{tr}(S_1 Q) + \alpha\|S_2\|_1 \\ \text{s.t.} & S_1 - S_2 = 0 \end{array}$$

- Set $p = 1000$, $q = 100$. Generate sparse $S \in \mathcal{S}_{++}^q$ with around 10% nonzero entries. Calculate Q using p i.i.d. samples from $\mathcal{N}(0, S^{-1})$. $\alpha = 0.001 \sup_{i \neq j} |Q_{ij}|$
- The proximal operator is evaluated with complexity $O(q^3)$.
- A2DR(1h, 10^{-3}; 2.6h, 10^{-3}) beats SCS(11h, 10^{-1}; ?) on $q = 1200$ and $q = 2000$.



Example 3: Multitask Regularized Logistic Regression

Let $Z \in \mathbb{R}^{p\times L}$, $Y \in \{-1,1\}^{p\times L}$, the problem is

$$\begin{array}{ll} \min & f_1(Z) + f_2(\theta) + f_2(\tilde{\theta}) \\ \text{s.t.} & Ax = 0 \end{array}$$

where

$$A = \left[\begin{array}{ccc} I & -W & 0 \\ 0 & I & -I \end{array}\right], \quad x = \left[\begin{array}{c} Z \\ \theta \\ \tilde{\theta} \end{array}\right], \quad b = 0, \quad \theta = [\theta_1 \cdots \theta_L] \in \mathbb{R}^{s\times L}$$

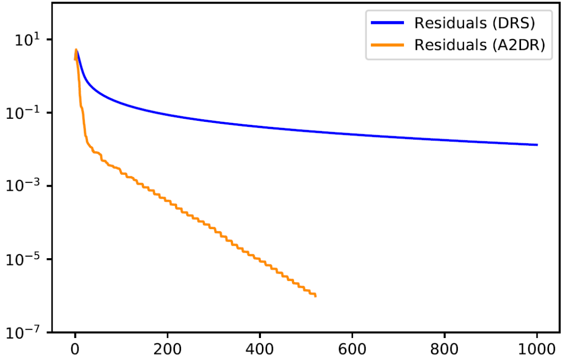
and

$$\begin{array}{l} f_1(Z) = \phi(Z,Y) = \sum_{l=1}^{L}\sum_{i=1}^{p} \log\left(1 + \exp(-Y_{il}Z_{il})\right) \\ f_2(\theta) = \alpha\|\theta\|_{2,1} = \alpha\sum_{l=1}^{L}\|\theta_l\|_2 \\ f_3(\tilde{\theta}) = \beta\|\tilde{\theta}\|_* \text{ is the nuclear norm} \end{array}$$

- The proximal operator of f_1 is evaluated via Newton type methods (`scipy.optimize.minimize`) in parallel.

Example 3: Multitask Regularized Logistic Regression

Let $p = 300$, $s = 500$, $L = 10$, $\alpha = \beta = 0.1$. The entries of $W \in \mathbb{R}^{p \times s}$ and $\theta^* \in \mathbb{R}^{s \times L}$ are drawn i.i.d. from $\mathcal{N}(0, 1)$. Calculate $Y = \mathbf{sign}\,(W\theta^*)$.



Concluding Remarks

- Combine DRS and type-II AA on convex problems of prox-affine form
- Adaptive regularization for stability
- Global convergence
- A Python software `https:github.com/cvxgrp/a2dr`
    - Fast, parallelized, scalable and memory-efficient.

Concluding Remarks

- Combine DRS and type-II AA on convex problems of prox-affine form
- Adaptive regularization for stability
- Global convergence
- A Python software https:github.com/cvxgrp/a2dr
    - Fast, parallelized, scalable and memory-efficient.

Thank You! Questions?