Convergence rate analysis of several splitting schemes[1]

Damek Davis[2]

Department of Mathematics
University of California, Los Angeles

INFORMS annual meeting 2014

[1] Joint work with Prof. Wotao Yin (UCLA) (http://arxiv.org/abs/1406.4834)
[2] http://www.math.ucla.edu/~damek

Background/outline

- **Topic:** Convergence rates of splitting algorithms
- Convergence rates of these algorithms were unknown for many years.
- **Today:** I'll present a simple procedure for convergence rate analysis that generalizes to a wide class of algorithms.
- **Outline**:
  - Algorithms
  - Our Question
  - Challenges/Techniques

What is a splitting?

- **We want to:**

$$\underset{x \in \mathcal{H}}{\text{minimize}} \; f(x) + g(x).$$

  - $\mathcal{H}$ is a Hilbert space, may be infinite dimensional.
  - f **and** g **are closed, proper, and convex** (not necessarily differentiable).
  - **Focus of all algorithms today**

Basic operations in splitting algorithms

- **The proximal operator:** For all $x \in \mathcal{H}$ and $\gamma > 0$

$$\mathbf{prox}_{\gamma h}(\mathbf{x}) := \operatorname*{arg\,min}_{y \in \mathcal{H}} \; h(y) + \frac{1}{2\gamma}\|y - \mathbf{x}\|^2$$
$$= \mathbf{x} - \gamma \widetilde{\nabla} h(\mathbf{prox}_{\gamma h}(\mathbf{x})) \longleftarrow \textbf{implicit subgradient}.$$

  - For all $z \in \mathcal{H}$, the vector $\widetilde{\nabla} h(z) \in \partial h(z)$ is a subgradient.
  - $\mathbf{prox} =$ **Main subproblem in splitting algorithms**.
  - Many functions in machine learning and signal processing have **simple or closed form** proximal operators (e.g., ℓ_1 and matrix norms, indicator functions, quadratic functions,...).

Major first order algorithms: subgradient form

implicit semi-implicit explicit

- **(Sub)gradient method:**

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

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

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

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

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

- **Douglas Rachford splitting (DRS):**

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

- $\implies \|z^{k+1} - z^k\|$ controls size of subgradients!

Major first order algorithms: diagram form

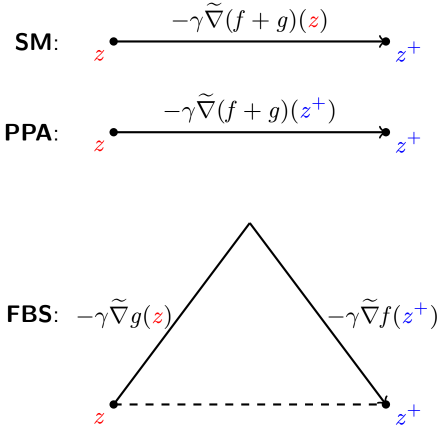
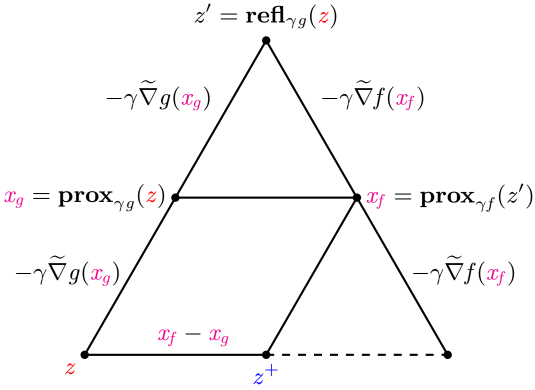


Diagram of DRS



$$z^+ = \frac{1}{2}z + \frac{1}{2}\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\gamma g}(z), \quad \text{where } \mathbf{refl} := 2\mathbf{prox} - I$$

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

Our main question

How fast and how slow are splitting algorithms?

- For simplicity, let's consider objective error and the unconstrained problem

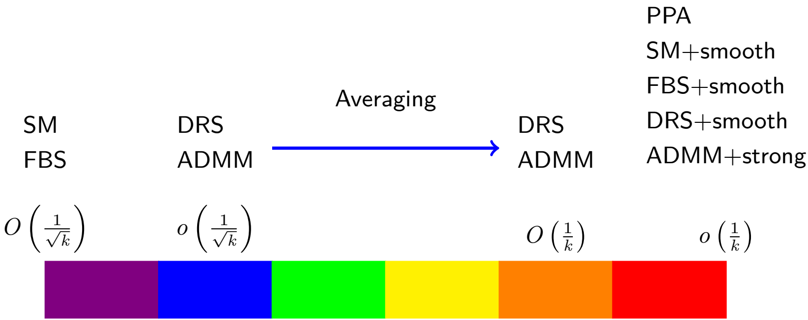
$$\underset{x \in \mathcal{H}}{\text{minimize}}\ f(x) + g(x).$$

  - Let $x^* \in \mathcal{H}$ be a minimizer of $f + g$.
  - Our goal is to measure

$$f(x^k) + g(x^k) - f(x^*) - g(x^*)$$

  for certain natural sequences $(x^j)_{j \geq 0}$.
- **Note:** This talk is not comprehensive.
  - The paper analyzes other algorithms and convergence measures.

Results: spectrum of objective error convergence rates



- The rates are sharp. (new result)
- **Counterintuitive result:** DRS is nearly as slow as subgradient method...
- ...but averaging: $(x^j)_{j\geq 0} \mapsto (\frac{1}{j+1}\sum_{i=0}^{j} x^i)_{j\geq 0}$
  - Smooths objective value sequence.
  - Nearly as fast as PPA.
- For DRS, the smooth results only require f OR g to be smooth, not both. (FBS needs g smooth and SM needs $f+g$ smooth.)

Should we always average?

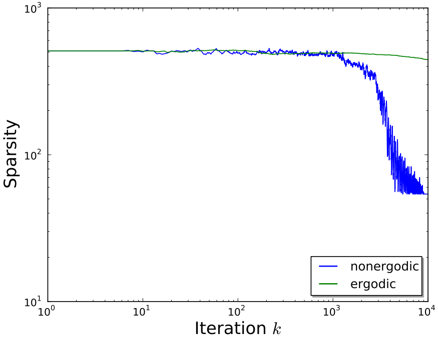
- Convergence rates improve when we average.

$$o(1/\sqrt{k+1}) \to O(1/(k+1)).$$

- Should we always average?
  - No. Can ruin sparsity patterns in the solution/prolong convergence
  - Consider DRS applied to basis pursuit problem

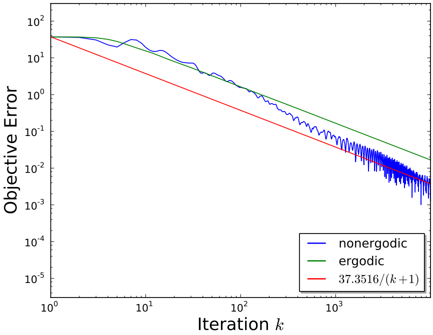
$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \ \|x\|_1$$

$$\text{subject to: } Ax = b$$



Challenges of convergence analysis

- In splitting algorithms, implicit/explicit subgradients are generated at two different points
  - Should make Lipschitz continuity assumption.
    - **Example:** $C \subseteq \mathcal{H}$, $f = \chi_C$ (0 in C, ∞ outside), $g = \|\cdot\|_2$, only natural point to evaluate is in C.
- **The objective does not decrease monotonically**
  - $\Longrightarrow$ The classical approaches to obtain convergence rates fail!



Other forms of monotonicity

- **Other quantities decrease monotonically**:
  - z^* be a fixed point of one of the above algorithms.

$$\|z^{k+1} - z^*\|^2 \leq \|z^k - z^*\|^2 - \|z^{k+1} - z^k\|^2$$

$$\|z^{k+1} - z^k\|^2 \leq \|z^k - z^{k-1}\|^2$$

  - The above inequalities are key to the convergence analysis.
  - Implies that $\|z^{k+1} - z^k\|^2$ is monotonic and summable! (Important)
  - true in PPA/FBS/DRS/ADMM/forward-Douglas-Rachford splitting/Chambolle and Pock's primal-dual algorithm....
- Recall: $\|z^{k+1} - z^k\|$ controls subgradient size

Our techniques: nonsmooth case

Our results follow from three tools.

- A lemma that estimates convergence rates of sequences
  - **Roughly:** $(a_j)_{j\geq 0} \subseteq \mathbf{R}$ summable and monotonic $\implies a_k = o(1/(k+1))$.
- A theorem that estimates convergence rates of subgradients in splitting algorithms
  - **Recall:** $\|z^{k+1} - z^k\|^2$ is monotonic and summable, and so

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

  - $\implies$ In DRS:

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

- An inequality that bounds objective values by subgradient norms.
  - $\implies$ nonergodic rate $o(1/\sqrt{k+1})$.
  - Ergodic $O(1/(k+1))$ rates follows from this inequality + Jensen's inequality.

Conclusions

- We also analyze ADMM and other splitting algorithms.
- All of the obtained rates are sharp! (new result)
- **Applications in the paper**: New convergence rates for feasibility, distributed model fitting, linear programming, semidefinite programming, and decentralized ADMM problems.
- In a followup paper, we study these algorithms when f and g are *regular* (e.g., strongly convex or differentiable).[3]
  - The rates automatically improve without knowledge of Lipschitz constants or strong convexity modulus.
  - e.g., for differentiable f or g $o(1/\sqrt{k+1}) \to o(1/(k+1))$.
- We also generalized these techniques to prove convergence rates of wide class of primal-dual algorithms[4].

[3] http://arxiv.org/abs/1407.5210

[4] http://arxiv.org/abs/1408.4419

References

- Damek Davis, and Wotao Yin.
  Convergence rate analysis of several splitting schemes.
  arXiv:1406.4834 (2014).
- Damek Davis, and Wotao Yin.
  Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions.
  arXiv:1407.5210 (2014).
- Damek Davis
  Convergence rate analysis of primal-dual splitting schemes.
  arXiv:1408.4419 (2014).
- More: http://www.math.ucla.edu/∼damek