Accelerating First-Order Methods via Linear Prediction

Clarice Poon
Department of Mathematical Sciences, University of Bath
Claverton Down, Bath, BA2 7AY, United Kingdom
Email: C.M.H.S.Poon@maths.cam.ac.uk

Jingwei Liang
DAMTP, University of Cambridge
Wilberforce Road, Cambridge, CB3 0WA, United Kingdom
Email: jl993@cam.ac.uk

***Abstract*—In this paper, we propose a novel linear prediction framework for accelerating first-order methods. We first discuss the trajectory of the sequence generated by first-order methods, showing that eventually the trajectory obeys certain low-dimensional regularity. Then building on top of such a property, we design a trajectory following linear prediction scheme for accelerating first-order methods. Numerical experiments on problems arising from inverse problems and image processing demonstrate the state-of-the-art performance of the proposed algorithm.**

I. Introduction

First-order methods are widely used in many fields through science and engineering, such as inverse problems, image processing, compressed sensing and machine learning, etc. Let **FoM** be the class of first-order methods, and $\mathcal{F} \in \mathbf{FoM}$ be a first-order method, the iteration of $\mathcal{F}$ then can be written as

$$z_{k+1} = \mathcal{F}(z_k), \tag{I.1}$$

where $\{z_k\}_{k\in\mathbb{N}}$ is the fixed-point sequence. Examples of (I.1) include, for instance, Forward–Backward splitting (FB) [1], Douglas–Rachford splitting (DR) [2] and Primal–Dual splitting (PD) [3].

II. Trajectory of first-order methods

In this part, we use Forward–Backward and Douglas–Rachford splitting to demonstrate two distinct types of trajectories.

Forward–Backward splitting is designed to handle the problem

$$\min_{x\in\mathbb{R}^n} \{\Phi(x) \stackrel{\text{def}}{=} R(x) + F(x)\}, \tag{$\mathcal{P}$}$$

where $R : \mathbb{R}^n \to]-\infty, +\infty]$ is proper convex and lower semi-continuous, and $F \in C^{1,1}(\mathbb{R}^n)$ is convex and has L-Lipschitz gradient ∇F. The standard iteration of Forward–Backward reads

$$x_{k+1} = \mathrm{prox}_{\gamma R}(x_k - \gamma\nabla F(x_k)),\ \gamma \in]0, 2/L[,$$

where $\mathrm{prox}_{\gamma R}(\cdot) \stackrel{\text{def}}{=} \mathrm{argmin}_x \gamma R(x) + \frac{1}{2}\|x - \cdot\|^2$ is the proximal mapping of γR. If we assume that F is non-smooth, then ($\mathcal{P}$) can be solved by Douglas–Rachford spitting: let $\gamma > 0$

$$z_{k+1} = z_k + \mathrm{prox}_{\gamma R}(2x_k - z_k) - x_k,$$
$$x_{k+1} = \mathrm{prox}_{\gamma F}(z_{k+1}).$$

Assume $\mathrm{Argmin}(\Phi) \neq \emptyset$, let $z^\star$ be the point z_k converges to for DR and $x^\star \in \mathrm{Argmin}(\Phi)$ be a global minimiser. Let $v_k \stackrel{\text{def}}{=} x_k - x_{k-1}$ for FB and $v_k \stackrel{\text{def}}{=} z_k - z_{k-1}$ for DR, define the angle $\theta_k \stackrel{\text{def}}{=} \angle(v_k, v_{k-1})$.

Theorem II.1. *For problem ($\mathcal{P}$) and Forward–Backward splitting, suppose R is locally polyhedral around $x^\star$ and F is C^2 around $x^\star$, then there exists an $\eta \in]0,1[$ such that $1 - \cos(\theta_k) = O(\eta^k)$.*

Suppose that both R and F are locally polyhedral around $x^\star$, then for Douglas–Rachford splitting, there exist $\omega \in]0, \pi/2]$ and $\eta \in]0,1[$ such that $\cos(\omega) - \cos(\theta_k) = O(\eta^k)$.

For Douglas–Rachford, let T_R, T_J be the minimal subspaces that are parallel to $\partial R(x^\star), \partial F(x^\star)$, respectively, then ω is the so-called Friedrichs angle between T_R and T_J [7].

Theorem II.1 implies that the eventual trajectory is a straight line for Forward–Backward, and a spiral for Douglas–Rachford; See the very first figure of Figure 1-3 for examples in $\mathbb{R}^3$.

III. Linear prediction for first-order methods

The convergence of $\cos(\theta_k)$ in Theorem II.1 implies that the eventual trajectories of Forward–Backward/Douglas–Rachford enjoy a low-dimensional regularity. Such a property is not only limited to these two methods and locally polyhedral functions, but rather many popular first-order methods and non-smooth optimisation problems. As a consequence, we can use such a regularity to design trajectory following acceleration scheme for first-order methods.

Given $c \in \mathbb{R}^q$, define the mapping H by,

$$H(c) = \left[\begin{array}{c|c} c_{1:q-1} & \mathrm{Id}_{q-1} \\ c_q & 0_{1,q-1} \end{array}\right] \in \mathbb{R}^{q\times q}.$$

Given $C \in \mathbb{R}^{q\times q}$, denote C^k its power and $C_{:,1}$ its first column.

Let $\mathcal{F}$ be a first-order method as in (I.1), we propose the following scheme for accelerating (I.1).

Algorithm 1: A linear prediction scheme for **FoM**

Initial: Let $s \geq 1, q \geq 2$ be integers, $D = 0 \in \mathbb{R}^{n\times(q+1)}$;
Repeat:

- For $k \geq 1$:
$$z_k = \mathcal{F}(z_{k-1}),\ \ D = [z_k - z_{k-1}, D(:, 1:q)].$$
- If $\mathrm{mod}(k, q+1) = 0$: let $V_{k-1} = D(:, 2:q+1)$:
$$c_k = (V_{k-1}^T V_{k-1})^{-1} V_{k-1}^T v_k \ \text{ and } \ C_k = H(c_k),$$
$$y_k = z_k + V_k(\textstyle\sum_{i=1}^s C_{:,1}^i),\ \ z_{k+1} = \mathcal{F}(y_k).$$

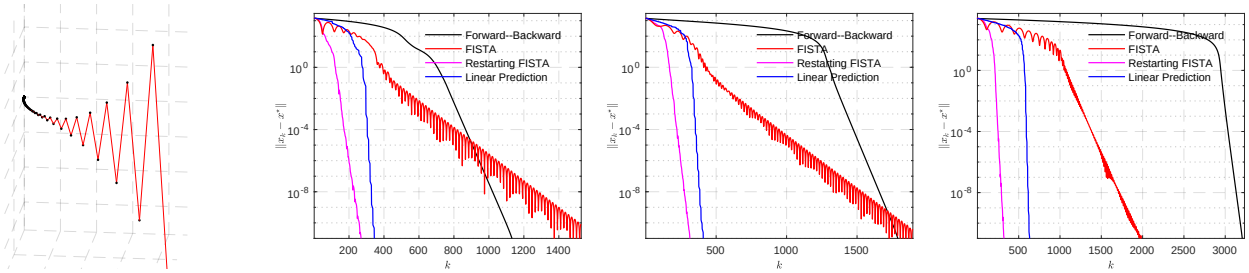
The above scheme is equivalent to *minimal polynomial extrapolation* if we set $s = +\infty$. By treating Algorithm 1 as a perturbation of (I.1), a safeguard rule can be added to guarantee its convergence [4, Eq. 4.2.6]. For specific types of problems, thanks to the sequence trajectories, explicit acceleration rates can be obtained. For example, in the case of polyhedral R, F for DR, $\|y_k - z^\star\| = \mathcal{O}(\sigma^k\|z_k - z^\star\|)$ for some $\sigma < 1$ under the choice of $q = 3$.

IV. Numerical experiments

To demonstrate the performance of Algorithm 1, linear inverse problems are considered for numerical experiments. For Forward–Backward splitting, we consider solving $\min_{x\in\mathbb{R}^n} \mu R(x) + \frac{1}{2}\|y - Ax\|^2$, where $y \in \mathbb{R}^m$ is the noise contaminated observation, $A \in \mathbb{R}^{m\times n}$ is drawn from the standard Gaussian ensemble, and $\mu > 0$ is the regularisation parameter. The results are depicted in Figure 1.

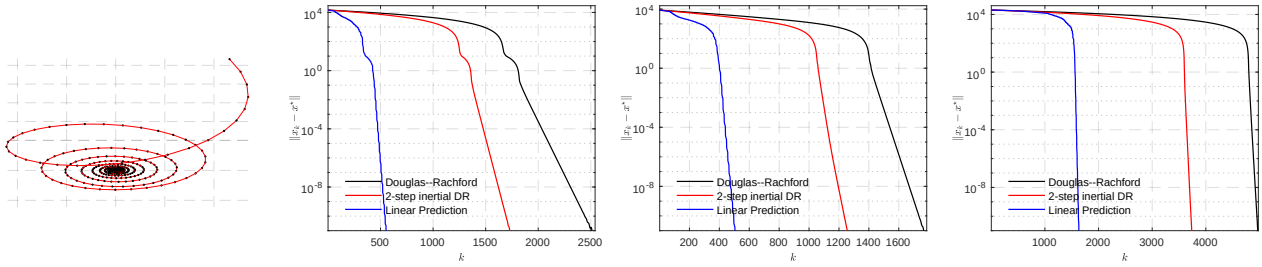
For Douglas–Rachford and Primal–Dual splitting methods, we consider solving $\min_{x\in\mathbb{R}^n} R(x)$ s.t. $Ax = y$, where $y \in \mathbb{R}^m$ is the observation, $A \in \mathbb{R}^{m\times n}$ is drawn from the standard Gaussian ensemble. The results are provided in Figure 2 and 3, respectively.

For all numerical comparison, three choices of R are considered: ℓ_1-norm, total variation and nuclear norm. For Algorithm 1, we fixed $q = 4$ and $s = 80$ for all tests. Except the trajectory plots which are in $\mathbb{R}^3$, all the other numerical experiments are conducted in high dimension.



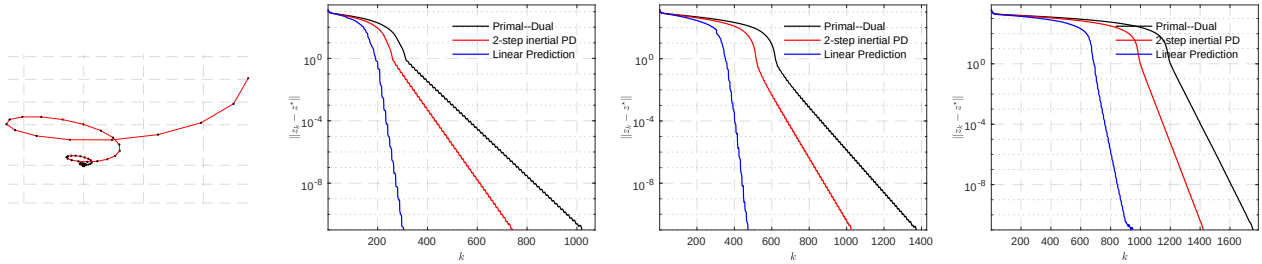
(a) Trajectory of FB (b) R is ℓ_1-norm (c) R is total variation (d) R is Nuclear norm

Fig. 1: Trajectory of Forward–Backward splitting in $\mathbb{R}^3$ and performance comparison between Forward–Backward, FISTA [5], restarting FISTA [6] and Algorithm 1 in terms of $\|x_k - x^\star\|$. In terms of step size, we use $\gamma = 1.95/L$ for Forward–Backward, $\gamma = 1/L$ for (restarting) FISTA and $\gamma = 1.95/L$ for Algorithm 1. It can be observed that, restarting FISTA is the fastest, and Algorithm 1 is the second. Subfigure (a) is the trajectory of Forward–Backward in $\mathbb{R}^3$. For the other three subfigures, we solve the problem $\min_{x\in\mathbb{R}^n} \mu R(x) + \frac{1}{2}\|y - Ax\|^2$.



(a) Trajectory of DR (b) R is ℓ_1-norm (c) R is total variation (d) R is Nuclear norm

Fig. 2: Trajectory of Douglas–Rachford splitting in $\mathbb{R}^3$ and performance comparison between Douglas–Rachford, 2-step inertial Douglas–Rachford [4] and Algorithm 1 in terms of $\|z_k - z^\star\|$. It can be observed that, Algorithm 1 is significantly faster than the other two. We fix $\gamma = 1$ for all methods. Subfigure (a) is the trajectory of Douglas–Rachford in $\mathbb{R}^3$. For the other three subfigures, we solve a problem $\min_{x\in\mathbb{R}^n} R(x) + \iota_{\{x\in\mathbb{R}^n : y = Ax\}}(x)$ where $\iota_{\{x\in\mathbb{R}^n : y=Ax\}}(x)$ is the indicator function fo the affine constraint set $\{x \in \mathbb{R}^n : y = Ax\}$.



(a) Trajectory of PD (b) R is ℓ_1-norm (c) R is total variation (d) R is Nuclear norm

Fig. 3: Trajectory of Primal–Dual splitting in $\mathbb{R}^3$ and performance comparison between Primal–Dual, 2-step inertial Primal–Dual [4] and Algorithm 1 in terms of $\|z_k - z^\star\|$. All schemes use the same step-size setting. Similar to the above two comparisons, Algorithm 1 is significantly faster than the other two. Subfigure (a) is the trajectory of Primal–Dual in $\mathbb{R}^3$. For the other three subfigures, we solve a saddle problem $\min_{x\in\mathbb{R}^n} \max_{w\in\mathbb{R}^m} R(x) + \langle Ax - b, w\rangle - \iota^*_{\{0\}}(w)$ where $\iota^*_{\{0\}}(w)$ is the convex conjugate of $\iota_{\{0\}}(w)$. For this case, we have $z = (x; w)$ is the fixed-point sequence of Primal–Dual splitting.

References

[1] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*. SIAM Journal on Numerical Analysis, 16(6):964–979, 1979.

[2] J. Douglas and Jr. H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*. Transactions of the American mathematical Society, 82(2):421–439, 1956.

[3] A. Chambolle and T. Pock, *A first-order Primal–Dual algorithm for convex problems with applications to imaging*. Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011.

[4] J. Liang, *Convergence rates of first-order operator splitting methods*. PhD thesis, Normandie Université; GREYC CNRS UMR 6072, 2016.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.

[6] B. O'Donoghue and E. Candès, *Adaptive restart for accelerated gradient schemes*. Foundations of computational mathematics, 15(3):715–732, 2015.

[7] H. H. Bauschke, et al, *Optimal rates of linear convergence of relaxed alternating projections and generalized Douglas–Rachford methods for two subspaces*. Numerical Algorithms, 73(1):33–76, 2016.