Local Linear Convergence of Primal–Dual Splitting Methods for Low Complexity Regularization

Jingwei Liang*, Jalal M. Fadili*, Gabriel Peyré†
*Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Email: {Jingwei.Liang, Jalal.Fadili}@ensicaen.fr
†CNRS, DMA, École Normale Supérieure, Paris, Email: Gabriel.Peyre@ens.fr

***Abstract*—Primal–Dual (PD) splitting method has become very popular for solving sparse recovery problems and beyond (see for instance the review [9]). The goal of this work is to understand the local convergence behaviour of PD which has been observed in practice to exhibit local linear rate of convergence. In this paper, we show that when the involved non-smooth functions are partly smooth, the PD algorithm identifies the associated active manifolds in finite time, and then locally converges linearly with a rate determined by the properties of the primal and dual active manifolds. The result is illustrated by several concrete examples and supported by numerical experiments.**

I. Introduction

Consider the following composite optimization problem,

$$\min_{x\in\mathbb{R}^n} R(x) + F(x) + (J \mathbin{\dot{\vee}} G)(Lx), \tag{1}$$

where R, J are proper, lower semi-continuous (lsc) and convex on $\mathbb{R}^n$ and $\mathbb{R}^m$, F is convex differentiable with ∇F being $1/\beta_{_F}$-Lipschitz continuous and G is $\beta_{_G}$-strongly convex, $L : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator and $(J \mathbin{\dot{\vee}} G)(\cdot) \stackrel{\text{def}}{=} \inf_{v\in\mathbb{R}^m} J(\cdot) + G(\cdot - v)$ denotes the infimal convolution of J and G. In this work, we focus on the saddle point formulation of problem (1) which reads

$$\min_{x\in\mathbb{R}^n}\max_{v\in\mathbb{R}^m} R(x) + F(x) + \langle Lx, v\rangle - \big(J^*(v) + G^*(v)\big), \tag{2}$$

and assume that strong duality holds (*i.e.* the duality gap is 0 [11]). Denote $\mathcal{X}$ and $\mathcal{V}$ the sets of primal and dual solutions, and assume that both $\mathcal{X}$ and $\mathcal{V}$ are non-empty. We also assume that both R, J^* are simple, *i.e.* $\mathrm{prox}_{\gamma R}, \mathrm{prox}_{\gamma J^*}, \gamma > 0$ are easy to compute, where $\mathrm{prox}_{\gamma R}$ denotes the proximity operator which is defined by $\mathrm{prox}_{\gamma R}(\cdot) = \mathrm{argmin}_{x\in\mathbb{R}^n} \frac{1}{2}\|x - \cdot\|^2 + \gamma R(x), \gamma > 0$.

In the literature, an efficient and provably convergent algorithm for solving (2) is the following Primal–Dual splitting method proposed in [8], [4] which covers [1], [5] as special cases,

$$\left\lfloor\begin{aligned} x_{k+1} &= \mathrm{prox}_{\tau R}(x_k - \tau L^* v_k - \tau\nabla F(x_k)),\\ \bar{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k),\\ v_{k+1} &= \mathrm{prox}_{\sigma J^*}(v_k + \sigma L\bar{x}_{k+1} - \sigma\nabla G^*(v_k)),\end{aligned}\right. \tag{3}$$

where $\tau, \sigma > 0$ are the step-size, and $\theta \in [0, 1]$. When $\theta = 1$, the convergence of the sequences generated by (3) is guaranteed if τ, σ are chosen such that $2\min\{\beta_{_F}, \beta_{_G}\}\min\{\frac{1}{\sigma}, \frac{1}{\tau}\}(1 - \sqrt{\sigma\tau\|L\|^2}) > 1$. Moreover, the convergence rate of the sequences is $o(1/\sqrt{k})$ which is sub-linear, see [10] and reference therein for the global sub-linear rate of convergence.

II. Partly Smooth Functions and Finite Identification

Beside being proper convex lsc, our central assumption for R, J^* is that they are partly smooth, a concept originally defined in [2]. Here we specialize it to the case of proper convex lsc functions. Denote $\mathrm{par}(C)$ the subspace parallel to the non-empty convex set $C \subset \mathbb{R}^n$.

Definition II.1. Let $R : \mathbb{R}^n \to]-\infty, +\infty]$ be proper convex and lsc, and $x \in \mathbb{R}^n$ such that $\partial R(x) \neq \emptyset$. R is *partly smooth at x relative to a set $\mathcal{M}$* containing x if

(Smoothness) $\mathcal{M}$ is a C^2-manifold, $R|_{\mathcal{M}}$ is C^2 around x.
(Sharpness) The tangent space $\mathcal{T}_{\mathcal{M}}(x) = T_x \stackrel{\text{def}}{=} \mathrm{par}(\partial R(x))^{\perp}$.
(Continuity) The ∂R is continuous at x relative to $\mathcal{M}$.

We denote $\mathrm{PSF}_x(\mathcal{M})$ the set of partly smooth functions at x relative to $\mathcal{M}$. Partly smooth functions are ubiquitous in imaging and machine learning, and popular examples include $\ell_1, \ell_{1,2}, \ell_\infty$-norms, TV semi-norm and nuclear norm. See the descriptions below, and also [6], [3], [7]. Given $x \in \mathbb{R}^n$, let $s = \mathrm{sign}(x)$ and ∇ be the gradient operator,

$$\begin{aligned}
\ell_1 &: \mathcal{M}_x = \{u \in \mathbb{R}^n : \mathrm{supp}(u) \subseteq \mathrm{supp}(x)\},\\
\ell_{1,2} &: \mathcal{M}_x = \{u \in \mathbb{R}^n : I_u \subseteq I_x\},\ I_x = \big\{i :\ x_{b_i} \neq 0\big\},\\
\ell_\infty &: \mathcal{M}_x = \big\{u : u_I = rs_I,\ r \in \mathbb{R}\big\},\ I = \{i : |x_i| = \|x\|_\infty\},\\
\mathrm{TV} &: \mathcal{M}_x = \big\{u \in \mathbb{R}^n : \mathrm{supp}(\nabla u) \subseteq I\big\},\ I = \mathrm{supp}(\nabla x),\\
\text{Nuclear} &: \mathcal{M}_x = \{u \in \mathbb{R}^{n_1\times n_2} : \mathrm{rank}(u) = \mathrm{rank}(x) = r\},
\end{aligned}$$

A desirable property of partly smooth functions is that the underlying low-dimensional smooth manifold "attracts" the iterates generated by first-order algorithms, as we show below for the PD algorithm.

Theorem II.2 (Finite activity identification). *Suppose that the PD algorithm (3) is run such that $(x_k, v_k) \to (x^\star, x^\star) \in \mathcal{X}\times\mathcal{V}$. Assume moreover that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star})$, $J \in \mathrm{PSF}_{v^\star}(\mathcal{M}^{J^*}_{v^\star})$, and*

$$\begin{aligned} -L^*v^\star - \nabla F(x^\star) &\in \mathrm{ri}\big(\partial R(x^\star)\big),\\ Lx^\star - \nabla G^*(v^\star) &\in \mathrm{ri}\big(\partial J^*(v^\star)\big).\end{aligned} \tag{4}$$

Then, the PD algorithm has the finite activity identification property, i.e. for all k sufficiently large, $(x_k, v_k) \in \mathcal{M}^R_{x^\star} \times \mathcal{M}^{J^*}_{v^\star}$.

Condition (4) can be viewed as a geometric generalization of the strict complementarity of non-linear programming.

III. Local Convergence of PD

We now turn to local linear convergence properties of PD. To deliver the result, we need to define an augmented variable $\boldsymbol{z}_k = \binom{x_k}{v_k}$.

Theorem III.1. *Suppose that the PD algorithm (3) is run under the assumptions of Theorem II.2, then there exists a matrix $\boldsymbol{M}$ such that for all k large enough, the iteration (3) can be written as*

$$\boldsymbol{z}_{k+1} - \boldsymbol{z}^\star = \boldsymbol{M}(\boldsymbol{z}_k - \boldsymbol{z}^\star) + o(\|\boldsymbol{z}_k - \boldsymbol{z}^\star\|), \tag{5}$$

where $\boldsymbol{M}$ is convergent (i.e. $\lim_{k\to\infty}\boldsymbol{M}^k$ exists). Then,

1) *given any $\rho \in]\rho(\boldsymbol{M} - \boldsymbol{M}^\infty), 1[$, there exists a K large enough such that for all $k \geq K$,*

$$\|(\mathbf{Id} - \boldsymbol{M}^\infty)(\boldsymbol{z}_k - \boldsymbol{z}^\star)\| = O(\rho^{k-K}). \tag{6}$$

2) *If moreover, R, J^* are locally polyhedral around $(x^\star, v^\star)$, there exists a K large enough such that for all $k \geq K$,*

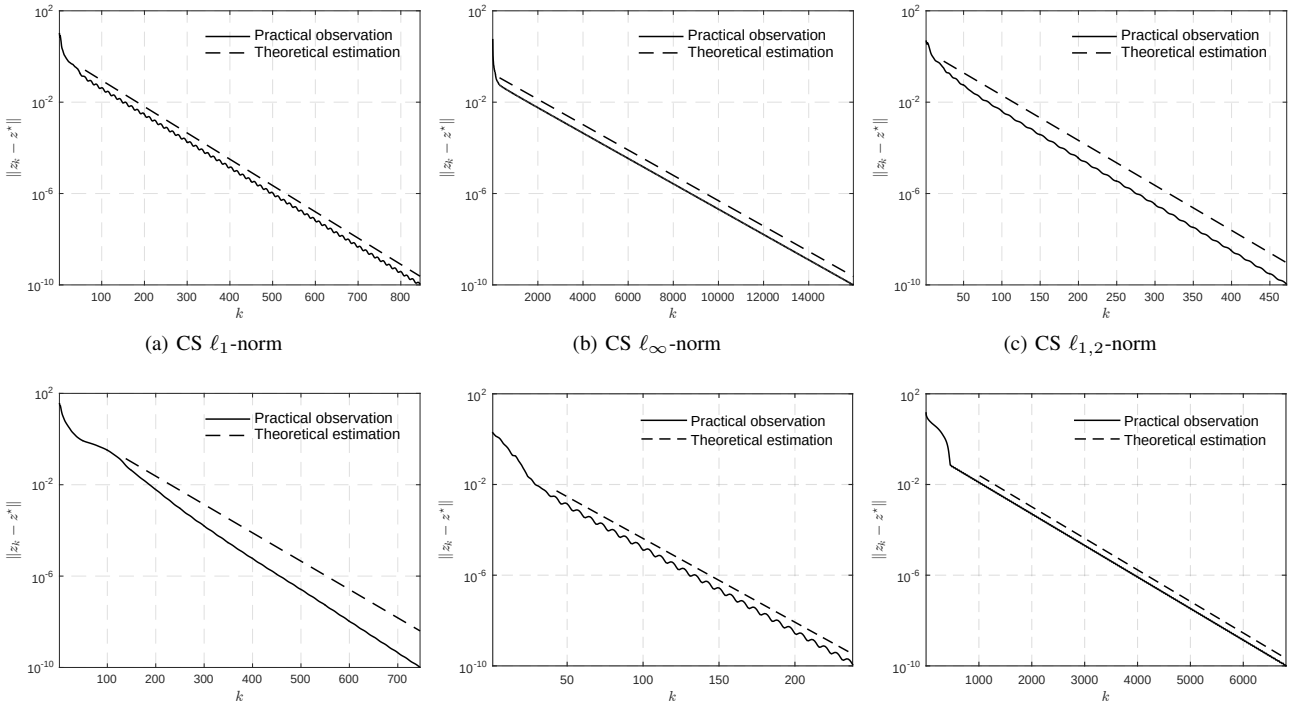
$$\|\boldsymbol{z}_k - \boldsymbol{z}^\star\| = O(\rho^{k-K}),\ \rho \in [\rho(\boldsymbol{M} - \boldsymbol{M}^\infty), 1[. \tag{7}$$

The result (7) holds also for $\|x_k - x^\star\|$ and $\|v_k - v^\star\|$.

If F, G^* vanish and R, J^* are locally polyhedral around $(x^\star, v^\star)$, then the convergence rate can be given in terms of the principal angle between the tangent spaces $T^R_{x^\star}$ and $T^{J^*}_{v^\star}$ [7].

IV. Numerical Experiments

To demonstrate the above result, we consider several problem instances, including compressed sensing, denoising and inpainting, which are well fitted for PD algorithm. The observed and predicted convergence profiles of PD are shown in Figure 1.



(a) CS ℓ_1-norm (b) CS ℓ_∞-norm (c) CS $\ell_{1,2}$-norm

(d) CS Nuclear norm (e) Uniform noise removal (f) Outliers removal

Fig. 1. Observed (solid) and predicted (dashed) convergence profiles of PD (3) in terms of $\|\boldsymbol{z}_k - \boldsymbol{z}^\star\|$. It can be observed that the rate estimates we obtain are very sharp, especially for the case where all the involved functions are polyhedral (*e.g.* figure (a), (b) , (e) and (f)). For the first 4 subfigures, we solve a problem of the form $\min_{x\in\mathbb{R}^n} J(x) + \iota_{\{0\}}(Lx - b)$ whose saddle point problem reads $\min_{x\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} J(x) + \langle Lx - b, v\rangle - \iota^*_{\{0\}}(v)$, where L is either drawn randomly from the standard Gaussian ensemble (CS) or random binary (inpainting), and $\iota_{\{0\}}(\cdot)$ is the indicator function. (a) CS with $R = \|\cdot\|_1$, $L \in \mathbb{R}^{48\times 128}$. (b) CS with $R = \|\cdot\|_\infty$, $L \in \mathbb{R}^{63\times 64}$. (c) CS with $R = \|\cdot\|_{1,2}$, $L \in \mathbb{R}^{48\times 128}$. (d) Nuclear norm, $L \in \mathbb{R}^{500\times 1024}$. For the two TV denoising examples, we consider the problem $\min_{x\in\mathbb{R}^n} \|x\|_{\mathrm{TV}}$ subject to $\|b - x\|_p \leq \tau$ whose corresponding saddle point problem reads $\min_{x\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} \iota_{\|b-\cdot\|_p\leq\tau} + \langle D_{\mathrm{DIF}}\, x, v\rangle - \iota_{\|\cdot\|_\infty\leq 1}(v)$. (e) Uniform noise removal with $p = \infty$. (f) Outliers removal with $p = 1$. The starting points of the dashed lines are the iteration at which the active manifolds are identified.

ACKNOWLEDGMENT

This work has been partly supported by the European Research Council (ERC project SIGMA-Vision). JF was partly supported by Institut Universitaire de France.

REFERENCES

[1] 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.

[2] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003.

[3] S. Vaiter, C. Deledalle, J. M. Fadili, G. Peyré, and C. Dossal. The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 2015. to appear.

[4] P. L. Combettes and B. C. Vũ. Variable metric Forward–Backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[5] L. Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, pages 1–20, 2012.

[6] S. Vaiter, G. Peyré, and M. J. Fadili. Model consistency of partly smooth regularizers. Technical Report arXiv:1307.2342, submitted, 2015.

[7] J. Liang. Convergence Rates of First-order Splitting Methods. PhD Thesis, Normandie Université, GREYC CNRS UMR 6072, 2016.

[8] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, pages 1–15, 2011.

[9] N. Komodakis, and J.-C. Pesquet. Playing with Duality: An overview of recent primal–dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32.6 (2015): 31-54.

[10] J. Liang, J. Fadili, and G. Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1):403–434, September 2016.

[11] S. Boyd, and L. Vandenberghe. Convex Optimization. *Cambridge university press*, 2004.