On the Convergence Rates of Proximal Splitting Algorithms
Jingwei Liang, Jalal M. Fadili, Gabriel Peyré

To cite this version:

HAL Id: hal-01080059
https://hal.archives-ouvertes.fr/hal-01080059
Submitted on 4 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ON THE CONVERGENCE RATES OF PROXIMAL SPLITTING ALGORITHMS

Jingwei Liang¹, Jalal M. Fadili³ and Gabriel Peyré²

¹ GREYC, CNRS-ENSICAEN  ² CEREMADE, CNRS-Paris-Dauphine

ABSTRACT

In this work, we first provide iteration–complexity bounds (pointwise and ergodic) for the inexact Krasnosel’skiĭ–Mann iteration built from nonexpansive operators. Moreover, under an appropriate regularity assumption on the fixed point operator, local linear convergence rate is also established. These results are then applied to analyze the convergence rate of various proximal splitting methods in the literature, which includes the Forward–Backward, generalized Forward–Backward, Douglas–Rachford, ADMM and some primal–dual splitting methods. For these algorithms, we develop easily verifiable termination criteria for finding an approximate solution, which is a generalization of the termination criterion for the classical gradient descent method. We illustrate the usefulness of our results on a large class of problems in signal and image processing.

Index Terms— Convex optimization, Proximal splitting, Convergence rates, Inverse problems.

1. INTRODUCTION

Problem statement and overview Many convex optimization problems in image and signal processing can be solved via the inexact Krasnosel’skiĭ–Mann iteration which is defined by

\[ x^{k+1} = x^k + \lambda_k (T x^k + \varepsilon^k - x^k), \]  

(1.1)

where \( T : \mathcal{H} \rightarrow \mathcal{H} \) is a nonexpansive operator on a Hilbert space \( \mathcal{H} \), \( (\lambda_k)_{k \in \mathbb{N}} \in [0, 1] \), and for \( x^k \in \mathcal{H} \), \( \varepsilon^k \in \mathcal{H} \) is the error when computing \( T x^k \).

One simple instance of such an algorithm is that of relaxed inexact gradient descent for solving

\[ \min_{x \in \mathcal{H}} f(x), \]

where \( f \) is proper convex and has \( \beta^{-1} \)-Lipschitz continuous gradient, in which case \( T = \text{Id} - \gamma \nabla f \), for \( \gamma \in [0, 2\beta] \). In this scenario, the error \( \varepsilon^k \) is that when evaluating \( \nabla f \) at \( x^k \).

Many structured convex optimization problems boil down to implementing an iteration that can be cast in the form of (1.1) for an appropriate \( T \). Consider for example

\[ \min_{x \in \mathcal{H}} \{ \Phi(x) = f(x) + \sum_{i=1}^{n} h_i \circ L_i(x) \}, \]

(1.2)

where \( f \in \Gamma_0(\mathcal{H}) \) has \( \beta^{-1} \)-Lipschitz continuous gradient, \( h_i \in \Gamma_0(\mathcal{H}) \), \( L_i \) is a bounded linear operator, \( \Gamma_0(\mathcal{H}) \) is the class of lower semicontinuous, proper, convex functions from \( \mathcal{H} \) to \( (-\infty, +\infty] \). Assume that some appropriate domain qualification conditions are verified for (1.2) to be well–posed. Problem (1.2) has been considered in e.g. [1, 2, 3], who proposed iterative schemes in the form (1.1).

In the last decades, based on the notion of the proximity operator [4], and assuming that the functions \( h_i \) are simple (i.e. their proximity operators are easily computable), a wide range of proximal splitting algorithms have been proposed to solve problems of the form (1.2). One can cite for instance the Forward–Backward splitting method (FBS) [5] valid for \( n = 1 \) and \( L_1 = \text{Id} \), the Douglas–Rachford splitting method (DRS) [6] that applies for \( f = 0 \) and \( L_i = \text{Id} \), generalized Forward–Backward splitting method (GFB) [3] when \( L_i = \text{Id} \), or primal–dual splitting methods. See [7] for a comprehensive account.

Contributions In this paper, we first establish iteration complexity bounds of the inexact relaxed fixed point iteration (1.1). Under a regularity assumption on \( \text{Id} - T \), a local linear convergence rate is established. We then build upon these results to provide rates for several proximal splitting algorithms, from which an easily verifiable termination criterion for finding an approximate solution will be given. For space limitations, the proofs of the results can be found in the long version of the paper [8].

2. ITERATION COMPLEXITY BOUNDS

Denote \( T' = \text{Id} - T \), \( \text{fix} T \) is the set of fixed points of \( T \), and

\[ e^k = T' x^k. \]

Let \( d_0 \) be the distance from a starting point \( x^0 \) to the solution set \( \text{fix} T \). Denote \( \tau_k = \lambda_k (1 - \lambda_k) \), \( \tau = \inf_{k \in \mathbb{N}} \tau_k \), \( \overline{\tau} = \sup_{k \in \mathbb{N}} \tau_k \), \( \nu_1 = 2 \sup_{k \in \mathbb{N}} \|[T_k x^k - x^*]\| + \sup_{k \in \mathbb{N}} \| \lambda_k e^k \| \), \( \nu_2 = 2 \sup_{k \in \mathbb{N}} \| e^k - e^{k+1} \| \), where \( x^* \in \text{fix} T \). Denote \( \ell_+ \) the set of summable sequences in \( [0, +\infty[ \).

This work has been supported by the ERC project SIGMA-Vision and l’Institut Universitaire de France.
Theorem 2.1 (Pointwise iteration complexity bound). Assume that
(a) \( \text{fix} T \neq \emptyset \);
(b) \( 0 < \inf k \in N \lambda_k \leq \sup k \in N \lambda_k < 1 \);
(c) \((k+1)\|e^k\| \in \ell^+_1\).

Let \( C_1 = \nu_1 \sum_{j \in N} \lambda_j \|e^j\| + \nu_2 T \sum_{\ell \in N}(\ell + 1)\|e^j\| < +\infty \), then
\[ \|e^k\| \leq \sqrt{\frac{d_0^2 + C_1}{\sum_{j=0}^k \ell_j}}. \]

Remark 2.2. When the fixed point iteration (1.1) is exact, we get
\[ \|e^k\| \leq \frac{d_0}{\sqrt{\sum_{j=0}^k \ell_j}}, \]
which recovers the result of [9, Proposition 11].

Next we present the ergodic iteration complexity bound of (1.1), define \( \Lambda_k = \sum_{j=0}^k \lambda_j \), and \( e^k = \frac{1}{\Lambda_k} \sum_{j=0}^k \lambda_j e^j \).

Theorem 3.1 (Metric subregularity [10]). A set–valued mapping \( F : \mathcal{H} \to 2^\mathcal{H} \) is called metrically subregular at \( \hat{x} \) for \( \hat{u} \in F(\hat{x}) \) if there exists \( \kappa \geq 0 \) along with neighbourhood \( \mathcal{X} \) of \( \hat{x} \) such that
\[ d(\hat{x}, F^{-1}\hat{u}) \leq \kappa d(\hat{x}, \hat{x}), \quad \forall x \in \mathcal{X}. \]

Remark 2.3. When the fixed point iteration (1.1) is exact, we have
\[ d(\hat{x}, \text{fix} T) \leq \kappa \|T^\kappa \hat{x}\|, \quad \forall x \in \mathcal{X}. \]

3. LOCAL LINEAR RATE

We now turn to a local convergence analysis of (1.1).

Definition 3.1 (Metric subregularity [10]). A set–valued mapping \( F : \mathcal{H} \to 2^\mathcal{H} \) is called metrically subregular at \( \hat{x} \) for \( \hat{u} \in F(\hat{x}) \) if there exists \( \kappa \geq 0 \) along with neighbourhood \( \mathcal{X} \) of \( \hat{x} \) such that
\[ d(\hat{x}, F^{-1}\hat{u}) \leq \kappa d(\hat{x}, \hat{x}), \quad \forall x \in \mathcal{X}. \]

Theorem 3.2 (Local linear rate). Assume that conditions (a-b) in Theorem 2.1 hold, and \( C_3 = \sum_{j \in N} \lambda_j \|e^j\| < +\infty \). Then,
\[ \|e^k\| \leq \frac{2(d_0 + C_3)}{\Lambda_k}. \]

Let’s specialize this definition to the operator \( T^\kappa \) and \( \hat{u} = 0 \). \( T^\kappa \) is single–valued and \( T^{\kappa-1}(0) = \text{fix} T \). Thus if \( T^\kappa \) is metrically subregular at some \( x^* \in \text{fix} T \) for 0, then from (3.1) we have
\[ d(x, \text{fix} T) \leq \kappa \|T^\kappa x\|, \quad \forall x \in \mathcal{X}. \]

Metric subregularity implies that (3.1) gives an estimate for how far a point \( x \) is from being the fixed point set of \( T \) in terms of the residual \( \|x - T x\| \). This is the rationale behind using such a regularity assumption on the operator \( T^\kappa \) to quantify the convergence rate on \( d(x^k, \text{fix} T) \). Thus, starting from \( x^0 \in \mathcal{H} \), and by virtue of Theorem 2.1, one can recover an \( O(1/\sqrt{k}) \) rate on \( d(x^k, \text{fix} T) \). In fact, we can do even better than shown in the following result. We use the shorthand notation \( d_k = d(x^k, \text{fix} T) \).

Theorem 3.2 (Local linear rate). Assume that conditions (a-b) in Theorem 2.1 hold, and \( T^\kappa \) is metrically subregular at \( x^* \in \text{fix} T \) for 0, with \( \kappa > \text{subreg}(T^\kappa; x^*; 0) \). If \( C_3 \) is sufficiently small and there exists a ball \( B_0(\mathcal{X}) \) such that
\[ B_\kappa(\mathcal{X}) \subseteq \mathcal{X}. \]

Then, for any starting point \( x^0 \in B_0(\mathcal{X}) \), there holds:
(i) \( (d_k^2)_{k \in N} \in \ell^+_1 \) and
\[ d_{k+1}^2 \leq \kappa d_k^2 + c_k, \]
where \( \kappa = \left\{ \begin{array}{ll} 1 - \frac{\tau_k}{\kappa^2}, & \text{if } \tau_k/\kappa^2 \in [0, 1] \in [0, 1], \ c_k = \\
\nu_1 \lambda_k \|e^k\|. & \text{otherwise} \end{array} \right. \)
(ii) If \( e^k = 0 \), then \( \lim_{k \to +\infty} \sqrt{k} d_k < 1 \).
(iii) If \( \text{fix} T = \{x^*\} \) and \( e^k = 0 \), then \( x^k \) converges linearly to \( x^* \).

Remark 3.3. For simplicity, suppose the iteration is exact, and let \( x^k \in \text{fix} T \) such that \( d_k = \|x^k - x^*\| \). Then we have
\[ \|e^k\|^2 = \|x^k - x^* + T x^k - T x^k\|^2 \leq 4d_k^2 \leq 4\kappa^2 d_0^2, \]
which means that locally, \( \|e^k\| \) also converges linearly to 0.

Again, if \( T \) is \( \alpha \)-averaged, then we can afford \( \lambda_k \in [0, 1/\alpha] \) and all statements of Theorem 3.2 remain valid with \( \zeta_k = \kappa^2/(\kappa^2 + \alpha \lambda_k (1 - \alpha \lambda_k)). \)

Since metric regularity [10] implies metric subregularity, equivalent characterizations of the latter and its modulus can be given, for instance in terms of derivative criteria. In particular, as \( T^\kappa \) is single–valued, its metric regularity holds if it is differentiable on a neighbourhood of \( x^* \) with nonsingular derivatives at \( x \) around \( x^* \), and the operator norms of their inverses are uniformly bounded and serve as an estimate of the (sub)regularity modulus \( \kappa \) [11, Theorem 1.2]. Computing \( \kappa \) in practice is however far from obvious in general even in for the differentiable case and these details will be left to a future work.
4. APPLICATIONS TO PROXIMAL SPLITTING

FBS Suppose that \( n = 1 \) and \( L_1 = \text{Id} \) in (1.2). Then FBS with a fixed step–size corresponds to \( T = \text{Prox}_{\cdot h_1} \circ (\text{Id} - \gamma \nabla f) \), \( \gamma \in [0, 2\beta[ \), \( \lambda_k \in [0, \frac{4\beta - \gamma}{2\beta}] \), and \( \varepsilon^k \) is the error when evaluating both the proximity operator and the gradient. In this case, setting \( g^k = \frac{1}{2}(x^{k+1} - x^k) - \nabla f(x^{k+1}) \) it then follows from Theorem 2.1 that
\[
g^k \in \partial h_1(x^k), \quad \| g^k + \nabla f(x^k) \|^2 = O(1/k).
\]

In plain words, this means that \( O(1/\varepsilon) \) iterations are needed to find a pair \( (x, g \in \partial h_1(x)) \) with the termination criterion
\[
\| g + \nabla f(x) \|^2 \leq \varepsilon.
\]
This is the best–known complexity bound possessed by first–order methods to solve general non-linear problems (e.g. the gradient descent method [12]). From Theorem 2.3, this iteration–complexity improves to \( O(1/\sqrt{\varepsilon}) \) in ergodic sense for the same termination criterion.

GFB Consider now \( n > 1 \) and \( L_i = \text{Id} \) in (1.2). [3] proposed a generalization in a product space \( \mathcal{H}^n \) of the FBS scheme to solve (1.2). For the GFB method, \( T \) takes a more intricate form, omitted here for space limitation, and \( \varepsilon^k \) absorbs the error when computing both the proximity operators \( \text{Prox}_{\cdot h_i} \), and the gradient. Let \( \gamma \) and \( \lambda_k \) be chosen as for FBS. One can show that \( O(1/\varepsilon) \) iterations are needed to find a pair \( (\{u_i\}_{1 \leq i \leq n}, g) \) with the termination criterion
\[
\| g + \nabla f(\Sigma_i u_i/n) \|^2 \leq \varepsilon, \quad \text{where} \ g \in \Sigma_i \partial h_i(u_i), \text{see [8] for details. Again, ergodic iteration complexity is} \ O(1/\sqrt{\varepsilon}).
\]

DRS, ADMM When \( f = 0 \) and \( n = 2 \), (1.2) can be solved by DRS for \( L_1 = L_2 = \text{Id} \), or the Alternating Direction method of Multipliers (ADMM) [13] if e.g. \( L_1 = \text{Id} \). ADMM is known to be equivalent to DRS applied to the dual problem [14]. It can be shown that in at most \( O(1/\varepsilon) \) iterations, DRS finds a subgradient \( g \) of \( h_1 + h_2 \) at \( x \) with \( \| g \|^2 \leq \varepsilon \).

Other splitting schemes Similar complexity bounds can be established for several primal–dual splitting methods, e.g. [1], for solving (1.2) (and even more general). One crucial property of these methods, is that they can be reformulated into the form of (1.1), and the corresponding fixed point operator is \( \alpha \)–averaged; see [8] for details.

5. NUMERICAL EXPERIMENTS

In this section, we illustrate the obtained convergence results on two applications, the nonnegative matrix completion (NMC) problem [15], and the principal component pursuit (PCP) problem [16].

5.1. Nonnegative matrix completion

Suppose we observe measurements \( y \in \mathbb{R}^p \) of a low rank matrix \( X_0 \in \mathbb{R}^{m \times n} \) with nonnegative entries
\[
y = A(X_0) + w
\]
where \( A : \mathbb{R}^{m \times n} \to \mathbb{R}^p \) is a measurement operator, and \( w \) is the noise. In our experiment here, \( A \) selects \( p \) entries of its argument uniformly at random. The matrix completion problem consists in recovering \( X_0 \), or an approximation of it, by solving a convex optimization problem, namely the minimization of the nuclear norm \([17, 18, 15]\). In penalized form, the problem reads
\[
\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2}\| y - A(X) \|^2 + \mu \| X \|_* + \iota_+(X), \quad (5.1)
\]
where \( \iota_+ \) is the indicator function of the nonnegative orthant to account for the nonnegativity constraint, and \( \mu \) is a regularization parameter chosen proportional to the noise level. Identifying \( f(X) = \frac{1}{2}\| y - A(X) \|^2 \), \( h_1(X) = \mu \| X \|_* \), and \( h_2(X) = \iota_+(X) \), (5.1) is nothing but an instance of (1.2). Both \( h_1 \) and \( h_2 \) are simple, since \( \text{Prox}_{\cdot h_i} (X) \) amounts to soft–thresholding the singular values of \( X \), and \( \text{Prox}_{\cdot h_2} (X) = (\max(\Sigma_{i \in [1]} X_{ij}^0, 0)_{i,j}) \) is the projector on the nonnegative orthant. Thus, (5.1) can be solved using e.g. GFB or the primal–dual (PD) scheme of [1].

Fig. 1: Observed convergence curve (solid blue) of \( \| e^k \| \) for GFB (left) and PD (right) to solve (5.1), theoretical global bound (solid red), and fitted local linear curve.
5.2. Principal component pursuit

In this part, we consider the PCP problem [16], and apply it to decompose a video sequence into its background and foreground components. The rationale behind this is that since the background is virtually the same in all frames, if the latter are stacked as columns of a matrix, it is likely to be low–rank (even of rank 1 for perfectly constant background). On the other hand, moving objects appear occasionally on each frame and occupy only a small fraction of it. Thus the corresponding component would be sparse.

Assume that a real matrix \( M \) can be written as

\[
M = X_{1,0} + X_{s,0} + W,
\]

where a \( X_{1,0} \) is low–rank, \( X_{s,0} \) is sparse and \( W \) is a perturbation matrix that accounts for model imperfection. The PCP proposed in [16] attempts to provably recover \( (X_{1,0},X_{s,0}) \), to a good approximation, by solving a convex optimization. Here, toward an application to video decomposition, we also add a non-negativity constraint to the low–rank component, which leads to the convex problem (\( \| \cdot \|_F \) is the Frobenius norm)

\[
\min_{X_1,X_s} \frac{1}{2} \| M - X_1 - X_s \|_F^2 + \mu_1 \| X_s \|_1 + \mu_2 \| X_1 \|_s + \iota_+ (X_1),
\]

(5.2)

One can observe that for fixed \( X_1 \), the minimizer of (5.2) is \( X_s^* = \text{Prox}_{\mu_1 \| \cdot \|_1} (M - X_1) \). Thus, (5.2) is equivalent to

\[
\min_{X_1} \frac{1}{2} \mu_1 \| \cdot \|_1 \| M - X_1 \|_F^2 + \mu_2 \| X_1 \|_s + \iota_+ (X_1),
\]

(5.3)

where \( \frac{1}{2} \mu_1 \| \cdot \|_1 \| M - X_1 \|_F^2 + \mu_1 \| Z \|_1 \) is the Moreau Envelope of \( \mu_1 \| \cdot \|_1 \) of index 1. Since the Moreau envelope is differentiable with a 1–Lipschitz continuous gradient [19], (5.3) is a special instance of (1.2) and can be solved using GFB and PD schemes.

We first used a synthetic example to illustrate the convergence property of GFB and PD. Fig. 2 displays the observed pointwise convergence rate of \( \| e^k \| \) for the GFB to solve (5.3) with synthetic data, theoretical global bound (solid red), and fitted local linear curve.

Now we consider a video sequence introduced in [20], whose resolution is \( 128 \times 160 \), each frame is stacked as a column of the matrix \( M \) and 300 frames in total. Hence \( M \) is of size \( 20480 \times 300 \). We then solve (5.3) to decompose the video into its foreground and background.

Fig. 3 displays the observed pointwise convergence rate and the one predicted by Theorem 2.1. Both the global and local convergence behaviours are similar to those observed in Fig. 1.

Fig. 2: Observed convergence curve (solid blue) of \( \| e^k \| \) for GFB (left) and PD (right) to solve (5.3) with synthetic data, theoretical global bound (solid red), and fitted local linear curve.

Fig. 3: Observed convergence curve (solid blue) of \( \| e^k \| \) for the GFB to solve (5.3) with a real video sequence, theoretical global bound (solid red), and fitted local linear curve.

Fig. 4: Left: Original frames of a video sequence (300 frames). Middle and Right: recovered background (low–rank) and foreground (sparse) components.
6. REFERENCES


