Modified Fejér sequences and applications

Citation: Lin, Junhong, et al. “Modified Fejér Sequences and Applications.” Computational Optimization and Applications, vol. 71, no. 1, Sept. 2018, pp. 95–113.

As Published: https://doi.org/10.1007/s10589-017-9962-1

Publisher: Springer US

Persistent URL: http://hdl.handle.net/1721.1/117359

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Modified Fejér Sequences and Applications[*]

Junhong Lin[†], Lorenzo Rosasco[†, °], Silvia Villa[⋆], and Ding-Xuan Zhou[*]

[†] *LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[⋆] *Dipartimento di Matematica, Politecnico di Milano, Milano 20133, Italy*
[°] *DIBRIS, Universitá degli Studi di Genova, Genova 16146, Italy*
[*] *Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China*

October 30, 2017

Abstract

In this note, we propose and study the notion of modified Fejér sequences. Within a Hilbert space setting, this property has been used to prove ergodic convergence of proximal incremental subgradient methods. Here we show that indeed it provides a unifying framework to prove convergence rates for objective function values of several optimization algorithms. In particular, our results apply to forward-backward splitting algorithm, incremental subgradient proximal algorithm, and the Douglas-Rachford splitting method including and generalizing known results.

[*]This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. The work by D. X. Zhou described in this paper is supported by a grant from the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006]. L. R. acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC, and S. V. the support of INDAM-GNAMPA, project 2017: "Algoritmi di ottimizzazione ad equazioni di evoluzione ereditarie".

1 Introduction

We are interested in the study of convergence properties of optimization algorithms to solve the problem

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

where $\mathcal{H}$ is a Hilbert space and $f\colon \mathcal{H} \to]-\infty, +\infty]$ is a proper function. Let $(x_t)_{t\in\mathbb{N}}$ be the sequence generated by a chosen algorithm. The sequence is said to be Fejér monotone if, for every x_* minimizer of f, $\|x_{t+1}-x_*\|^2 \leq \|x_t - x_*\|^2$. The notion of Fejér monotonicity captures essential properties of $(x_t)_{t\in\mathbb{N}}$ generated by a wide range of optimization methods and provides a common framework to analyze their convergence [11]. Quasi-Fejér monotonicity is a relaxation of the above notion that allows for an additional error term [13, 22]. Generalizations of the above notion have been proposed, that allow to deal with variable metric algorithms [18, 40], and stochastic perturbations [15, 16, 22, 37, 38, 39].

In this paper, we propose and study a novel, related notion, to analyze the convergence of the objective function values $f(x_t)$, in addition to that of the iterates. More precisely, we modify the notion of quasi-Fejér monotonicity, by adding a term involving the objective function and say that a sequence satisfying the new requirement is modified Fejér monotone (modified Fejér for short). This property is the key step to derive ergodic convergence of the iterates generated by the proximal incremental gradient algorithm in [7]. In this paper, we show the wider usefulness of this new notion of monotonicity by deriving convergence rates for several optimization algorithms in a unified way. Based on this approach, we not only recover known results, such as the sublinear convergence rate for the proximal forward-backward splitting algorithm, but also derive new results. Interestingly, our results show that for projected subgradient, incremental proximal subgradient, and Douglas-Rachford algorithms, considering the last iterate leads to essentially the same convergence rate as considering the best iterate selection rule [36, 41], or ergodic means [8, 42], as typically done.

2 Modified Fejér Sequences

Throughout this paper, we assume that $\mathcal{H}$ is a Hilbert space, and $f : \mathcal{H} \to]-\infty, \infty]$ is a proper function. We assume that the set of minimizers of f

$$\mathcal{X} = \{z \in \mathcal{H} \,|\, f(z) = \min_{x \in \mathcal{H}} f(x)\}$$

is nonempty. We are interested in solving the following optimization problem

$$f_* = \min_{x \in \mathcal{H}} f(x). \tag{1}$$

Given $x \in \mathcal{H}$ and a subset $S \subset \mathcal{H}$, $d(x, S)$ denotes the distance between x and S, i.e., $d(x, S) = \inf_{x' \in S} \|x - x'\|$. $\mathbb{R}_+$ is the set of all non-negative real numbers and $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$.

The following definition introduces the key notion we propose in this paper.

Definition 1. *A sequence $(x_t)_{t\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ is modified Fejér monotone with respect to the objective function f and the sequence $((\eta_t, \xi_t))_{t\in\mathbb{N}}$ in $\mathbb{R}^2_+$, if*

$$(\forall x \in \mathrm{dom} f) \qquad \|x_{t+1} - x\|^2 \le \|x_t - x\|^2 - \eta_t(f(x_{t+1}) - f(x)) + \xi_t. \tag{2}$$

Remark 1.

(i) Choosing $x \in \mathcal{X}$ in (2), we get

$$\eta_t f(x_{t+1}) \le \xi_t + \eta_t f_* + \|x_t - x\|^2 < \infty.$$

This implies that $x_t \in \mathrm{dom} f$ for every $t \in \mathbb{N}$.

(ii) All the subsequent results hold if condition (2) *is replaced by the following weaker condition*

$$(\forall x \in \mathcal{X} \cup \{x_t\}_{t\in\mathbb{N}}) \quad \|x_{t+1} - x\|^2 \le \|x_t - x\|^2 - \eta_t(f(x_{t+1}) - f(x)) + \xi_t. \tag{3}$$

However, in the proposed applications, condition (2) *is always satisfied for every $x \in \mathrm{dom} f$.*

(iii) Inequality (2) *has been proved to be satisfied and implicitly used to derive convergence rate for many algorithms, considering the best iterate selection rule, e.g., [36, 41], or ergodic means [8, 42, 7]. More precisely, for not descending methods, common approaches keep track of the best point found so far, i.e. they study the following quantity,*

$$(\forall T \in \mathbb{N}^*) \quad b_T = \min_{1 \le t \le T} f(x_t) - f_*,$$

or this one:

$$(\forall T \in \mathbb{N}^*) \quad f\left(\sum_{t=1}^{T} x_t/T\right) - f_*.$$

The main novelty of this paper is to show that considering the last iterate, i.e. $f(x_T) - f_*$, leads to essentially the same convergence rate as that for considering the best iterate selection rule, or ergodic means. See Theorem 2.

(iv) A similar condition has been considered in [1] to study convergence properties of several optimization methods, and in [40] to study a variable metric forward-backward algorithm under relaxed differentiability assumptions.

(v) Extensions of the proposed notion could be considered, e.g. resembling variable metric and stochastic quasi-Fejér monotonicity properties, which have been recently proposed and investigated in [18, 15, 37, 39].

In the following remark we discuss the relation with classical Fejér sequences.

Remark 2 (Comparison with quasi-Fejér sequences).
If $\sum_{t\in\mathbb{N}} \xi_t < +\infty$, Definition 1 implies that the sequence $(x_t)_{t\in\mathbb{N}}$ is quasi-Fejér monotone with respect to $\mathcal{X}$ [13, 22]. Indeed, (2) implies

$$(\forall x \in \mathcal{X}) \qquad \|x_{t+1} - x\|^2 \le \|x_t - x\|^2 + \xi_t.$$

Note that, in the study of convergence properties of quasi-Fejér sequences corresponding to a minimization problem, the property is considered with *respect to the set of solutions $\mathcal{X}$, while here we will consider modified Fejér monotonicity for the entire space $\mathcal{H}$.*

We next present two main results to show how modified Fejér sequences are useful to study the convergence of optimization algorithms. The first result shows that if a sequence is modified Fejér monotone, one can bound its corresponding excess function values in terms of $((\eta_t, \xi_t))_{t\in\mathbb{N}}$ explicitly.

Theorem 1. *Let $(x_t)_{t\in\mathbb{N}} \subset \mathcal{H}^{\mathbb{N}}$ be a modified Fejér sequence with respect to f and $((\eta_t, \xi_t))_{t\in\mathbb{N}}$ in $\mathbb{R}^2_+$. Let $(\eta_t)_{t\in\mathbb{N}}$ be a non-increasing sequence. Let $T \in \mathbb{N}^*$. Then*

$$\eta_T(f(x_{T+1}) - f_*) \leq \frac{d(x_1, \mathcal{X})^2}{T} + \sum_{t=1}^{T-1} \frac{1}{T-t+1}\xi_t. \tag{4}$$

Proof. Let $(u_t)_{t\in\mathbb{N}}$ be a sequence in $\mathbb{R}$. For every $k \in \{1, \cdots, T-1\}$, let $s_k = \sum_{j=T-k}^{T} u_j$. Then, since $s_k = s_{k-1} + u_{T-k}$,

$$\begin{aligned}&\frac{1}{k}s_{k-1} - \frac{1}{k+1}s_k \\ =&\frac{1}{k(k+1)}\left((k+1)s_{k-1} - ks_k\right) \\ =&\frac{1}{k(k+1)}(s_{k-1} - ku_{T-k}).\end{aligned}$$

Summing over $k = 1, \cdots, T-1$, and rearranging terms, we get

$$u_T = \frac{1}{T}s_{T-1} + \sum_{k=1}^{T-1} \frac{1}{k(k+1)}(s_{k-1} - ku_{T-k}). \tag{5}$$

Let $x \in \mathrm{dom} f$ and choose $(\forall t \in \mathbb{N})$ $u_t = \eta_t(f(x_{t+1}) - f(x))$. Then, we derive the following error decomposition [28]:

$$\begin{aligned}\eta_T(f(x_{T+1}) - f(x)) &= \frac{1}{T}\sum_{t=1}^{T} \eta_t(f(x_{t+1}) - f(x)) \\ &+ \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k+1}^{T} \eta_t(f(x_{t+1}) - f(x_{T-k+1})) \\ &+ \sum_{k=1}^{T-1} \frac{1}{k+1}\left[\left(\frac{1}{k}\sum_{t=T-k+1}^{T} \eta_t\right) - \eta_{T-k}\right]\left(f(x_{T-k+1}) - f(x)\right).\end{aligned}$$

Let $x = x_* \in \mathcal{X}$. Since $(\eta_t)_{t\in\mathbb{N}}$ is non-increasing and $f(x_{T-k+1}) - f_* \geq 0$, the last term of the above inequality is less than or equal to 0. Thus, we derive that

$$\begin{aligned} \eta_T(f(x_{T+1}) - f_*) \leq \frac{1}{T}\sum_{t=1}^{T}\eta_t(f(x_{t+1}) - f(x_*)) \\ +\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\eta_t(f(x_{t+1}) - f(x_{T-k+1})). \end{aligned} \tag{6}$$

For every $j \in \{1, \ldots, T\}$, and for every $x \in \mathrm{dom} f$, summing up (2) over $t = j, \cdots, T$, we get

$$\sum_{t=j}^{T}\eta_t(f(x_{t+1}) - f(x)) \leq \|x_j - x\|^2 + \sum_{t=j}^{T}\xi_t. \tag{7}$$

The above inequality with $x = x_*$ and $j = 1$ implies

$$\frac{1}{T}\sum_{t=1}^{T}\eta_t(f(x_{t+1}) - f(x_*)) \leq \frac{1}{T}\|x_1 - x_*\|^2 + \frac{1}{T}\sum_{t=1}^{T}\xi_t. \tag{8}$$

Inequality (7) with $x = x_{T-k+1}$ and $j = T - k + 1$ yields

$$\sum_{t=T-k+1}^{T}\eta_t(f(x_{t+1}) - f(x_{T-k+1})) \leq \sum_{t=T-k+1}^{T}\xi_t,$$

and thus

$$\begin{aligned} &\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\eta_t(f(x_{t+1}) - f(x_{T-k+1})) \\ \leq\ &\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\xi_t. \end{aligned} \tag{9}$$

Exchanging the order in the sum, we obtain

$$\begin{aligned}\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}\xi_t &= \sum_{t=2}^{T}\sum_{k=T-t+1}^{T-1}\frac{1}{k(k+1)}\xi_t \\ &= \sum_{t=2}^{T}\left(\frac{1}{T-t+1}-\frac{1}{T}\right)\xi_t \\ &= \sum_{t=2}^{T}\frac{1}{T-t+1}\xi_t - \frac{1}{T}\sum_{t=2}^{T}\xi_t. \end{aligned} \tag{10}$$

The result follows by plugging (8) and (10) into (6). □

Remark 3. *Féjer monotonicity of a sequence is often useful not only for obtaining convergence rate estimates, but also for proving the convergence of the iterates to a minimizer, see e.g. [2, Proposition 2]. Here we mainly focus directly on convergence rates for the objective function values.*

In the special case when, for every $t \in \mathbb{N}$, $\xi_t = 0$, we derive the following result.

Corollary 1. *Let $(x_t)_{t\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ be a modified Fejér sequence with respect to f and a sequence $((\eta_t, 0))_{t\in\mathbb{N}}$ in $\mathbb{R}^2_+$. Suppose that $(\eta_t)_{t\in\mathbb{N}}$ is non-increasing. Then for any $T \in \mathbb{N}^*$,*

$$f(x_{T+1}) - f_* \leq \frac{1}{\eta_T T} d(x_1, \mathcal{X})^2.$$

In the case that ξ_t is non-increasing, we can get the following result, which simplifies the upper bound in (4) from Theorem 1.

Corollary 2. *Let $(x_t)_{t\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ be a modified Fejér sequence with respect to an objective function f and a sequence $((\eta_t, \xi_t))_{t\in\mathbb{N}}$ in $\mathbb{R}^2_+$. Suppose that $(\xi_t)_{t\in\mathbb{N}}$ and $(\eta_t)_{t\in\mathbb{N}}$ are non-increasing. Let $T \in \mathbb{N}^*$. Then*

$$f(x_{T+1}) - f_* \leq \left(d(x_1, \mathcal{X})^2 + 2\sum_{t=1}^{T}\xi_t\right)(T\eta_T)^{-1} + \xi_{\lfloor \frac{T}{2}+1\rfloor}\eta_T^{-1}\log(T/2) \tag{11}$$

Proof. Since ξ_t is non-increasing,

$$\begin{aligned}\sum_{t=1}^{T-1}\frac{\xi_t}{T-t+1} &\leq \xi_{\lfloor\frac{T}{2}+1\rfloor}\sum_{T/2+1\leq t\leq T-1}\frac{1}{T-t+1}+2T^{-1}\sum_{1\leq t<T/2+1}\xi_t\\ &\leq \xi_{\lfloor\frac{T}{2}+1\rfloor}\log(T/2)+2T^{-1}\sum_{t=1}^{T}\xi_t.\end{aligned}$$

The result follows directly from Theorem 1. □

The next main result shows how to derive explicit rates for the objective function values corresponding to a modified Fejér sequence with respect to a polynomially decaying sequence $((\eta_t,\xi_t))_{t\in\mathbb{N}}$ in $\mathbb{R}^2_+$. Interestingly, the following result (as well as the previous ones) does not require convexity of f.

Theorem 2. *Let $(x_t)_{t\in\mathbb{N}}\in\mathcal{H}^{\mathbb{N}}$ be a modified Fejér sequence with respect to an objective function f and a sequence $((\eta_t,\xi_t))_{t\in\mathbb{N}}$ in $\mathbb{R}^2_+$. Let $\eta\in\,]0,+\infty[$, let $\theta_1\in[0,1[$, and set $\eta_t=\eta t^{-\theta_1}$. Let $(\theta_2,\xi)\in\mathbb{R}^2_+$ and suppose that $\xi_t=\xi t^{-\theta_2}$ for all $t\in\mathbb{N}$. Let $T\in\mathbb{N}^*$. Then, setting $c=2^{\theta_2}+2/(1-\theta_2)$:*

$$f(x_{T+1})-f_*\leq\begin{cases}\dfrac{d(x_1,\mathcal{X})^2}{\eta}T^{\theta_1-1}+\dfrac{\xi}{\eta}\left(2^{\theta_2}+\frac{2}{1-\theta_2}\right)T^{\theta_1-\theta_2}\log T & \textit{if } \theta_2<1\\ \dfrac{d(x_1,\mathcal{X})^2}{\eta}T^{\theta_1-1}+\dfrac{4\xi}{\eta}T^{\theta_1-1}\log T & \textit{if } \theta_2=1\\ \left(\dfrac{d(x_1,\mathcal{X})^2}{\eta}+\dfrac{\xi}{(\theta_2-1)\eta}\right)T^{\theta_1-1} & \textit{otherwise.}\end{cases}\tag{12}$$

Proof. Let $q\in\,]0,+\infty[$. For $n\geq 2$ we have (see [25, Theorem 3.3.3])

$$\sum_{t=2}^{n}t^{-q}\leq\int_1^n u^{-q}du\leq\begin{cases}n^{1-q}/(1-q), & \text{when } q<1,\\ \log n, & \text{when } q=1,\\ 1/(q-1), & \text{when } q>1.\end{cases}\tag{13}$$

We derive from Corollary 2 that

$$f(x_{T+1})-f_*\leq\left(d(x_1,\mathcal{X})^2+2\xi\sum_{t=1}^{T}t^{-\theta_2}\right)\eta^{-1}T^{\theta_1-1}+2^{\theta_2}\xi\eta^{-1}T^{\theta_1-\theta_2}\log(T/2).\tag{14}$$

If $\theta_2 < 1$, equation (13) yields

$$f(x_{T+1}) - f_* \leq \eta^{-1} d(x_1, \mathcal{X})^2 T^{\theta_1 - 1} + \xi \eta^{-1} \left(\frac{2}{1-\theta_2} + 2^{\theta_2} \right) T^{\theta_1 - \theta_2} \log T. \quad (15)$$

The case $\theta_2 = 1$ is analogous. If $\theta_2 > 1$, from (13), it follows that

$$f(x_{T+1}) - f_* \leq \left(d(x_1, \mathcal{X})^2 + 2\xi (1-\theta_2)^{-1} \right) \eta^{-1} T^{\theta_1 - 1} + 2^{\theta_2} \xi \eta^{-1} T^{\theta_1 - \theta_2} \log T. \quad (16)$$

Since $2^{\theta_2} T^{-\theta_2} \log T \leq T^{-1}/(\theta_2 - 1)$, the result follows. □

Remark 4. *If a sequence $(y_t)_{t \in \mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$ satisfies*

$$(\forall x \in \operatorname{dom} f)(\forall t \in \mathbb{N}) \quad \|y_{t+1} - x\|^2 \leq \|y_t - x\|^2 - \eta_t (f(y_t) - f(x)) + \xi_t,$$

under the same assumptions of Corollary 2, it is possible to derive an inequality analogous to (12)*.*

3 Applications in Convex Optimization

In this section, we apply previous results to some convex optimization algorithms, including forward-backward splitting, projected subgradient, incremental proximal subgradient, and Douglas-Rachford splitting method. Convergence rates for the objective function values are obtained by using Theorem 2. The key observation is that the sequences generated by these algorithms are modified Fejér monotone.

Throughout this section, we assume that $f : \mathcal{H} \to]-\infty, \infty]$ is a proper, lower semicontinuous convex function. Recall that the subdifferential of f at $x \in \mathcal{H}$ is

$$\partial f(x) = \{u \in \mathcal{H} : (\forall y \in \mathcal{H}) \ \ f(x) + \langle u, y - x \rangle \leq f(y)\}. \quad (17)$$

The elements of the subdifferential of f at x are called subgradients of f at x. More generally, for $\epsilon \in]0, +\infty[$, the ϵ-subdifferential of f at x is the set $\partial_\epsilon f(x)$ defined by

$$\partial_\epsilon f(x) = \{u \in \mathcal{H} : (\forall y \in \mathcal{H}) \ \ f(x) + \langle u, y - x \rangle - \epsilon \leq f(y)\}. \quad (18)$$

The proximity operator of f [30] is

$$\operatorname{prox}_f(x) = \operatorname*{argmin}_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2} \|y - x\|^2 \right\}. \quad (19)$$

3.1 Forward-Backward Splitting

In this subsection, we consider a forward-backward splitting algorithm for solving Problem (1), with objective function

$$f = \ell + r \tag{20}$$

where $r\colon \mathcal{H} \to \left]-\infty, \infty\right]$ and $\ell\colon \mathcal{H} \to \mathbb{R}$ are proper, lower semicontinuous, and convex. Since ℓ is real-valued, we have $\operatorname{dom} \partial\ell = \mathcal{H}$ [3, Proposition 16.14].

Algorithm 1. *Given $x_1 \in \mathcal{H}$, a sequence of stepsizes $(\alpha_t)_{t\in\mathbb{N}} \subset \left]0, +\infty\right[$, and a sequence $(\epsilon_t)_{t\in\mathbb{N}} \subset \left[0, +\infty\right[$ set, for every $t \in \mathbb{N}$,*

$$x_{t+1} = \operatorname{prox}_{\alpha_t r}(x_t - \alpha_t g_t) \tag{21}$$

with $g_t \in \partial_{\epsilon_t}\ell(x_t)$.

The forward-backward splitting algorithm has been well studied [43, 10, 12, 9] and a review of this algorithm can be found in [14] under the assumption that ℓ is differentiable with a Lipschitz continuous gradient. Convergence is proved using arguments based on Fejér monotonicity of the generated sequences [13]. Under the assumption that ℓ is a differentiable function with Lipschitz continuous gradient, the algorithm exhibits a sublinear convergence rate $O(T^{-1})$ on the objective f [4]. If ℓ is not smooth, the algorithm has been studied first in [34], and has a convergence rate $O(T^{-1/2})$, considering the best point selection rule [42]. A comprehensive study of proximal subgradient methods can be found in [5]. The use of ϵ-subgradients in a scaled version of algorithm (21) has been investigated in [6], in the special case where r is the indicator function of a convex and closed set, without focusing on convergence rates. Our objective here is to provide a convergence rate for the algorithm considering the last iterate, which shares the same rate (up-to logarithmic factors) and to allow the use of ϵ-subgradients, instead of subgradients. Before stating our main results, we introduce the following novel lemma for the forward-backward splitting for a (possibly) non-smooth ℓ. It recovers previous result (e.g. [4]) when ℓ is smooth.

Lemma 1. *Let $(x_t)_{t\in\mathbb{N}^*}$ be the sequence generated by Algorithm 1. Then for all $t \in \mathbb{N}^*$, there holds*

$$\begin{aligned}2\alpha_t[f(x_{t+1}) - f(x)] \leq \|x_t - x\|^2 - \|x_{t+1} - x\|^2 - \|x_t - x_{t+1}\|^2 \\ + 2\alpha_t[\langle x_{t+1} - x_t, g_{t+1} - g_t\rangle + \epsilon_{t+1} + \epsilon_t]. \quad (22)\end{aligned}$$

Proof. Let $t \in \mathbb{N}^*$. By Fermat's rule (see e.g. [3, Theorem 16.2]),

$$0 \in x_{t+1} - x_t + \alpha_t g_t + \alpha_t \partial r(x_{t+1}).$$

Thus, there exists $q_{t+1} \in \partial r(x_{t+1})$, such that x_{t+1} in (21) can be written as

$$x_{t+1} = x_t - \alpha_t g_t - \alpha_t q_{t+1}. \tag{23}$$

Let $x \in \text{dom} f$. The convexity of r implies

$$r(x_{t+1}) - r(x) \leq \langle x_{t+1} - x, q_{t+1}\rangle.$$

Multiplying both sides by $2\alpha_t$, and combining with (23), we get

$$\begin{aligned}2\alpha_t[r(x_{t+1}) - r(x)] &\leq 2\alpha_t\langle x_{t+1} - x, q_{t+1}\rangle \\ &= 2\langle x_{t+1} - x, x_t - x_{t+1} - \alpha_t g_t\rangle \\ &= 2\langle x_{t+1} - x, x_t - x_{t+1}\rangle + 2\alpha_t\langle x - x_{t+1}, g_t\rangle.\end{aligned}$$

A direct computation yields

$$\begin{aligned}2\langle x_{t+1} - x, x_t - x_{t+1}\rangle &= 2\langle x_{t+1} - x, x_t - x\rangle - 2\|x_{t+1} - x\|^2 \\ &= \|x_t - x\|^2 - \|x_{t+1} - x\|^2 - \|x_t - x_{t+1}\|^2. \quad (24)\end{aligned}$$

Therefore,

$$\begin{aligned}&2\alpha_t[r(x_{t+1}) - r(x)] \\ \leq\ &\|x_t - x\|^2 - \|x_{t+1} - x\|^2 - \|x_t - x_{t+1}\|^2 + 2\alpha_t\langle x - x_{t+1}, g_t\rangle.\end{aligned}$$

Moreover, by (18), we have

$$\langle x - x_t, g_t\rangle \leq \ell(x) - \ell(x_t) + \epsilon_t,$$

and

$$\langle x_t - x_{t+1}, g_{t+1}\rangle \leq \ell(x_t) - \ell(x_{t+1}) + \epsilon_{t+1},$$

and thus

$$\begin{aligned}\langle x - x_{t+1}, g_t\rangle &= \langle x - x_t, g_t\rangle + \langle x_t - x_{t+1}, g_{t+1}\rangle + \langle x_t - x_{t+1}, g_t - g_{t+1}\rangle\\ &\leq \ell(x) - \ell(x_t) + \epsilon_t + \ell(x_t) - \ell(x_{t+1}) + \epsilon_{t+1} + \langle x_t - x_{t+1}, g_t - g_{t+1}\rangle\\ &= \ell(x) - \ell(x_{t+1}) + \langle x_{t+1} - x_t, g_{t+1} - g_t\rangle + \epsilon_{t+1} + \epsilon_t.\end{aligned}$$

Consequently, we get

$$\begin{aligned}2\alpha_t[r(x_{t+1}) - r(x)] \leq\ & \|x_t - x\|^2 - \|x_{t+1} - x\|^2 - \|x_t - x_{t+1}\|^2\\ &+2\alpha_t[\ell(x) - \ell(x_{t+1}) + \langle x_{t+1} - x_t, g_{t+1} - g_t\rangle + \epsilon_{t+1} + \epsilon_t].\end{aligned}$$

Rearranging terms and recalling that $f = \ell + r$, the desired result thus follows.
□

Theorem 3. *Let $\alpha \in]0, +\infty[$, let $\theta \in [0, 1[$, and let, for every $t \in \mathbb{N}^*$, $\alpha_t = \alpha t^{-\theta}$. Let $\epsilon \in]0, +\infty[$, $(\epsilon_t)_{t\in\mathbb{N}^*} \subset [0, +\infty]$, and assume that $\epsilon_t \leq \epsilon\alpha_t$. Let $(x_t)_{t\in\mathbb{N}^*}$ be the sequence generated by Algorithm 1. Let $T \in \mathbb{N}$, and assume that there exists $B \in]0, +\infty[$ such that*

$$(\forall T \in \{1, \ldots, T\}) \quad \|g_t\| \leq B, \tag{25}$$

Then, there exists $c \in [0, +\infty[$ such that

$$f(x_{T+1}) - f_* \leq \begin{cases} \dfrac{d(x_1, \mathcal{X})^2}{2\alpha} T^{\theta-1} + 2\alpha c(B^2 + \epsilon)T^{-\theta}\log T & \textit{if } \theta \leq 1/2\\ \left[\dfrac{d(x_1, \mathcal{X})^2}{2\alpha} + 2\alpha c(B^2 + \epsilon)\right] T^{\theta-1} & \textit{otherwise.}\end{cases}$$

Proof. Let $t \in \mathbb{N}^*$. By (22) and Cauchy-Schwartz inequality

$$\begin{aligned}&\|x_{t+1} - x\|^2 - \|x_t - x\|^2\\ \leq& -\|x_t - x_{t+1}\|^2 + 2\alpha_t[\langle x_{t+1} - x_t, g_{t+1} - g_t\rangle + \epsilon_{t+1} + \epsilon_t] - 2\alpha_t[f(x_{t+1}) - f(x)]\\ \leq& -\|x_t - x_{t+1}\|^2 + 2\alpha_t[\|x_{t+1} - x_t\|\|g_{t+1} - g_t\| + \epsilon_{t+1} + \epsilon_t] - 2\alpha_t[f(x_{t+1}) - f(x)]\\ \leq& \alpha_t^2\|g_{t+1} - g_t\|^2 + 2\alpha_t[\epsilon_{t+1} + \epsilon_t] - 2\alpha_t[f(x_{t+1}) - f(x)].\end{aligned}$$

Using the assumptions $\|g_t\| \leq B$ and $\epsilon_t \leq \epsilon\alpha_t$,

$$\begin{aligned}
&\|x_{t+1} - x\|^2 - \|x_t - x\|^2 \\
\leq& 4B^2\alpha_t^2 + 2\epsilon\alpha_t[\alpha_t + \alpha_{t+1}] - 2\alpha_t[f(x_{t+1}) - f(x)] \\
\leq& 4(B^2 + \epsilon)\alpha_t^2 - 2\alpha_t[f(x_{t+1}) - f(x)]. \qquad (26)
\end{aligned}$$

Thus, $(x_t)_{t\in\mathbb{N}^*}$ is a modified Fejér sequence with respect to the objective function f and $\big((2\alpha_t, 4(B^2+\epsilon)\alpha_t^2)\big)_{t\in\mathbb{N}^*}$. The statement follows from Theorem 2, applied with $\theta_1 = \theta$, $\theta_2 = 2\theta$, $\eta = 2\alpha$ and $\xi = 4(B^2+\epsilon)\alpha^2$. □

The following remark collects some comments on the previous result.

Remark 5.

(i) Theorem 3 suggests the optimal choice for the stepsize decay rate $\theta = 1/2$. *In such a way, we get a convergence rate* $O(T^{-1/2}\log T)$ *for forward-backward algorithm applied to a sum of nonsmooth functions with nonsummable diminishing stepsizes, considering the last iterate. As usual in this setting, no convergence is obtained using a fixed stepsize.*

(ii) In Theorem 3, the assumption on bounded approximate subgradients, which implies Lipschitz continuity of ℓ, *is satisfied for several practical optimization problems. For example, when* r *is the indicator function of a closed, bounded, and convex set* $D \subset \mathbb{R}^N$, *it follows that* $(x_t)_{t\in\mathbb{N}}$ *is bounded. If* ℓ *is bounded on bounded sets, this implies that* ℓ *is Lipschitz continuous on bounded sets [44, Corollary 2.2.12] and thus that* $(g_t)_{t\in\mathbb{N}}$ *is bounded as well [35, Proposition 1.11]. Similar results to those proved here may be obtained by imposing a less restrictive growth condition on* ∂f, *using a similar approach to that in [28] to bound the sequence of subgradients.*

(iii) An inequality similar to (26) *has been obtained in [5, Lemma 2.1]. Theorem 3 improves [5, Corollary 2.4] in two aspects. First, the assumption (25) is weaker than the assumption* $\|g_t + u_t\| \leq B$ *for some* $u_t \in \partial r(x_t)$,

in [5]. Second, [5] shows convergence rate only for the best point, i.e, the one with smallest function value:

$$(\forall T \in \mathbb{N}^*) \qquad b_T = \operatorname*{argmin}_{1\leq t\leq T} f(x_t). \tag{27}$$

whereas our result holds for any iterate.

If the function ℓ in (20) is differentiable, with a Lipschitz differentiable gradient, we recover the following well-known convergence result. We include the proof for completeness.

Proposition 1. *[4, Theorem 3.1] Let $\beta \in [0, +\infty[$ and assume that $\nabla\ell$ is β-Lipschitz continuous. Consider Algorithm 1 with $\epsilon_t = 0$ and $(\alpha_t)_{t\in\mathbb{N}}$ non-increasing, with $\alpha_t \in]0, 1/\beta[$ for all $t \in \mathbb{N}^*$. Then, for every $T \in \mathbb{N}^*$,*

$$f(x_{T+1}) - f_* \leq \frac{d(x_1, \mathcal{X})^2}{\alpha_T T} \tag{28}$$

Proof. Since $\nabla\ell$ is β-Lipschitz continuous, it follows from the descent lemma [3, Theorem 18.15] and the definition of the forward-backward algorithm that

$$\ell(x_{t+1}) - \ell(x_t) \leq \langle \nabla\ell(x_t), x_{t+1} - x_t\rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2. \tag{29}$$

Since $(x_t - x_{t+1})/\alpha_t - \nabla f(x_t) \in \partial r(x_{t+1})$, it follows from convexity of ℓ and r, that

$$(\forall y \in \mathcal{H}) \quad \ell(x_t) - \ell(y) \leq \langle \nabla\ell(x_t), x_t - y\rangle \tag{30}$$

$$(\forall y \in \mathcal{H}) \quad r(x_{t+1}) - r(y) \leq \Big\langle \frac{x_{t+1} - x_t}{\alpha_t} + \nabla\ell(x_t), y - x_{t+1}\Big\rangle. \tag{31}$$

Summing up (29),(30),(31) we derive

$$f(x_{t+1}) - f(y) \leq \Big\langle \frac{x_t - x_{t+1}}{\alpha_t}, x_{t+1} - y\Big\rangle + \frac{\beta}{2}\|x_t - x_{t+1}\|^2.$$

Therefore

$$2\alpha_t(f(x_{t+1}) - f(y)) \leq \|x_t - y\|^2 - \|x_{t+1} - y\|^2 + (\beta\alpha_t - 1)\,\|x_t - x_{t+1}\|^2,$$

which implies, since $\alpha_t \leq 1/\beta$, that $(x_t)_{t\in\mathbb{N}}$ is modified Fejér monotone with respect to f and the sequence $\big((2\alpha_t, 0)\big)$. The statement follows from Corollary 1.

Remark 6. *For the forward-backward algorithm, it has been proved in [17] that* $f(x_{T+1}) - f_* = o(1/T)$ *also for* $\alpha_t \in\,]0, 2/\beta[$*, even in the presence of errors, and with relaxation. The concept of modified Fejér monotonicity allows to derive nonasymptotic bounds on the sequence of iterates, but only if* $\alpha_t \in\,]0, 1/\beta[$ *for every* $t \in \mathbb{N}^*$*.*

□

3.2 Projected Approximate Subgradient Method

Let D be a convex and closed subset of $\mathcal{H}$, and let ι_D be the indicator function of D. In this subsection, we consider Problem (1) with objective function given by

$$f = \ell + \iota_D \tag{32}$$

where $\ell\colon \mathcal{H} \to \mathbb{R}$ is lower semicontinuous and convex. It is clear that (32) is a special case of (20) corresponding to a given choice of r. The forward-backward algorithm in this case reduces to the following projected subgradient method (see e.g. [8, 26, 36, 41] and references therein), which allows to use ϵ-subgradients, see [2, 11].

Algorithm 2. *Given* $x_1 \in \mathcal{H}$*, a sequence of stepsizes* $(\alpha_t)_{t\in\mathbb{N}} \subset\,]0, +\infty[$*, and a sequence* $(\epsilon_t)_{t\in\mathbb{N}} \subset [0, +\infty[$*, set, for every* $t \in \mathbb{N}$*,*

$$x_{t+1} = P_D(x_t - \alpha_t g_t) \tag{33}$$

with $g_t \in \partial_{\epsilon_t}\ell(x_t)$*.*

The algorithm has been studied using different rules for choosing the stepsizes. Here, as a corollary of Theorem 3, we derive the convergence rate for the objective function values, for a nonsummable diminishing stepsize.

Theorem 4. *For some* $\alpha_1 > 0$*,* $\epsilon \geq 0$ *and* $\theta \in [0, 1)$*, let* $\alpha_t = \eta t^{-\theta}$ *and* $\epsilon_t \leq \epsilon\alpha_t$ *for all* $t \in \mathbb{N}^*$*. Let* $(x_t)_{t\in\mathbb{N}}$ *be a sequence generated by Algorithm 2. Assume that for all* $t \in \mathbb{N}^*$*,* $\|g_t\| \leq B$*. Then, there exists* $c \in\,]0, +\infty[$ *such that, for every* $T \in \mathbb{N}^*$

$$f(x_{T+1}) - f^* \leq \begin{cases} \dfrac{d(x_1, \mathcal{X})^2}{2\alpha_1} T^{\theta-1} + \alpha_1 c(B^2 + 2\epsilon)T^{-\theta}\log T & \text{if } \theta \leq 1/2 \\ \left[\dfrac{d(x_1, \mathcal{X})^2}{2\alpha_1} + \alpha_1 c(B^2 + 2\epsilon)\right] T^{\theta-1} & \text{otherwise} \end{cases}$$

Choosing $\theta = 1/2$, we get a convergence rate of order $O(T^{-1/2} \log T)$ for projected approximate subgradient method with nonsummable diminishing stepsizes, which is optimal up to a log factor without any further assumption on f [19, 33]. Since the subgradient method is not a descent method, a common approach keeps track of the best point found so far, see (27). The projected subgradient method with diminishing stepsizes of the form $(\alpha t^{-\theta})_{t \in \mathbb{N}}$, with $\theta \in \,]0, 1]$, satisfies $f(x_{T+1}) - f_* = O(T^{-1/2})$. Our result shows that considering the last iterate for projected approximate subgradient method essentially leads to the same convergence rate, up to a logarithmic factor, as the one corresponding to the best iterate, even if the function value may not decrease at each iteration. To the best of our knowledge, our result is the first of this kind, without any assumption on strong convexity of f, or on a conditioning number with respect to subgradients (as in [23] using stepsizes $\{\gamma_t/\|g_t\|\}_t$). Note that, using nonsummable diminishing stepsizes, convergence rate $O(T^{-1/2})$ was shown, but only for a subsequence of $(x_t)_{t \in \mathbb{N}^*}$ [2]. Finally, let us mention that using properties of quasi-Fejér sequences, convergence properties were proved in [11].

3.3 Incremental Subgradient Proximal Algorithm

In this subsection, we consider an incremental subgradient proximal algorithm [7, 31] for solving (1), with objective function f given by, for some $m \in \mathbb{N}^*$,

$$\sum_{i=1}^{m} (\ell_i + r_i),$$

where for each i, $\ell_i : \mathcal{H} \to \mathbb{R}$ and $r_i : \mathcal{H} \to \,]-\infty, +\infty]$ are convex, proper, and lower semicontinuous. The algorithm is similar to the proximal subgradient method, the main difference being that at each iteration, x_t is updated incrementally, through a sequence of m steps.

Algorithm 3. *Let $t \in \mathbb{N}^*$. Given $x_t \in \mathcal{H}$, an iteration of the incremental proximal subgradient algorithm generates x_{t+1} according to the recursion,*

$$x_{t+1} = \psi_t^m, \tag{34}$$

where ψ_t^m is obtained at the end of a cycle, namely as the last step of the recursion

$$\psi_t^0 = x_t, \qquad \psi_t^i = \operatorname{prox}_{\alpha_t r_i}(\psi_t^{i-1} - \alpha_t g_t^i), \qquad \forall g_t^i \in \partial \ell_i(\psi_t^{i-1}), \quad i = 1, \cdots, m \tag{35}$$

for a suitable sequence of stepsizes $\{\alpha_t\}_{t\in\mathbb{N}^*} \subset]0, +\infty[$.

Several versions of incremental subgradient proximal algorithms have been studied in [7], where convergence results for various stepsizes rules and both for stochastic of cyclic selection of the components are given. Concerning the function values, the results are stated in terms of the best iterate, i.e., (27). See also [32] for the study of the special case of incremental subgradient methods under different stepsizes rules. The paper [24] provides convergence results using approximate subgradients instead of gradients.

In this section, we derive a sublinear convergence rate for the incremental subgradient proximal algorithm in a straightforward way, relying on the properties of modified Fejér sequences assuming a boundedness assumption on the subdifferentials, already used in [32].

Theorem 5. *Let $\alpha \in]0, +\infty[$, let $\theta \in [0, 1[$, and let, for every $t \in \mathbb{N}^*$, $\alpha_t = \alpha t^{-\theta}$. Let $(x_t)_{t\in\mathbb{N}^*}$ be the sequence generated by Algorithm 3. Let $B \in]0, +\infty[$ be such that*

$$(\forall t \in \mathbb{N}^*)(\forall g \in \partial \ell_i(x_t) \cup \partial r_i(x_t)) \qquad \|g\| \le B.$$

Then, there exists $c \in]0, +\infty[$ such that, for every $T \in \mathbb{N}^*$,

$$f(x_T) - f_* \le \begin{cases} \dfrac{d(x_1, \mathcal{X})^2}{2\alpha} T^{\theta-1} + \dfrac{c\alpha(4m+5)mB^2}{2} T^{-\theta} \log T & \text{if } \theta \le 1/2 \\ \left[\dfrac{d(x_1, \mathcal{X})^2}{2\alpha} + \dfrac{c\alpha(4m+5)mB^2}{2}\right] T^{\theta-1} & \text{otherwise.} \end{cases} \tag{36}$$

Proof. It was shown in [7, Proposition 3 (Equation 27)] that,

$$\|x_{t+1} - x\|^2 \le \|x_t - x\|^2 - 2\alpha_t[f(x_t) - f(x)] + \alpha_t^2 (4m+5) mB^2.$$

Thus, $(x_t)_{t\in\mathbb{N}^*}$ is a modified Fejér sequence with respect to the objective function f, and $((2\alpha_t, \alpha_t^2 (4m+5) mB^2))_{t\in\mathbb{N}^*}$. The proof is concluded by applying Remark 4 with $\theta_1 = \theta, \theta_2 = 2\theta$, $\eta = 2\alpha$ and $\xi = \alpha^2\,(4m+5)\,mB^2$. □

Remark 7.

(i) The choice $\theta = 1/2$ in Theorem 5, yields a convergence rate of order $O(T^{-1/2}\log T)$ for the objective function values.

(ii) In [7, Proposition 5] a bound similar to (36) is derived for the best iterate (27) with a fixed stepsize. In contrast to this previous result, our result holds for any last iterate, considering both the fixed and diminishing stepsize setting. Note that, neither [7, Proposition 5] nor Theorem 5 imply convergence for the fixed stepsize.

As in Theorem 5, we can derive convergence rates for the projected incremental subgradient method. Analogously to what we have done for the forward-backward algorithm in Section 3.1, Theorem 5 can be extended to analyze convergence of the approximate and incremental subgradient method in [24].

3.4 Douglas-Rachford splitting method

In this subsection, we consider Douglas-Rachford splitting algorithm for solving (1). Given $\ell\colon \mathcal{H} \to \,]-\infty,+\infty]$ and $r\colon \mathcal{H} \to \,]-\infty,+\infty]$ convex and lower semicontinuous functions, we suppose that $f = \ell + r$ in (1).

Algorithm 4. *Let $(\alpha_t)_{t\in\mathbb{N}^*} \in \,]0,+\infty[^{\mathbb{N}}$. Let $t \in \mathbb{N}^*$. Given $x_t \in \mathcal{H}$, an iteration of Douglas-Rachford algorithm generates x_{t+1} according to*

$$\left\{\begin{array}{l} y_{t+1} = \operatorname{prox}_{\alpha_t \ell}(x_t) \\ z_{t+1} = \operatorname{prox}_{\alpha_t r}(2y_{t+1} - x_t), \\ x_{t+1} = x_t + z_{t+1} - y_{t+1}. \end{array}\right. \tag{37}$$

The algorithm has been introduced in [21] to solve matrix equations. Then it has been extended for solving the minimization problem of the sum of two convex functions [27], and then to monotone inclusions involving the sum of two nonlinear operators [29]. A review of this algorithm can be found in [14]. The convergence of the iterates is established using the theory of

Fejér sequences [13]. Our objective here is to establish a new result, namely a convergence rate for the objective function values.

Theorem 6. *Let $\alpha \in \,]0, +\infty[$, and let $\theta \in [0, 1[$. For every $t \in \mathbb{N}^*$, let $\alpha_t = \alpha t^{-\theta}$. Let $\big((y_t, x_t, z_t)\big)_{t\in\mathbb{N}^*}$ be the sequences generated by Algorithm 4. Assume that there exists $B \in \,]0, +\infty[$ such that, for every $t \in \mathbb{N}^*$:*

$$(\forall v \in \partial\ell(y_t))(\exists u_t \in \partial\ell(x_t))(\exists s_t \in \partial r(x_t)) \qquad \|v\| \leq B, \quad \|u_t\| \leq B \quad \textit{and} \quad \|s_t\| \leq B. \tag{38}$$

Then, there exists $c \in \,]0, +\infty[$, such that, for every $T \in \mathbb{N}^*$,

$$f(x_{T+1}) - f_* \leq \begin{cases} \dfrac{d(x_1, \mathcal{X})^2}{2\alpha} T^{\theta-1} + \dfrac{5c\alpha B^2}{2} T^{-\theta} \log T & \textit{if } \theta \leq 1/2 \\ \left[\dfrac{d(x_1, \mathcal{X})^2}{2\alpha} + \dfrac{5c\alpha B^2}{2}\right] T^{\theta-1} & \textit{otherwise.} \end{cases}$$

Proof. Let $t \in \mathbb{N}^*$, set $v_{t+1} = (x_t - y_{t+1})/\alpha_t$ and $w_{t+1} = (2y_{t+1} - x_t - z_{t+1})/\alpha_t$. By Fermat's rule,

$$v_{t+1} \in \partial\ell(y_{t+1}) \text{ and } w_{t+1} \in \partial r(z_{t+1}). \tag{39}$$

We can rewrite (37) as

$$\begin{cases} y_{t+1} = x_t - \alpha_t v_{t+1}, \\ z_{t+1} = (2y_{t+1} - x_t) - \alpha_t w_{t+1}, \\ x_{t+1} = x_t + z_{t+1} - y_{t+1}, \end{cases} \tag{40}$$

By (40), we have for any $x \in \operatorname{dom} f$,

$$\ell(y_{t+1}) - \ell(x) \leq \langle y_{t+1} - x, v_{t+1} \rangle.$$

Multiplying both sides by $2\alpha_t$,

$$2\alpha_t[\ell(y_{t+1}) - \ell(x)] \leq 2\alpha_t\langle y_{t+1} - x, v_{t+1}\rangle = 2\langle y_{t+1} - x, x_t - y_{t+1}\rangle.$$

Similarly, we have

$$2\alpha_t[r(z_{t+1}) - r(x)] \leq 2\alpha_t\langle z_{t+1} - x, w_{t+1}\rangle = 2\langle z_{t+1} - x, 2y_{t+1} - x_t - z_{t+1}\rangle.$$

Combining the above two estimates, we get

$$\begin{array}{rl} & 2\alpha_t[\ell(y_{t+1}) + r(z_{t+1}) - \ell(x) - r(x)] \\ \leq & 2\langle y_{t+1} - x, x_t - y_{t+1}\rangle + 2\langle z_{t+1} - x, 2y_{t+1} - x_t - z_{t+1}\rangle. \end{array}$$

The third equality in (40) implies that that $z_{t+1} = x_{t+1} - x_t + y_{t+1}$, and thus

$$\begin{array}{rl} & 2\alpha_t[\ell(y_{t+1}) + r(z_{t+1}) - \ell(x) - r(x)] \\ \leq & 2\langle y_{t+1} - x, x_t - y_{t+1}\rangle + 2\langle x_{t+1} - x_t + y_{t+1} - x, 2y_{t+1} - x_{t+1} - y_{t+1}\rangle \\ = & 2\langle x_t - x_{t+1}, x_{t+1} - x\rangle. \\ = & \|x_t - x\|^2 - \|x_{t+1} - x\|^2 - \|x_t - x_{t+1}\|^2. \end{array}$$

Adding $2\alpha_t[\ell(x_{t+1}) + r(x_{t+1}) - \ell(y_{t+1}) - r(z_{t+1})]$ to both sides, and recalling that $f = \ell + r$,

$$\begin{aligned} 2\alpha_t[f(x_{t+1}) - f(x)] \leq \|x_t - x\|^2 - \|x_{t+1} - x\|^2 - \|x_t - x_{t+1}\|^2 & \\ + 2\alpha_t[l(x_{t+1}) + r(x_{t+1}) - l(y_{t+1}) - r(z_{t+1})]. & \quad (41) \end{aligned}$$

Let $u_{t+1} \in \partial\ell(x_{t+1})$ and $s_{t+1} \in \partial r(x_{t+1})$ such that $\|u_{t+1}\| \leq B$ and $\|s_{t+1}\| \leq B$. Convexity of ℓ and r, and (40) yield,

$$\begin{array}{rcl} \ell(x_{t+1}) - \ell(y_{t+1}) & \leq & \langle x_{t+1} - y_{t+1}, u_{t+1}\rangle \\ & = & \langle x_{t+1} - x_t, u_{t+1}\rangle + \langle x_t - y_{t+1}, u_{t+1}\rangle \\ & = & \langle x_{t+1} - x_t, u_{t+1}\rangle + \alpha_t\langle v_{t+1}, u_{t+1}\rangle \\ & \leq & \|x_{t+1} - x_t\|\|u\| + \alpha_t\|v_{t+1}\|\|u_{t+1}\| \\ & \leq & \|x_{t+1} - x_t\|B + \alpha_t B^2 \\ & \leq & \|x_{t+1} - x_t\|^2/(2\alpha_t) + B^2\alpha_t/2 + \alpha_t B^2, \end{array}$$

and

$$\begin{array}{rcl} r(x_{t+1}) - r(z_{t+1}) & \leq & \langle x_{t+1} - z_{t+1}, s_{t+1}\rangle \\ & = & \alpha_t\langle v_{t+1}, s_{t+1}\rangle \leq \alpha_t\|v_{t+1}\|\|s_{t+1}\| \leq \alpha_t B^2. \end{array}$$

Introducing the last two estimates into (41), and by a direct calculation,

$$2\alpha_t[f(x_{t+1}) - f(x)] \leq \|x_t - x\|^2 - \|x_{t+1} - x\|^2 + 5B^2\alpha_t^2.$$

Thus, $(x_t)_{t\in\mathbb{N}^*}$ is a modified quasi-Fejér sequence with respect to the objective function f and $\big((2\alpha_t, 5\alpha_t^2 B^2)\big)_{t\in\mathbb{N}^*}$. The statement follows from Theorem 2 with $\theta_1 = \theta$ and $\theta_2 = 2\theta$. □

Remark 8.

(i) Condition (38) *is verified if ℓ and r are Lipschitz continuous on $\mathcal{H}$. In this case the subdifferential is nonempty at every point and uniformly bounded, see [35, Proposition 1.11].*

(ii) Choosing $\theta = 1/2$, we get a convergence rate $O(T^{-1/2}\log T)$ for the algorithm with nonsummable diminishing stepsizes. Convergence does not follow using a fixed stepsize.

(iii) In the case where ℓ is the indicator function of a linear subspace of $\mathcal{H}$ (thus taking also the value $+\infty$) and r is Lipschitz continuous, nonergodic convergence rates for the objective function values corresponding to the Douglas-Rachford iteration can be derived by [20, Corollary 3.5].

4 Concluding remarks

We studied a modified notion of Fejér monotonicity, providing various convergence results for different optimization algorithms. Possible generalization and extension of the proposed notion can be considered, for instance including the stochastic and the variable metric settings.

References

[1] Attouch, H., Bolte, J., and Svaiter, B., Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, Math. Program. Ser. A 137, 91-129, 2011.

[2] Alber, Y. I., Iusem, A. N., and Solodov, M. V.: On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. Math. Program. 81, 23-35 (1998).

[3] Bauschke, H. H., and Combettes, P. L.: Convex Analysis and Monotone Operator Theory in Hilbert spaces, Springer, New York, 2011.

[4] Beck, A., and Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sciences 2, 183-202 (2009).

[5] Bello-Cruz, J.-Y., On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions, Set-Valued Var. Anal. 25, 245-263 (2017).

[6] Bonettini, S., Benfenati, A., and Ruggiero, V.: Scaling techniques for ϵ-subgradient methods, SIAM J. Optim. 26, 891-921 (2016).

[7] Bertsekas, D. P.: Incremental proximal methods for large scale convex optimization. Math. Program., Ser. B 129, 163-195 (2009).

[8] Boyd, S., Xiao, L., and Mutapcic, A.: Subgradient methods. https://web.stanford.edu/class/ee392o/subgrad_method.pdf. Accessed 14 October 2015.

[9] Bredies, K. and Lorenz, D. A.: Linear convergence of iterative soft-thresholding. J. Fourier Anal. Appl. 14, 813-837 (2008).

[10] Chen, G. H., and Rockafellar, R. T.: Convergence rates in forward–backward splitting. SIAM J. Optim. 7, 421-444 (1997).

[11] Combettes, P. L.: Quasi-Fejérian analysis of some optimization algorithms. Stud. Comput. Math. 8, 115-152 (2001).

[12] Combettes, P.L. and Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. 4, 1168-1200 (2005).

[13] Combettes, P. L.: Fejér monotonicity in convex optimization. In: C. A. Floudas and P. M. Pardalos (eds.) Encyclopedia of Optimization (pp. 1016-1024). Springer, New York (2009).

[14] Combettes, P. L., and Pesquet, J. C.: Proximal splitting methods in signal processing. In: Fixed-point algorithms for inverse problems in science and engineering (pp. 185-212). Springer, New York, 2011.

[15] Combettes, P.L., and Pesquet, J.C.: Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping, SIAM J. Optim. 25 1221-1248 (2015).

[16] Combettes, P.L., and Pesquet, J.C.: Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping II:Mean-square and linear convergence, https://arxiv.org/abs/1704.08083, (2017).

[17] Combettes, P.L., Salzo, S., and Villa, S.: Consistent learning by composite proximal thresholding, Mathematical Programming, published online 2017-03-25.

[18] Combettes, P. L., and Vũ, B.C.: Variable metric quasi Fejér monotonicity. Nonlinear Anal. 78, 17-31 (2013).

[19] Darzentas, J.: Problem complexity and method efficiency in optimization. J. Oper. Res. Soc., 35, 455-455 (1984).

[20] Davis, D.: Convergence rate analysis of the forward-Douglas-Rachford splitting scheme, SIAM J. Optim. 25, 1760-1786 (2015).

[21] Douglas, J., and Rachford, H. H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Amer. Math. Soc. 82, 421-439 (1956).

[22] Ermol'ev, Yu. M. and Tuniev, A. D.: Random Fejér and quasi-Fejér sequences, Theory of Optimal Solutions – Akademiya Nauk Ukrainskoĭ SSR Kiev 2, 76–83 (1968) ; translated in: American Mathematical Society Selected Translations in Mathematical Statistics and Probability 13,143-148 (1973).

[23] Goffin, J. L.: On convergence rates of subgradient optimization methods. Math. Program. 13, 329-347 (1977).

[24] Kiwiel, K. C.: Convergence of approximate and incremental subgradient methods for convex optimization. SIAM J. Optim. 14, 807-840 (2004).

[25] Knopp, K. Infinite Sequences and Series, Dover Publications Inc., New York 1956.

[26] Larsson, T., Patriksson, M., Stromberg, A.-B.: On the convergence of conditional ϵ-subgradient methods for convex programs and convex-concave saddle-point problems, EJOR 151, 461-473, 2003.

[27] Lieutaud, J.: Approximation d'Opérateurs par des Méthodes de Décomposition. Thèse, Université de Paris (1969).

[28] Lin J., Rosasco L., and Zhou D. X.: Iterative regularization for learning with convex loss functions. Journal of Machine Learning Research 17, 1-38 (2016).

[29] Lions, P. L., and Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. 16, 964-979, (1979).

[30] Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien, C. R. Acad. Sci. Paris 255, 2897–2899, (1962).

[31] Nedic, A., and Bertsekas, D. P.: Incremental subgradient methods for nondifferentiable optimization. SIAM J. Optim. 12, 109-138, (2001).

[32] Nedic, A., and Bertsekas, D.: Convergence rate of incremental subgradient algorithms. In Stochastic optimization: algorithms and applications (pp. 223-264). Springer, New York 2001.

[33] Nesterov, Y.: Introductory Lectures on Convex Optimization. Springer, New York, 2004.

[34] Passty, G. B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. 72, 383–390 (1979).

[35] R. R. Phelps, Convex Functions, Monotone Operators and Differentiability, 2nd ed. Springer, New York, 1993.

[36] Polyak, B. T.: Introduction to Optimization. Optimization Software, New York 1987.

[37] Rosasco, L., Villa, S., and Vũ, B.C., A stochastic forward-backward splitting method for solving monotone inclusions in Hilbert spaces, J. Optim. Theory Appl. 169, 388-406 (2016).

[38] Rosasco, L., Villa, S., and Vũ, B.C., A stochastic inertial forward-backward splitting algorithm for multivariate monotone inclusions, Optim. 65, 1293-1314 (2016).

[39] Rosasco, L., Villa, S., and Vũ, B.C., A first-order stochastic primal-dual algorithm with correction step, Numer. Funct. Anal. Optim. 38, 602-626 (2017).

[40] Salzo, S.: The variable metric forward-backward splitting algorithm under mild differentiability assumptions arXiv1605.00952v1, 2016

[41] Shor, N. Z.: Minimization Methods for Non-Differentiable Functions. Springer, New York, 1979.

[42] Singer, Y., and Duchi, J. C.: Efficient learning using forward-backward splitting. In Advances in Neural Information Processing Systems (pp. 495-503) (2009).

[43] Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim. 29, 119-138 (1991).

[44] C. Zălinescu, Convex Analysis in General Vector Spaces. World Scientific, River Edge, 2002.