Research Article

Cristian Vega*, Cesare Molinari, Lorenzo Rosasco and Silvia Villa

Fast iterative regularization by reusing data

https://doi.org/10.1515/jiip-2023-0009
Received January 27, 2023; revised June 23, 2023; accepted August 7, 2023

Abstract: Discrete inverse problems correspond to solving a system of equations in a stable way with respect to noise in the data. A typical approach to select a meaningful solution is to introduce a regularizer. While for most applications the regularizer is convex, in many cases it is neither smooth nor strongly convex. In this paper, we propose and study two new iterative regularization methods, based on a primal-dual algorithm, to regularize inverse problems efficiently. Our analysis, in the noise free case, provides convergence rates for the Lagrangian and the feasibility gap. In the noisy case, it provides stability bounds and early stopping rules with theoretical guarantees. The main novelty of our work is the exploitation of some a priori knowledge about the solution set: we show that the linear equations determined by the data can be used more than once along the iterations. We discuss various approaches to reuse linear equations that are at the same time consistent with our assumptions and flexible in the implementation. Finally, we illustrate our theoretical findings with numerical simulations for robust sparse recovery and image reconstruction. We confirm the efficiency of the proposed regularization approaches, comparing the results with state-of-the-art methods.

Keywords: Primal-dual splitting algorithms, iterative regularization, early stopping, Landweber method, stability and convergence analysis

MSC 2020: 90C25, 65K10, 49M29

1 Introduction

Many applied problems require the estimation of a quantity of interest from noisy linear measurements, for instance compressed sensing [24, 25, 33, 65, 77], image processing [26, 28, 56, 58, 66, 67, 79], matrix completion [19, 22, 23, 49], and various problems in machine learning [6, 34, 53, 64, 70, 79, 80]. In all these problems, we are interested in finding stable solutions to a system of equations where the accessible data are corrupted by noise. This is classically achieved by regularization. The most popular procedure in the literature is Tikhonov (or variational) regularization [35], and consists in minimizing the sum of a data fidelity term plus a regularizer, which is explicitly added to the objective function and entails some a priori knowledge or some desired property on the solutions that we want to select. A trade-off parameter is then introduced to balance the fidelity term and the regularizer. In practice, this implies that the optimization problem has to be solved many times for different values of the parameter. Finally, a parameter - and the correspondent solution - is chosen accordingly to the performance with respect to some criterion, such as Morozov discrepancy principle in inverse problems [35] or cross-validation on left-out data in machine learning [38, 74].

A computationally efficient alternative to explicit regularization is iterative regularization, also known as implicit regularization [2, 13, 18, 35]. The minimization of the regularizer under the noisy data constraints is

***Corresponding author: Cristian Vega,** MaLGa Center, DIMA – Dipartimento di eccellenza 2023-27, Università di Genova, Via Dodecaneso 35, Genoa, Italy, e-mail: cristian.vega@edu.unige.it. https://orcid.org/0000-0001-7792-0137
Cesare Molinari, Silvia Villa, MaLGa Center, DIMA – Dipartimento di eccellenza 2023-27, Università di Genova, Via Dodecaneso 35, Genoa, Italy, e-mail: molinari@dima.unige.it, silvia.villa@unige.it. https://orcid.org/0000-0003-0864-5682, https://orcid.org/0000-0002-6232-5631
Lorenzo Rosasco, CBMM, Massachusets Institute of Technology, Cambridge, MA, USA; and Istituto Italiano di Tecnologia, Genoa, Italy, e-mail: lorenzo.rosasco@unige.it. https://orcid.org/0000-0003-3098-383X

considered and a numerical algorithm to solve the optimization problem is chosen and early stopped, to avoid convergence to the noisy solution. In this setting, it is known that the number of iterations plays the role of the regularization parameter [35]. As for Tikhonov regularization, the best performing iterate is chosen according to some a priori criterion and then considered as the regularized solution. Compared to Tikhonov regularization, this procedure is very efficient, since only one optimization problem is solved, and not even until convergence.

The main novelty of this work is the design and analysis of two new iterative regularization methods for convex regularizers, which are not necessarily smooth nor strongly convex. The new iterative regularization methods are based on primal-dual algorithms [29, 31, 78] combined with the idea of reusing the linear equations determined by the data at every iteration [16]. Primal-dual algorithms perform one minimization step on the primal variable followed by one on the dual and are well-suited for the large scale setting, as only matrix-vector multiplications and the calculation of a proximity operator are required. The idea of exploiting redundant information was presented in [16] and turned out to be very effective in practice. The first method that we propose is a primal-dual algorithm (PDA) with additional activations of the linear equations: we propose different variants, depending on the extra activation steps. For instance, we are able to exploit the data constraints more than once at every iteration via gradient descent, with a fixed or an adaptive step size. The second method is a dual-primal algorithm (DPA) where a subset containing the dual solutions is activated at each step. This subset is not affected by the noise in the data and is usually determined by a finite number of constraints.

These additional steps may seem artificial or inefficient. However, while maintaining an easy implementation, our methods achieve better numerical performances and considerable speed-ups with respect to the vanilla primal-dual algorithm. We extend to the noisy case the techniques studied in [16, 17] for the exact case. The assumptions on the noise are the classical ones in inverse problems, see, e.g., [18, 21, 47, 49], and the proposed results generalize the ones in [49], by including in the primal-dual procedure a diagonal preconditioning and an extra activation step. For the noisy case, we provide an early stopping criterion to recover a stable approximation of an ideal solution, in the same spirit of [5, 12, 18, 21, 47, 63, 80, 83]. The early stopping rule is derived from theoretical stability bounds and feasibility gap rates for both algorithms, obtaining implicit regularization properties similar to those stated in [47, 49]. Theoretical results are complemented by numerical experiments for robust sparse recovery and total variation, showing that state-of-the-art performances can be achieved with considerable computational speed-ups.

Related works. In this section, we briefly discuss the literature about variational and iterative regularization techniques. Tikhonov regularization has been introduced in [76]; see also [11, 35] and the references therein for an extensive treatment of the topic. The most famous iterative regularization method is the Landweber algorithm [35, 44], namely gradient descent on the least squares problem. Duality theory in optimization gives another interpretation which sheds light on the regularizing properties of this procedure. Indeed, consider the problem of minimizing the squared norm under linear constraints. Running gradient descent on its dual problem and mapping back to the primal variable, we obtain exactly the Landweber method. This provides another explanation of why the iterates of the Landweber algorithm converge to the minimal norm solution of the linear equation [47]. Stochastic gradient descent on the previous problem is the generalization of the Kaczmarz method [42, 46, 69, 75], which consists in applying cyclic or random projections onto single equations of the linear system. Accelerated and diagonal versions are also discussed in [35, 55] and [4, 43, 68], respectively. The regularization properties of other optimization algorithms for more general regularizers have also been studied. If strong convexity is assumed, mirror descent [8, 54] can also be interpreted as gradient descent on the dual problem, and its regularization properties (and those of its accelerated variant) have been studied in [47]. Diagonal approaches [3] with a regularization parameter that vanishes along the iterations have been studied in [37]; see [21] for an accelerated version. Another common approach relies on the linearized Bregman iteration [56, 79, 81, 82], which has found applications in compressed sensing [20, 57, 82] and image deblurring [20]. However, this method requires to solve non-trivial minimization problems at each iteration. For convex, but not strongly convex regularizers, the regularization properties of primal-dual algorithms have been investigated in [49]. Other optimization techniques are available to solve this kind of minimization problem (for instance, [51, 52] and [15, 48, 61], see also [40, 71, 72]), but no iterative regularization properties have been studied so far for these algorithms.

The rest of the paper is organized as follows. In Section 2, we introduce the notation jointly with its mathematical background. In Section 3, we present the main problem and propose five classes of algorithms to solve it numerically. In Section 4, we derive stability and feasibility gap bounds and related early stopping rules. In Section 5, we verify the performance of the algorithm on two numerical applications: robust sparse recovery problem and image reconstruction by total variation. Finally, we provide some conclusions.

2 Notation and background

First we recall some well known concepts and properties used in the paper.

Let X, Y be two finite-dimensional real vector spaces equipped with an inner product $\langle \cdot \mid \cdot \rangle$ and the induced norm $\|\cdot\|$. We denote the set of convex, lower semicontinuous, and proper functions on X by $\Gamma_0(X)$. The subdifferential of $F \in \Gamma_0(X)$ is the set-valued operator defined by

$$\partial F\colon X \to 2^X, \quad x \mapsto \{u \in X : F(x) + \langle y - x \mid u \rangle \leq F(y) \text{ for all } y \in X\}.$$

If the function F is Gâteaux differentiable at the point x, then $\partial F(x) = \{\nabla F(x)\}$ (see [7, Proposition 17.31 (i)]). In general, for $F \in \Gamma_0(X)$, it holds that $(\partial F)^{-1} = \partial F^*$ (see [7, Corollary 16.30]), where $F^* \in \Gamma_0(X)$ is the conjugate function of F, defined by

$$F^*(x) := \sup_{u \in X} \langle x \mid u \rangle - F(u).$$

For every self-adjoint positive definite matrix Σ, we define the proximity operator of F relative to the metric induced by $\|\cdot\|_\Sigma^2 := \langle \cdot \mid \Sigma \cdot \rangle$ as $\operatorname{prox}_F^\Sigma = (\mathrm{Id} + \Sigma \partial F)^{-1}$. If $\Sigma = \sigma \mathrm{Id}$ for some real number $\sigma > 0$, it is customary to write $\operatorname{prox}_{\sigma F}$ rather than $\operatorname{prox}_F^\Sigma$. The projection operator onto a nonempty closed convex set $C \subseteq X$ is denoted by P_C. If we define the indicator $\iota_C \in \Gamma_0(X)$ as the function that is 0 on C, and $+\infty$ otherwise, then $\operatorname{prox}_{\iota_C} = P_C$. Moreover, if C is a singleton, say $C = \{b\}$, we have that $\iota_{\{b\}}^*(u) = \langle u \mid b \rangle$ (see [7, Example 13.3 (i)]). The relative interior of C is

$$\operatorname{ri}(C) = \{x \in C \mid \mathbb{R}_{++}(C - x) = \operatorname{span}(C - x)\},$$

where

$$\mathbb{R}_{++}C = \{\lambda y \mid (\lambda > 0) \wedge (y \in C)\}$$

and $\operatorname{span}(C)$ is the smallest linear subspace of X containing C.

Given $\alpha \in \left]0, 1\right[$, an operator $T : X \to X$ is α-averaged non-expansive if

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{1 - \alpha}{\alpha} \|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 \quad \text{for all } x \in X \text{ and all } y \in X,$$

and it is quasi-non-expansive if

$$\|Tx - y\|^2 \leq \|x - y\|^2 \quad \text{for all } x \in X \text{ and all } y \in \operatorname{Fix} T,$$

where $\operatorname{Fix} T = \{x \in X \mid Tx = x\}$ is the set of fixed points of T. For further results on convex analysis and operator theory, the reader is referred to [7].

The operator norm of a real matrix $A \in \mathbb{R}^{d \times p}$ is denoted by $\|A\|$, and the adjoint of A is A^*. We define the Frobenius norm of A by $\|A\|_F^2 := \sum_{i=1}^d \|a_i\|^2$, where, for every $i \in [d] := \{1, \ldots, d\}$, a_i denotes the i-th row of A. We also denote by A_i the i-th column of A. We denote by $\operatorname{ran}(A)$ and $\ker(A)$ the range and the kernel of A, respectively.

3 Main problem and algorithm

Many applied problems require to estimate a quantity of interest $x \in \mathbb{R}^p$ based on linear measurements $b = Ax$ for some matrix $A \in \mathbb{R}^{d \times p}$. For simplicity, we do the analysis in the finite-dimensional case, but note that it can be easily extended to the infinite-dimensional setting. A standard approach to recover the desired solution is to assume that it is a minimizer of the following linearly constrained optimization problem:

$$\min_{x\in\mathbb{R}^p}\{J(x) : Ax = b\}, \tag{$\mathcal{P}$}$$

where $J \in \Gamma_0(\mathbb{R}^p)$ encodes a priori information on the solution and is usually hand-crafted. Typical choices are the squared norm [35], the elastic net regularization [32, 41, 45, 47, 84, 85], the ℓ^1-norm [24, 25, 33, 77], and the total variation [26, 58, 66, 67]. Note that, in the previous examples, the first two regularizers are strongly convex, while the second two are just convex and non-smooth.

If we use the indicator function of $\{b\}$, the problem $(\mathcal{P})$ can be written equivalently as

$$\min_{x\in\mathbb{R}^p} J(x) + \iota_{\{b\}}(Ax).$$

We denote by μ the optimal value of $(\mathcal{P})$ and by $\mathcal{S}$ the set of its minimizers. We assume that $\mathcal{S} \neq \emptyset$. In order to build our regularization procedure, we consider the Lagrangian functional for problem $(\mathcal{P})$:

$$\mathcal{L}\colon \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}, \quad (x,u) \mapsto J(x) + \langle u \mid Ax - b\rangle.$$

This approach allows us to split the contribution of the non-smooth term J and the one of the linear operator A, without requiring to compute the projection on the set

$$C := \{x \in \mathbb{R}^p \mid Ax = b\}.$$

We define the set of saddle points of $\mathcal{L}$ by

$$\mathcal{Z} = \{(x,u) \in \mathbb{R}^p \times \mathbb{R}^d : \mathcal{L}(x,v) \le \mathcal{L}(x,u) \le \mathcal{L}(y,u) \text{ for all } (y,v) \in \mathbb{R}^p \times \mathbb{R}^d\}.$$

The set $\mathcal{Z}$ is characterized by the first-order optimality condition:

$$\mathcal{Z} = \{(x,u) \in \mathbb{R}^p \times \mathbb{R}^d : 0 \in \partial J(x) + A^*u \text{ and } Ax = b\}.$$

In the following, we always assume that $\mathcal{Z} \neq \emptyset$.

Remark 3.1 (Saddle points and primal-dual solutions). Observe that the objective function of $(\mathcal{P})$ is the sum of two functions in $\Gamma_0(\mathbb{R}^p)$ where one of the two is composed with a linear operator. This formulation is suitable to apply Fenchel–Rockafellar duality. Recalling that $\iota^*_{\{b\}}(u) = \langle u \mid b\rangle$, the dual problem of $(\mathcal{P})$ is given by

$$\min_{u\in\mathbb{R}^d} J^*(-A^*u) + \langle u \mid b\rangle. \tag{$\mathcal{D}$}$$

We denote its optimal value by μ_* and its set of minimizers by $\mathcal{S}^*$. Then $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{S}^*$, and equality holds if the qualification condition (see [7, Proposition 6.19] for special cases when it holds)

$$b \in \operatorname{ri}(A(\operatorname{dom} J)) \tag{3.1}$$

is satisfied [7, Proposition 19.21 (v)]. In addition, condition (3.1) implies that problem $(\mathcal{D})$ has a solution. Then, under (3.1), since we assumed that $S \neq \emptyset$, we derive also that $\mathcal{Z} \neq \emptyset$.

In practical situations, the exact data b is unknown and only a noisy version is accessible. Given a noise level $\delta \ge 0$, we consider a worst case scenario, where the error is deterministic and the accessible data b^δ is such that

$$\|b^\delta - b\| \le \delta.$$

This is the classical model in inverse problems [35, 43]. The solution set of the inexact linear system $Ax = b^\delta$ is denoted by C_δ. Analogously, we denote by $\mathcal{S}_\delta$ and $\mathcal{S}^*_\delta$ the sets of primal and dual solutions with noisy data, respectively. It is worth pointing out that, if $b^\delta \notin \operatorname{ran}(A)$, then $\mathcal{S}_\delta \subseteq C_\delta = \emptyset$, but our analysis and bounds still hold.

3.1 Primal-dual splittings with a priori information

In this section, we propose an iterative regularization procedure to solve problem ($\mathcal{P}$), based on a primal-dual algorithm with preconditioning and arbitrary activations of a predefined set of operators. While the use of primal-dual algorithms [29] as iterative regularization methods is somewhat established [49], in this paper we focus on the possibility of reusing the data constraints along the iterations. This idea was originally introduced in [16], where the authors studied the case in which the exact data are available, and consists in the activation of extra operators, which encode information about the solution set to improve the feasibility of the updates. In our setting, we reuse data constraints, and we project, in series or in parallel, onto some equations given by the (noisy) linear constraints. But we will show that other interesting choices are possible, as projections onto the set of dual constraints.

More formally, for $i \in [m]$, we consider a finite number of operators $T_i \colon \mathbb{R}^p \to \mathbb{R}^p$ or $T_i \colon \mathbb{R}^d \to \mathbb{R}^d$ such that the set of noisy primal (or dual) solutions is contained in $\operatorname{Fix} T_i$ for every $i \in [m]$. We refer to this as redundant a priori information. A list of operators suitable to our setting (and with a cheap practical implementation) can be found in Section 5.

The primal-dual algorithms with reuse of data which are given in Table 1 are a preconditioned and deterministic version of the one proposed in [16] applied to the case of linearly constrained minimization.

Primal-dual splitting with activations

```
Input: $(\bar{p}^0, x^0, u^0) \in \mathbb{R}^{2p} \times \mathbb{R}^d$.
For $k = 1, \dots, N$:
        $u^{k+1} = u^k + \Gamma(A\bar{p}^k - b^\delta)$
        $x^{k+1} = \operatorname{prox}_J^\Sigma(p^k - \Sigma A^* u^{k+1})$
        Choose $\epsilon_{k+1} \in [m]$ and set                (PDA)
        $p^{k+1} = T_{\epsilon_{k+1}} x^{k+1}$
        $\bar{p}^{k+1} = p^{k+1} + x^{k+1} - p^k$,
End
```

Dual-primal splitting with activations

```
Input: $(\bar{v}^0, u^0, x^0) \in \mathbb{R}^{2d} \times \mathbb{R}^p$.
For $k = 1, \dots, N$:
        $x^{k+1} = \operatorname{prox}_J^\Sigma(x^k - \Sigma A^* \bar{v}^k)$
        $u^{k+1} = v^k + \Gamma(Ax^{k+1} - b^\delta)$
        Choose $\epsilon_{k+1} \in [m]$ and set                (DPA)
        $v^{k+1} = T_{\epsilon_{k+1}} u^{k+1}$
        $\bar{v}^{k+1} = v^{k+1} + u^{k+1} - v^k$,
End
```

Table 1: Proposed algorithms for iterative regularization.

We first focus on the primal-dual splitting. It is composed by four different steps, to be performed in series. The first step is the update of the dual variable, in which the residuals to the linear equation $Ax = b^\delta$ are accumulated after preconditioning by the operator Γ. The second step is an implicit prox-step, with function J and norm $\|\cdot\|_{\Sigma^{-1}}$, on the primal variable. The third one is the activation of the operator related to reusing data constraints, on the primal variable. Finally, the last step is an extrapolation again on the primal variable. Notice that, if no operator is activated, it corresponds simply to $\bar{p}^{k+1} = 2x^{k+1} - x^k$, that is, the classical update in the primal-dual algorithm. On the other hand, the dual-primal splitting algorithm, except for permutation in the order of the steps, differs from the previous one because the activation of the operator is done not on the primal variable but on the dual one. Indeed, Lemma A.1 establishes that, without the activation of the operator, there is an equivalence between the primal variables generated by (PDA) and the ones generated by (DPA).

Remark 3.2. Observe that in the proofs of convergence and stability (Theorems 4.1 and 4.2), we will never use that x belongs to a finite-dimensional space. This is in line with previous research on the convergence guarantees of the plain methods in Hilbert and Banach spaces, as outlined in [31, 73, 78]. It follows that the primal-dual algorithms above can be formulated exactly in the same way when the unknown vector x belongs to an infinite-dimensional Hilbert space, and our analysis can be extended to that setting. Another possible extension of the algorithm, which we do not analyze explicitly in this work, is related with the stochastic version of the primal-dual algorithm; see [1, 27, 39].

In the following, we list the assumptions that we require on the parameters and the operators involved in the algorithm.

Assumption 3.3. Consider the setting of (PDA) or (DPA).

(A1) The preconditioners $\Sigma \in \mathbb{R}^{p\times p}$ and $\Gamma \in \mathbb{R}^{d\times d}$ are two diagonal positive definite matrices such that

$$0 < \alpha := 1 - \|\Gamma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}\|^2. \tag{3.2}$$

(A2) For every $k \in \mathbb{N}$, $\epsilon_k \in [m]$.

Consider the setting of (PDA).

(A3) $\{T_i\}_{i\in[m]}$ is a family of operators from $\mathbb{R}^p$ to $\mathbb{R}^p$ and, for every $i \in [m]$:
 (a) $\operatorname{Fix} T_i \supseteq \mathcal{S}_\delta$;
 (b) there exists $e_i \geq 0$ such that, for every $x \in \mathbb{R}^p$ and $\bar{x} \in \mathcal{S}$,

$$\|T_i x - \bar{x}\|^2_{\Sigma^{-1}} \leq \|x - \bar{x}\|^2_{\Sigma^{-1}} + e_i \delta^2. \tag{3.3}$$

We set $e = \max_{i\in[m]} e_i$.

Now, consider the setting of (DPA).

(A4) $\{T_i\}_{i\in[m]}$ is a family of operators from $\mathbb{R}^d$ to $\mathbb{R}^d$ and, for every $i \in [m]$:
 (a) $\operatorname{Fix} T_i \supseteq \mathcal{S}^*_\delta$.
 (b) For every $u \in \mathbb{R}^d$ and $\bar{u} \in \mathcal{S}^*_\delta$,

$$\|T_i u - \bar{u}\|^2_{\Gamma^{-1}} \leq \|u - \bar{u}\|^2_{\Gamma^{-1}}. \tag{3.4}$$

Remark 3.4 (Hypothesis about the operators). If assumption (A3) (a) holds and $\delta = 0$, then assumption (A3) (b) is implied by the quasi-nonexpansivity of T_i on $\mathcal{S}$. This is a weaker condition than the one proposed in [16], where, due to the generality of the setting, α-averaged non-expansive operators were needed. A similar reasoning applies to assumption (A4).

4 Main results

In this section, we present and discuss the main results of the paper. We derive convergence and stability properties of the primal-dual and dual-primal splitting algorithms for linearly constrained optimization with a priori information.

First, we define the averaged iterates and the square weighted norm induced by Σ and Γ on $\mathbb{R}^p \times \mathbb{R}^d$, namely

$$(X^n, U^n) := \frac{\sum_{k=1}^n z^k}{n} \quad \text{and} \quad V(z) := \frac{\|x\|^2_{\Sigma^{-1}}}{2} + \frac{\|u\|^2_{\Gamma^{-1}}}{2},$$

where $z^k := (x^k, u^k)$ is the k-th iterate and $z := (x, u)$ is a primal-dual variable. We also recall the definition of the Lagrangian as $\mathcal{L}(x, u) = J(x) + \langle u \mid Ax - b\rangle$.

The first result establishes the stability properties of the algorithm (PDA), both in terms of the Lagrangian and the feasibility gap. We recall that here we use activation operators based on the noisy data and corresponding constraints in the primal space, namely the set C_δ.

Theorem 4.1. *Consider the setting of* (PDA) *under assumptions* (A1), (A2)*, and* (A3)*. Let* $(\bar{p}^0, x^0, u^0) \in \mathbb{R}^{2p} \times \mathbb{R}^d$ *be such that* $x^0 = \bar{p}^0$*. Then, for every* $z = (x, u) \in \mathcal{Z}$ *and for every* $N \in \mathbb{N}$*, we have*

$$\mathcal{L}(X^N, u) - \mathcal{L}(x, U^N) \leq \frac{V(z^0 - z)}{N} + \frac{2N\|\Gamma^{\frac{1}{2}}\|^2\delta^2}{\alpha} + \delta\|\Gamma^{\frac{1}{2}}\|\Big(\frac{2V(z^0 - z)}{\alpha}\Big)^{\frac{1}{2}} + \delta\|\Gamma^{\frac{1}{2}}\|\Big(\frac{Ne\delta^2}{\alpha}\Big)^{\frac{1}{2}} + \frac{e\delta^2}{2} \tag{4.1}$$

and

$$\begin{aligned} \|AX^N - b\|^2 \leq{} & \frac{16N\|\Gamma\|\|\Gamma^{-1}\|\delta^2}{\alpha^2} + 8\delta\|\Gamma^{-1}\|\Big(\frac{2\|\Gamma\|V(z^0 - z)}{\alpha^3}\Big)^{\frac{1}{2}} + 8\delta^2\|\Gamma^{-1}\|\Big(\frac{\|\Gamma\|eN}{\alpha^3}\Big)^{\frac{1}{2}} \\ & + \frac{8\|\Gamma^{-1}\|V(z^0 - z)}{N\alpha} + 2\delta^2 + \frac{4\|\Gamma^{-1}\|e\delta^2}{\alpha}, \end{aligned} \tag{4.2}$$

where we recall that the constant α *and* e *are defined in assumptions* (A1) *and* (A3)*, respectively.*

The proof of Theorem 4.1 is given in Section A.2. The proof combines and extends the techniques developed in [16, 49], based on the firm non-expansivity of the proximal point operator and discrete Bihari's lemma to deal with the error, see also [62].

In the next result, we establish upper bounds for the Lagrangian and feasibility gap analogous to those proposed in Theorem 4.1, but for the algorithm (DPA). The main difference is that now the activation step is based on a priori information in the dual space $\mathbb{R}^d$, and not on C_δ. This set is represented by the intersection of fixed point sets of a finite number of operators and encodes some knowledge about the dual solution.

Theorem 4.2. *Consider the setting of* (DPA) *under assumptions* (A1), (A2), *and* (A4). *Let* $(\bar{v}^0, u^0, x^0) \in \mathbb{R}^{2d} \times \mathbb{R}^p$ *be such that* $u^0 = \bar{v}^0$. *Then, for every* $z = (x, u) \in \mathcal{Z}$ *and for every* $N \in \mathbb{N}$, *we have that*

$$\mathcal{L}(X^N, u) - \mathcal{L}(x, U^N) \le \frac{V(z^0 - z)}{N} + 2\|\Gamma^{\frac{1}{2}}\|^2 N\delta^2 + \|\Gamma^{\frac{1}{2}}\|\delta(2V(z^0 - z))^{\frac{1}{2}} \tag{4.3}$$

and

$$\|AX^N - b\|^2 \le \frac{8\|\Gamma^{\frac{1}{2}}\|^2\|\Gamma^{-1}\|N\delta^2}{\alpha} + \frac{4\|\Gamma^{\frac{1}{2}}\|\|\Gamma^{-1}\|\delta(2V(z^0 - z))^{\frac{1}{2}}}{\alpha} + \frac{4\|\Gamma^{-1}\|V(z^0 - z)}{N\alpha} + 2\delta^2, \tag{4.4}$$

where we recall that the constant α *is defined in assumption* (A1).

The proof is given in Section A.3.

First, we comment on the chosen optimality measures. As discussed in [49, 50, 62], the Lagrangian gap is equivalent to the Bregman distance of the iterates to the solution. If the penalty is strongly convex, the Bregman divergence is an upper bound of the squared norm of the difference between the reconstructed and the ideal solution, while if J is only convex, the Bregman divergence gives only limited information and in general it is a very weak convergence measure. For instance, in the exact case, a vanishing Lagrangian gap does not imply that cluster points of the generated sequence are primal solutions. However, as can be derived from [50], a vanishing Lagrangian gap coupled with vanishing feasibility gap implies that every cluster point of the primal sequence is a solution of the primal problem.

In both theorems, the established result ensures that the two optimality measures can be upper bounded with the sum of two terms. The first one, which can be interpreted as an optimization error, is of the order $\mathcal{O}(N^{-1})$, and so it goes to zero as N tends to $+\infty$. Note that, in the exact case $\delta = 0$, only this term is present and both the Lagrangian and the feasibility gap are indeed vanishing, guaranteeing that every cluster point of the sequence is a primal solution. The second term, which can be interpreted as a stability control, collects all errors due to the perturbation of the exact datum and takes also into account the presence of the activation operators T, when the reuse data constraints are noisy. It is an increasing function of the number of iterations and the noise level δ.

Remark 4.3. Theorems 4.1 and 4.2 are an extension of [16, Theorem 1], where it is proved that the sequence generated by the algorithms converges to an element in $\mathcal{Z}$ when $\delta = 0$, but no convergence rates neither stability bounds were given. In this work, we filled the gap for linearly constrained convex optimization problems. Moreover, in the noise free case, our assumptions on the additional operators T are weaker than those proposed in [16], where α-averagedness is required. For the noisy case, without the activation operators (and so with $e = 0$), our bounds are of the same order as in [49] in the number of iterations and noise level (δ).

As mentioned above, in (4.1) and (4.2), when $\delta > 0$ and $N \to +\infty$, the upper bounds for the (PDA) iterates tend to infinity and the iteration may not converge to the desired solution. The same comment can be made for the (DPA) iterates, based on (4.3) and (4.4). In both cases, to obtain a minimal reconstruction error, we need to impose a trade-off between convergence and stability. The next corollary introduces an early stopping criterion, depending only on the noise level and leading to stable reconstruction.

Corollary 4.4 (Early-stopping). *Under the assumptions of Theorem 4.1 or Theorem 4.2, choose* $N = c/\delta$ *for some* $c > 0$. *Then, for every* $z = (x, u) \in \mathcal{Z}$, *there exist constants* C_1, C_2, *and* C_3 *such that*

$$\begin{aligned}\mathcal{L}(X^N, u) - \mathcal{L}(x, U^N) &\le C_1\delta,\\ \|AX^N - b\|^2 &\le C_2\delta + C_3\delta^2.\end{aligned}$$

The early stopping rule prescribed above is computationally efficient, in the sense that the number of iterations is proportional to the inverse of the noise level. In particular, if the error δ is small, then more iterations are useful, while if δ is big, it is convenient to stop sooner. So, the number of iterations plays the role of a regularization parameter. Using the early stopping strategy proposed above, we can see that the error in the data transfers to the error in the solution with the same noise level, which is the best that one can expect for a general operator A.

Remark 4.5 (Comparison with Tikhonov regularization). The reconstruction properties of the proposed algorithms are comparable to the ones obtained using Tikhonov regularization [10, 35], with the same dependence on the noise level. We underline that in [10, Theorem 5.1] only the Bregman divergence is considered, and not the feasibility. In addition, iterative regularization is way more efficient from the computational point of view, as it requires the solution of only one optimization problem, while Tikhonov regularization amounts to solve a family of problems indexed by the regularization parameter. Let us also note that, when δ is unknown, any principle used to determine a suitable λ can be used to determine the stopping time.

5 Implementation details

In this section, we discuss some standard choices to construct non-expansive operators T that satisfy our assumptions and encode some redundant information on the solution set. We first present examples for (PDA), and later for (DPA).

To define the operators, we first recall how to compute the projection on the constraint determined by each datum. For every $j \in [d]$, we denote by a_j the j-th row of A, and by P_j the projection onto the j-th linear equation, namely

$$P_j\colon \mathbb{R}^p \mapsto \mathbb{R}^p, \quad x \mapsto x + \frac{b_j - \langle a_j \mid x \rangle}{\|a_j\|^2} a_j^*.$$

Analogously, for every $j \in [d]$, we denote by P_j^δ the projection operator as in the previous definition, but with the noisy data b^δ instead of b.

We proceed to define the four families of operators proposed in this paper for (PDA).

Definition 5.1. Consider the operator $T\colon \mathbb{R}^p \mapsto \mathbb{R}^p$.

(i) T is a *serial projection* if

$$T = P_{\beta_l}^\delta \circ \cdots \circ P_{\beta_1}^\delta,$$

where, for every $j \in [l]$, $\beta_j \in [d]$.

(ii) T is a *parallel projection* if

$$T = \sum_{j=1}^{l} \alpha_j P_{\beta_j}^\delta, \tag{5.1}$$

where, for every $j \in [l]$, $\beta_j \in [d]$ and $(\alpha_j)_{j=1}^l$ are real numbers in $[0,1]$ such that $\sum_{j=1}^l \alpha_j = 1$.

(iii) T is a *Landweber operator* with parameter α if

$$T : \mathbb{R}^p \mapsto \mathbb{R}^p, \quad x \mapsto x - \alpha A^*(Ax - b^\delta), \tag{5.2}$$

where

$$\alpha \in \left]0, \frac{2}{\|A\|^2}\right[.$$

(iv) T is a *Landweber operator with adaptive step* and parameter M if

$$T\colon \mathbb{R}^p \mapsto \mathbb{R}^p, \qquad x \mapsto \begin{cases} x - \beta(x)A^*(Ax - b^\delta) & \text{if } A^*Ax \neq A^*b^\delta, \\ x & \text{otherwise.} \end{cases} \tag{5.3}$$

where, for $M > 0$,

$$\beta(x) = \min\Big(\frac{\|Ax - b^\delta\|^2}{\|A^*(Ax - b^\delta)\|^2}, M\Big).$$

The next lemma states that the operators in Definition 5.1 satisfy assumption (A3).

Lemma 5.2. *Let $T\colon \mathbb{R}^p \to \mathbb{R}^p$ be one of the operators given in Definition 5.1. Then assumption* (A3) *holds with the following error coefficients:*

(i) *If T is a serial projection, then*

$$e_T = \sum_{j=1}^{l} \frac{1}{\|a_{\beta_j}\|^2}.$$

(ii) *If T is a parallel projection, then*

$$e_T = \sum_{j=1}^{l} \frac{\alpha_j}{\|a_{\beta_j}\|^2}.$$

(iii) *If T is the Landweber operator with parameter α, then*

$$e_T = \frac{\alpha}{2 - \alpha\|A\|^2}.$$

(iv) *If T is the Landweber operator with adaptive step and parameter M, then $e_T = M$.*

Remark 5.3 (Relationship between parallel projection and Landweber operator). A particular parallel projection is the one corresponding to $l = d$, $\beta_j = j$, and

$$\alpha_j = \frac{\|a_j\|^2}{\|A\|_F^2}.$$

Then (5.1) reduces to

$$T(x) = x - \frac{1}{\|A\|_F^2} A^*(Ax - b^\delta). \tag{5.4}$$

Observe that, since $\|A\| \leq \|A\|_F$, the previous is a special case of the Landweber operator with $\alpha = 1/\|A\|_F^2$.

Remark 5.4 (Steepest descent). Let $\bar{x} \in \mathbb{R}^p$ be such that $A\bar{x} = b$. Then, from (5.3), we derive (see also equation (A.37))

$$\|Tx - \bar{x}\|^2 = \|x - \bar{x}\|^2 - 2\beta(x)\langle b^\delta - b \mid Ax - b^\delta \rangle - 2\beta(x)\|Ax - b^\delta\|^2 + \beta(x)^2\|A^*(Ax - b^\delta)\|^2. \tag{5.5}$$

If $\delta = 0$, then the choice of $\beta(x)$ given in (5.3) minimizes the right-hand side of (5.5) if the minimizer is smaller than M. In this case, β is chosen in order to maximize the contractivity with respect to a fixed point of T. While we cannot repeat the same procedure for $\delta > 0$, since we do not know b, we still keep the same choice. If $b^\delta \in \operatorname{ran}(A)$, then

$$\sup_{x \in \mathbb{R}^p} \|Ax - b^\delta\|^2 / \|A^*(Ax - b^\delta)\|^2 < +\infty.$$

However, in general, if $\delta > 0$, this is not true and M is needed to ensure that $\beta(x)$ is bounded.

Remark 5.5. From a computational point of view, parallel projections and Landweber operators are more efficient than serial projections. In particular, note that the quantity $(Ax^k - b^\delta)$ needs to be computed anyway in the other steps of the algorithm.

While for the primal space the data constraints that we want to reuse are clearly given by the linear constraints, for the dual there is not always a natural choice. In the following, we present an example related to the ℓ^1-norm regularization. A similar implementation can be extended to the case of 1-homogenous penalty functions, for which the Fenchel conjugate is the indicator of a closed and convex subset of the dual space [7, Proposition 14.11 (ii)].

Example 5.6. Consider the noisy version of problem ($\mathcal{P}$) with $J(x) = \|x\|_1$. Then the dual is given by

$$\min_{u \in \mathbb{R}^d} \{\langle b^\delta, u \rangle : |(A^*u)_i| \leq 1 \text{ for every } i \in [p]\}.$$

For every $i \in [p]$, set $D_i = \{u \in \mathbb{R}^d : |(A^*u)_i| \leq 1\}$ and denote by T_i the projection over D_i. Note that this projection is easy to compute, see for example [7, Example 28.17], since it is the projection onto the intersection of two parallel half-hyperplanes. Clearly, assumption (A4) holds. Differently from the primal case, here we are projecting on exact constraints which are independent of the noisy data b^δ.

6 Numerical results

In this section, to test the efficiency of the proposed algorithms, we perform numerical experiments in two settings: sparse reconstruction with ℓ^1-norm regularization, and image denoising and deblurring with total variation regularization. For the ℓ^1-norm regularization, we compare our results with other regularization techniques. In the more complex problem of total variation, we explore the properties of different variants of our procedure.

Code statement: All numerical examples are implemented in MATLAB® on a laptop. In the second experiment, we also use the library Numerical tours [59]. The corresponding code can be downloaded at: https://github.com/cristianvega1995/L1-TV-Experiments-of-Fast-iterative-regularization-by-reusing-data-constraints

6.1 ℓ^1-norm regularization

In this section, we apply the routines (PDA) and (DPA) when J is equal to the ℓ^1-norm. We compare the results given by our method with two state-of-the-art regularization procedures: iterative regularization by vanilla primal-dual [49], and Tikhonov explicit regularization, solving each problem by using the forward-backward algorithm [30]. In addition, we compare to another classical optimization algorithm for the minimization of the sum of two non-differentiable functions, namely the Douglas–Rachford algorithm [14]. In the noise free case, this algorithm is very effective in terms of number of iterations, but at each iteration it requires the explicit projection on the feasible set. In the noisy case, a stability analysis of the latter is not available.

We use the four variants of the algorithm (PDA) corresponding to the different choices of the operators T in Definition 5.1 and the version of (DPA) described in Example 5.6. Unless otherwise stated, in all experiments we use as preconditioners $\Sigma = \Gamma = \frac{0.99}{\|A\|}\operatorname{Id}$, which both satisfy (3.2).

Let $d = 2260$, $p = 3000$, and let $A \in \mathbb{R}^{d\times p}$ be such that every entry of the matrix is an independent sample from $\mathcal{N}(0,1)$, then normalized column by column. We set $b := Ax^*$, where $x^* \in \mathbb{R}^p$ is a sparse vector with approximately 300 nonzero entries uniformly distributed in the interval $[0,1]$. It follows from [36, Theorem 9.18] that x^* is the unique minimizer of the problem with probability bigger than 0.99. Let b^δ be such that $b^\delta = b + \|b\|u$, where the vector u is distributed, entry-wise, as $U[-0.2, 0.2]$. In this experiment, to test the reconstruction capabilities of our method, we use the exact datum x^* to establish the best stopping time, i.e. the one minimizing $\|x^k - x^*\|$. The exact solution is also used for the other regularization techniques. In a real practical situation, if δ is unknown, we would need to use parameter tuning techniques in order to select the optimal stopping time when both x^* and δ are unknown, but we do not address this aspect here.

We detail the used algorithms and their parameters below:

- *Tikhonov regularization* (Tik). We consider a grid of penalty parameters

$$G = \left\{\left(1 - \frac{l-1}{5}\right)10^{1-d}\|Ab^\delta\|_\infty : l \in [5],\ d \in [6]\right\}$$

  and, for each value $\lambda \in G$, the optimization problem

$$\min_{x\in\mathbb{R}^p}\left\{\lambda\|x\|_1 + \frac{1}{2}\|Ax - b^\delta\|^2\right\}. \tag{6.1}$$

  We solve each one of the previous problems with 300 iterations of the forward-backward algorithm, unless the stopping criterion $\|x^{k+1} - x^k\| \le 10^{-3}$ is satisfied for $k < 300$. Moreover, to deal efficiently with the sequence of problems, we use warm restart [9]. We first solve problem (6.1) for the biggest value of λ in G. Then we initialize the algorithm for the next value of λ, in decreasing order, with the solution reached for the previous one, and so on.
- *Douglas–Rachford* (DR). See [14, Theorem 3.1].
- *Primal-dual* (PD). This corresponds to PDA with $m = 1$ and $T_1 = \operatorname{Id}$.
- *Primal-dual with serial projections* (PDS). At every iteration, we compute a serial projection using all equations of the noisy system, where the order of the projections is given by a random shuffle.

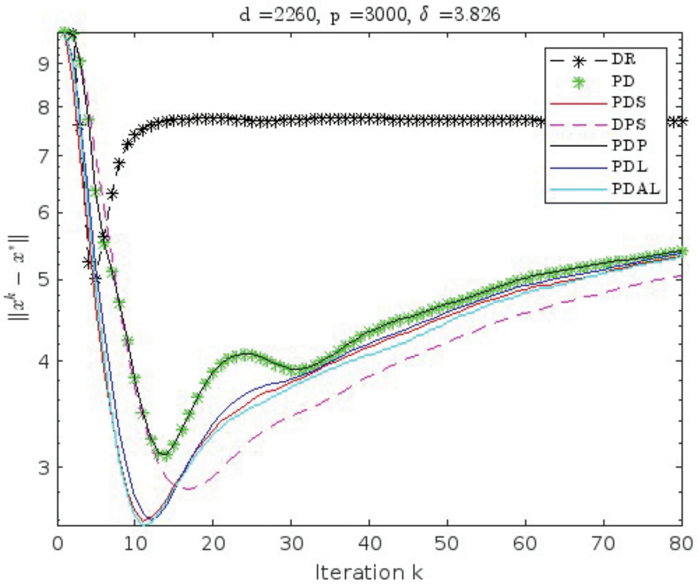

Figure 1: Graphical representation of early stopping. Note that the reconstruction error decreases and then increases, since the iterates first approach the exact solution and then converge to the noisy solution.

- *Primal-dual with parallel projections* (PDP). Set $m = 1$ and
$$T_1 x = x - \frac{1}{\|A\|_F^2} A^*(Ax - b^\delta);$$
see Remark 5.3.
- *Primal-dual Landweber* (PDL). Set $m = 1$ and
$$T_1 x = x - \frac{2}{\|A\|^2} A^*(Ax - b^\delta).$$
- *Primal-dual Landweber with adaptive step* (PDAL). Set $m = 1$ and $T_1 x = x - \beta(x)A^*(Ax - b^\delta)$, where
$$\beta(x) = \min\Big(\frac{\|Ax - b^\delta\|^2}{\|A^*(Ax - b^\delta)\|^2}, M\Big) \quad \text{for } M = 10^6.$$
- *Dual primal with serial projections* (DPS). At every iteration, we compute a serial projection over every inequality of $\|A^* u\|_\infty \leq 1$, where the order is given by a random shuffle of the rows of A^*.

	Time [in seconds]	Iteration	Reconstruction error
Tik	1.89	109	3.07
DR	3.08	5	5.01
PD	0.36	14	3.11
PDS	1.41	11	2.58
PDP	0.35	14	3.11
PDL	0.28	12	2.60
PDAL	0.27	11	2.56
DPS	0.54	17	2.83

Table 2: Run-time and number of iterations of each method until it reaches the best reconstruction error. We compare the proposed algorithms with Tikhonov regularization (Tik), Douglas–Rachford (DR), and iterative regularization (PD).

In Table 2, we reported also the number of iterations needed to achieve the best reconstruction error, but it is important to note that the iteration of each method has a different computational cost, so the run-time is a more appropriate comparison criterion.

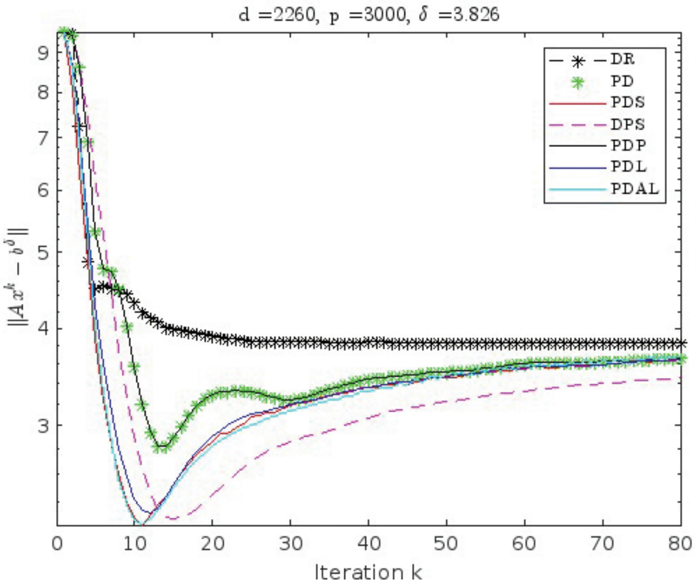

Figure 2: Early stopping with respect to the feasibility. Note that their behavior with respect to *k* is similar to that in Figure 1.

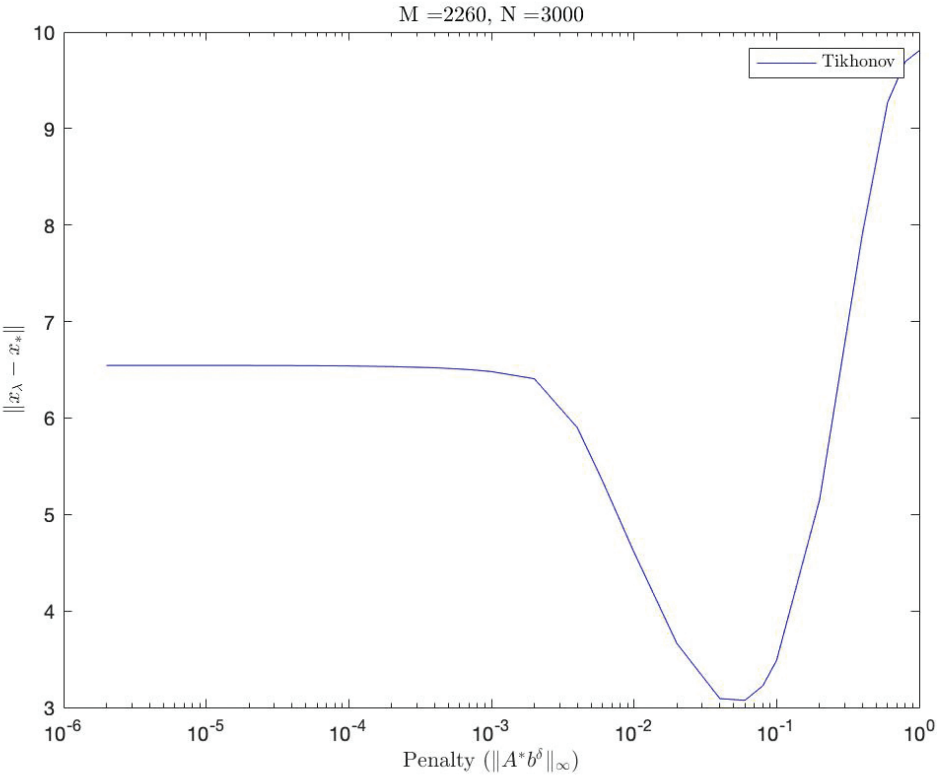

Figure 3: Reconstruction error of Tikhonov method with different penalties.

Douglas–Rachford with early stopping is the regularization method performing worst on this example, both in terms of time and reconstruction error. This behavior may be explained by the fact that this algorithm converges fast (meaning in few iterations) to the noisy solution, from which we infer that Douglas–Rachford is not a good algorithm for iterative regularization. Moreover, since we project on the noisy feasible set at every iteration, the resolution of a linear system is needed at every step. This also explains the cost of each iteration in terms of time. Note in addition that in our example b^δ is in the range of A, and so the noisy feasible set is nonempty. Tikhonov's regularization performs similarly in terms of time, but it requires many more (cheaper) iterations (see Figure 3). The achieved error is smaller than the one of DR, but bigger than the minimal one achieved by other methods.

Regarding our proposals, we observe that in Table 2 the proposed methods perform better than (PD). This supports the idea that reusing the constraints determined by the data is beneficial with respect to vanilla primal-dual. The benefit is not evident for (PDP), which achieves the worst reconstruction error, since $\|A\|_F^2$ is very big and so T_1 is very close to the identity. All other methods give better results in terms of reconstruction error. On the other hand, (PDS) is the slowest since it requires computing several projections at each iteration in a serial manner. We also observe that (PDL) and (PDAL) have better performance improving 22.2% and 25.0% in reconstruction error and 16.4% and 17.7% in run-time.

Figure 1 empirically shows the existence of the trade-off between convergence and stability for all algorithms, and therefore the advantage of early stopping. Similar results were obtained for the feasibility gap (see Figure 2).

6.2 Total variation

In this section, we perform several numerical experiments using the proposed algorithms for image denoising and deblurring. As done in the classical image denoising method introduced by Rudin, Osher, and Fantemi in [67], we rely on the total variation regularizer. See also [26, 28, 56, 58, 66, 67, 79]. We compare (PD) with (PDL) and (PDAL) algorithms, which were the algorithms performing the best in the previous application. In this section, we use two different preconditioners, which have been proved to be very efficient in practice [60].

Let $x^* \in \mathbb{R}^{N^2}$ represent an image with $N \times N$ pixels in $[0,1]$. We want to recover x^* from a blurry and noisy measurement y, i.e. from

$$y = Kx^* + \zeta,$$

where K is a linear blurring operator and ζ is a random noise vector. A standard approach is to assume that the original image is well approximated by the solution of the following constrained minimization problem:

$$\min_{u \in \mathbb{R}^{N \times N}} \{\|Du\|_{1,2} : Ku = y\}. \tag{TV}$$

Here,

$$\|\cdot\|_{1,2} \colon (\mathbb{R}^2)^{N \times N} \to \mathbb{R}, \quad p \mapsto \sum_{i=1}^{N}\sum_{j=1}^{N} \|p_{ij}\|,$$

and $D \colon \mathbb{R}^{N^2} \to (\mathbb{R}^2)^{N^2}$ is the discrete gradient operator for images, which is defined by

$$(Du)_{ij} = ((D_x u)_{ij}, (D_y u)_{ij})$$

with

$$(D_y u)_{ij} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if } 1 \le i \le N-1, \\ 0 & \text{if } i = N, \end{cases}$$

$$(D_x u)_{ij} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if } 1 \le j \le N-1, \\ 0 & \text{if } j = N. \end{cases}$$

In order to avoid the computation of the proximity operator of $\|D \cdot\|_{1,2}$, we introduce an auxiliary variable

$$v = Du \in Y := (\mathbb{R}^2)^{N^2}.$$

Since the value in each pixel must belong to $[0,1]$, we add the constraint $u \in X := [0,1]^{N^2}$. In this way, (TV) becomes

$$\min_{(u,v) \in X \times Y} \{\|v\|_{1,2} : Ku = y,\ Du = v\}. \tag{TV}$$

6.2.1 Formulation and algorithms

Problem (TV) is a special instance of ($\mathcal{P}$), with

$$\begin{cases} J\colon \mathbb{R}^{N^2} \times (\mathbb{R}^2)^{N^2} \to \mathbb{R} \cup \{+\infty\}, \quad x := (u, v) \mapsto \|v\|_{1,2} + \iota_X(u), \\ A = \begin{bmatrix} K & 0 \\ D & -\mathrm{Id} \end{bmatrix}, \quad b^\delta = \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad p = d = 3N^2. \end{cases}$$

Clearly, A is a linear nonzero operator, and $J \in \Gamma_0(\mathbb{R}^{N^2} \times (\mathbb{R}^2)^{N^2})$.

```
Primal-dual for total variation
Input: $(p^0, p^{-1}, x^0, v^0) \in \mathbb{R}^{6N^2} \times (\mathbb{R}^2)^{N^2}$ and $(q^0, q^{-1}, z^0, w^0) \in \mathbb{R}^{3N^2} \times \mathbb{R}^{N^2}$.
For $k = 1, \dots, L$:
    $v^{k+1} = v^k + \Gamma(K(p^k + x^k - p^{k-1})^k - y)$
    $w^{k+1} = w^k - \Gamma(q^k + z^k - q^{k-1}) + \Gamma D(p^k + x^k - p^{k-1})$
    $x^{k+1} = P_X(p^k - \Sigma K^* v^{k+1} + \Sigma w^{k+1})$
    $z^{k+1} = \mathrm{prox}_{\Sigma\|\cdot\|_{1,2}}(q^k - \Sigma D^* w^{k+1})$                (6.2)
    $p^{k+1} = x^k - \alpha(x^k)(K^*(Kx^k - y) + (Dx^k - z^k))$
    $q^{k+1} = q^k - \alpha(x^k)D^*(Dx^k - z^k)$
End
```

Table 3: General form of the algorithms.

We compare the algorithms listed below. Note that all proposed algorithms are different instances of the general routine described in Table 3, and each of them corresponds to a different choice of $\alpha(x^k)$:

(i) PD, the vanilla primal-dual algorithm, corresponding to $\alpha(x^k) = 0$.
(ii) PPD, the preconditioned primal-dual algorithm, obtained by $\alpha(x^k) = 0$ and Σ and Γ as in [60, Lemma 2].
(iii) PDL, corresponding to $\alpha(x^k) = 1/\|A\|^2$.
(iv) PDAL, corresponding to $\alpha(x^k) = \beta(x^k)$ as (5.3).

Initializing by $p^0 = \bar{p}^0 = x^0$ and $q^0 = \bar{q}^0 = z^0$, we recover the results of Theorem 4.1 and Corollary 4.4.

Remark 6.1. In order to implement the algorithm in (6.2), we first need to compute some operators:

(i) It follows from [7, Proposition 24.11] and [7, Example 24.20] that

$$\mathrm{prox}^{\Sigma}_{\|\cdot\|_{1,2}}(v) = (\mathrm{prox}^{\Sigma_i}_{\|\cdot\|}(v_i))_{i=1}^{N^2} = \left(\left(1 - \frac{\Sigma}{\max\{\Sigma, \|v\|\}}\right)v_i\right)_{i=1}^{N^2},$$

where $v_i \in \mathbb{R}^2$. The projection onto X can be computed as

$$P_X(u) = (P_{[0,1]}(u_i))_{i=1}^{N^2},$$

where

$$P_{[0,1]}(u_i) = \min\{1, \max\{u_i, 0\}\}.$$

(ii) It follows from [26] that

$$-D^* p = \operatorname{div} p = \begin{cases} (p_1)_{i,j} - (p_1)_{i-1,j} & \text{if } 1 < i < N \\ (p_1)_{i,j} & \text{if } i = 1, \\ -(p_1)_{i-1,j} & \text{if } i = N, \end{cases}$$

$$+ \begin{cases} (p_2)_{i,j} - (p_2)_{i,j-1} & \text{if } 1 < j < N, \\ (p_2)_{i,j} & \text{if } j = 1, \\ -(p_2)_{i,j-1} & \text{if } j = N. \end{cases}$$

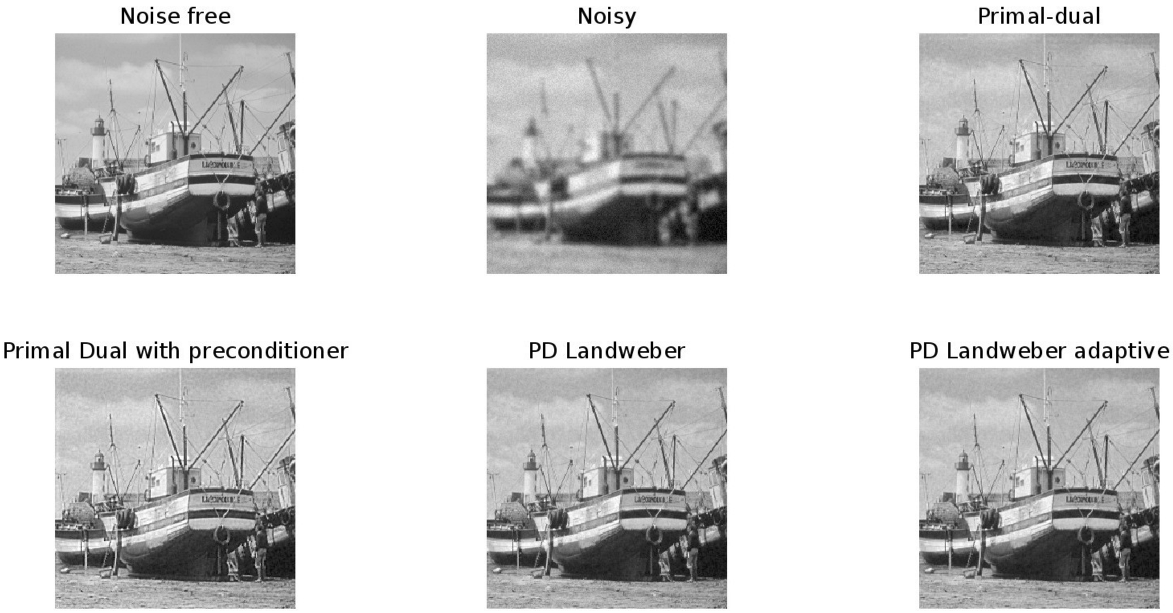

Figure 4: Qualitative comparison of the four proposed methods.

6.2.2 Numerical results

Set $N = 256$ and let x^* be the image "boat" in the library Numerical tours [59]. We suppose that K is an operator assigning to every pixel the average of the pixels in a neighborhood of radius 8 and that $\zeta \sim U([-0.025, 0.025])^{N^2}$. We use the original image as exact solution. For denoising and deblurring, we early stop the procedure at the iteration minimizing the mean square error (MSE), namely $\|x^k - x^*\|^2/N^2$, and we measure the time and the number of iterations needed to reach it. Another option for early stopping could be to consider the image with minimal structural similarity (SSIM). Numerically, in our experiments, this gives the same results. Additionally, we use the peak signal-to-noise ratio (PSNR) to compare the images. Note that the primal-dual algorithm with preconditioning is the method that needs less time and iterations among all procedures. Moreover, due to [29, Lemma 2], the condition (3.2) is automatically satisfied, while for the other methods we need to check it explicitly, which is computationally costly. However, (PPD) is the worst in terms of SSIM, PNSR, and MSE. We verify that all other algorithms have a superior performance in terms of reconstruction, with a small advantage for the Landweber with fixed and adaptive step-sizes, reducing the MSE of 94% with respect to the noisy image. In addition, compared to (PD), the algorithms (PDL) and (PDAL) require less iterations and time to satisfy the early stopping criterion. We believe that this is due to the fact that the extra Landweber operator improves the feasibility of the primal iterates. Visual assessment of the denoised and deblurred images are shown in Figure 4, which highlights the regularization properties achieved by the addition of the Landweber operator and confirms the previous conclusions.

	Iterations	Time	SSIM	PNSR	MSE
Noisy image	–	–	0.4468	21.4801	0.0071
PD	54	8.9773	0.8928	32.3614	0.0006
PD (precondition)	5	1.5515	0.8581	27.3753	0.0018
PDL	46	7.1846	0.9066	34.2174	0.0004
PDAL	31	5.4542	0.9112	34.3539	0.0004

Table 4: Quantitative comparison of the algorithms in terms of structural similarity (SSIM), peak signal-to-noise ratio (PSNR), mean square error (MSE), time, and iterations to reach the early stopping.

7 Conclusion and future work

In this paper, we studied two new iterative regularization methods for solving a linearly constrained minimization problem, based on an extra activation step reusing the data constraints. The analysis was carried out in the context of convex functions and worst-case deterministic noise. We proposed five instances of our algorithm and compared their numerical performance with state-of-the-art methods, and we observed considerable improvement in run-time.

In the future, we would like to extend Theorem 4.1 to structured convex problems and other algorithms. Possible extensions are: (1) the study of problems including, in the objective function, a L-smooth term and a composite linear term; (2) the analysis of random updates in the dual variable (see [27]) and stochastic approximations for the gradient; (3) the theoretical study of the impact of different preconditioners; (4) the improvement of the convergence and stability rates for strongly convex objective functions.

A Proofs

A.1 Equivalence between primal-dual and dual-primal algorithms

In the following lemma, we establish that, if $T = \mathrm{Id}$ and the initialization is the same, then there is an equivalence between the k-th primal variable of (PDA) and (DPA), denoted by u^k_{PD} and u^k_{DP}, respectively.

Lemma A.1. *Let*

$$(\bar{p}^0_{PD}, x^0_{PD}, u^0_{PD}) \in \mathbb{R}^{2p} \times \mathbb{R}^d \quad \textit{and} \quad (\bar{v}^0_{DP}, u^0_{DP}, x^0_{DP}) \in \mathbb{R}^{2d} \times \mathbb{R}^p$$

be the initialization (PDA) *and* (DPA)*, respectively, in the case when* $m = 1$ *and* $T = \mathrm{Id}$*. Suppose that* $x^0_{PD} = \bar{p}^0_{PD}$*,* $u^0_{DP} = \bar{v}^0_{DP}$*,* $u^0_{PD} = v^0_{DP}$*, and* $x^1_{PD} = x^1_{DP}$*. Then, for every* $k \in \mathbb{N}$*,* $x^k_{PD} = x^k_{DP}$*.*

Proof. Since $m = 1$ and $T = \mathrm{Id}$ in both algorithms, for every $k \in \mathbb{N}$, we have $x^k_{PD} = p^k_{PD}$ and $u^k_{DP} = v^k_{DP}$. On one hand, by the definition of (PDA), we have that

$$\begin{aligned} u^{k+1}_{PD} &= u^1_{PD} + \Gamma \sum_{i=1}^{k} (A\bar{p}^i_{PD} - b^\delta) \\ &= u^1_{PD} + \sum_{i=1}^{k} \Gamma A(p^i_{PD} - p^{i-1}_{PD}) + \Gamma \sum_{i=1}^{k} (Ax^i_{PD} - b^\delta) \\ &= u^1_{PD} + \Gamma A(p^k_{PD} - p^0_{PD}) + \Gamma \sum_{i=1}^{k} (Ax^i_{PD} - b^\delta) \\ &= u^0_{PD} + \Gamma(Ax^k_{PD} - b^\delta) + \Gamma \sum_{i=1}^{k} (Ax^i_{PD} - b^\delta), \end{aligned} \tag{A.1}$$

where the last equality is obtained since $p^0_{PD} = \bar{p}^0_{PD}$. Replacing (A.1) in the definition of x^{k+1}_{PD}, we obtain

$$x^{k+1}_{PD} = \operatorname{prox}^\Sigma_J \Big(x^k_{PD} - \Sigma A^* \Big(u^0_{PD} + \Gamma(Ax^k_{PD} - b^\delta) + \Gamma \sum_{i=1}^{k} (Ax^i_{PD} - b^\delta) \Big) \Big). \tag{A.2}$$

On the other hand, by (DPA) we have that

$$u^{k+1}_{DP} = v^{k+1}_{DP} = v^0_{DP} + \Gamma \sum_{i=1}^{k+1} (Ax^i_{DP} - b^\delta)$$

and

$$\bar{v}^k_{DP} = v^k_{DP} + u^k_{DP} - v^{k-1}_{DP} = v^0_{DP} + \Gamma(Ax^k_{DP} - b^\delta) + \Gamma \sum_{i=1}^{k} (Ax^i_{DP} - b^\delta). \tag{A.3}$$

Replacing (A.3) in (DPA), for every $k > 1$, we can deduce that

$$x_{DP}^{k+1} = \operatorname{prox}_J^{\Sigma}\Big(x_{DP}^k - \Sigma A^*\Big(v_{DP}^0 + \Gamma(Ax_{DP}^k - b^\delta) + \Gamma\sum_{i=1}^{k}(Ax_{DP}^i - b^\delta)\Big)\Big).$$

Since $u_{PD}^0 = v_{DP}^0$ and $x_{PD}^1 = x_{DP}^1$, the result follows by induction. □

Remark A.2. An analysis similar to that in the proof of Lemma A.1 shows that

$$x_{PD}^{k+1} = \operatorname{prox}_J^{\Sigma}\Big(x_{PD}^k - \Sigma A^*\Big(u_{PD}^0 + \Gamma(AT_{\epsilon_k}x_{PD}^k - b^\delta) + \Gamma\sum_{i=1}^{k}(Ax_{PD}^i - b^\delta)\Big)\Big),$$

which implies that the algorithm can be written in one step if we only care about the primal variable.

A.2 Proof of Theorem 4.1

Proof. From (PDA), we deduce that

$$\begin{aligned}\Sigma^{-1}(p^k - x^{k+1}) - A^*u^{k+1} &\in \partial J(x^{k+1}),\\ \Gamma^{-1}(u^k - u^{k+1}) + A\bar{p}^k &= b^\delta.\end{aligned} \tag{A.4}$$

Therefore, we have

$$J(x^{k+1}) + \langle \Sigma^{-1}(p^k - x^{k+1}) - A^*u^{k+1} \mid x - x^{k+1}\rangle \le J(x) \quad \text{for all } x \in \mathbb{R}^p, \tag{A.5}$$

and (A.5) yields

$$\begin{aligned}0 &\ge J(x^{k+1}) - J(x) + \langle \Sigma^{-1}(p^k - x^{k+1}) - A^*u^{k+1} \mid x - x^{k+1}\rangle\\ &= J(x^{k+1}) - J(x) + \frac{\|p^k - x^{k+1}\|_{\Sigma^{-1}}^2}{2} + \frac{\|x^{k+1} - x\|_{\Sigma^{-1}}^2}{2} - \frac{\|p^k - x\|_{\Sigma^{-1}}^2}{2} + \langle x^{k+1} - x \mid A^*u^{k+1}\rangle.\end{aligned} \tag{A.6}$$

Analogously, by (A.4), we get

$$\begin{aligned}0 &= \langle \Gamma^{-1}(u^k - u^{k+1}) + A\bar{p}^k - b^\delta \mid u - u^{k+1}\rangle\\ &= \frac{\|u^{k+1} - u^k\|_{\Gamma^{-1}}^2}{2} + \frac{\|u^{k+1} - u\|_{\Gamma^{-1}}^2}{2} - \frac{\|u^k - u\|_{\Gamma^{-1}}^2}{2} + \langle b^\delta - A\bar{p}^k \mid u^{k+1} - u\rangle.\end{aligned} \tag{A.7}$$

Recall that

$$z := (x, u) \in \mathcal{Z} \subset C \times \mathbb{R}^d, \quad z^k := (x^k, u^k), \quad V(z) := \frac{\|x\|_{\Sigma^{-1}}^2}{2} + \frac{\|u\|_{\Gamma^{-1}}^2}{2}.$$

Summing (A.6) and (A.7), and by assumption (A3), we obtain

$$\begin{aligned}J(x^{k+1}) - J(x) &+ \frac{\|x^{k+1} - p^k\|_{\Sigma^{-1}}^2}{2} + \frac{\|u^{k+1} - u^k\|_{\Gamma^{-1}}^2}{2} + V(z^{k+1} - z) - V(z^k - z)\\ &+ \langle A(x^{k+1} - x) \mid u^{k+1}\rangle + \langle b^\delta - A\bar{p}^k \mid u^{k+1} - u\rangle - \frac{e\delta^2}{2}\\ &\le 0.\end{aligned} \tag{A.8}$$

Now, compute

$$\begin{aligned}J(x^{k+1}) - J(x) &+ \langle A(x^{k+1} - x) \mid u^{k+1}\rangle + \langle b^\delta - A\bar{p}^k \mid u^{k+1} - u\rangle\\ &= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) - \langle Ax^{k+1} - b \mid u\rangle + \langle Ax - b \mid u^{k+1}\rangle\\ &\qquad + \langle A(x^{k+1} - x) \mid u^{k+1}\rangle + \langle b^\delta - A\bar{p}^k \mid u^{k+1} - u\rangle\\ &= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) - \langle Ax^{k+1} \mid u\rangle + \langle b \mid u\rangle + \langle Ax \mid u^{k+1}\rangle - \langle b \mid u^{k+1}\rangle\\ &\qquad + \langle Ax^{k+1} \mid u^{k+1}\rangle - \langle Ax \mid u^{k+1}\rangle + \langle b^\delta \mid u^{k+1} - u\rangle - \langle A\bar{p}^k \mid u^{k+1} - u\rangle\\ &= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) + \langle b^\delta - b \mid u^{k+1} - u\rangle + \langle Ax^{k+1} - A\bar{p}^k \mid u^{k+1} - u\rangle\\ &\ge \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) - \delta\|\Gamma^{\frac{1}{2}}\|\|u^{k+1} - u\|_{\Gamma^{-1}} + \langle Ax^{k+1} - A\bar{p}^k \mid u^{k+1} - u\rangle.\end{aligned} \tag{A.9}$$

From (A.9) and (A.8), we obtain

$$\begin{aligned}
\mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) &+ \frac{\|x^{k+1} - p^k\|^2_{\Sigma^{-1}}}{2} + \frac{\|u^{k+1} - u^k\|^2_{\Gamma^{-1}}}{2} \\
&+ V(z^{k+1} - z) - V(z^k - z) - \delta\|\Gamma^{\frac{1}{2}}\|\|u^{k+1} - u\|_{\Gamma^{-1}} - \frac{e\delta^2}{2} \\
&\leq -\langle A(x^{k+1} - \bar{p}^k) \mid u^{k+1} - u\rangle \qquad\qquad \text{(A.10)} \\
&= -\langle A(x^{k+1} - p^k) \mid u^{k+1} - u\rangle + \langle A(x^k - p^{k-1}) \mid u^k - u\rangle + \langle A(x^k - p^{k-1}) \mid u^{k+1} - u^k\rangle \\
&= -\langle A(x^{k+1} - p^k) \mid u^{k+1} - u\rangle + \langle A(x^k - p^{k-1}) \mid u^k - u\rangle \\
&\quad + \langle \Gamma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} (x^k - p^{k-1}) \mid \Gamma^{-\frac{1}{2}}(u^{k+1} - u^k)\rangle \\
&\leq -\langle A(x^{k+1} - p^k) \mid u^{k+1} - u\rangle + \langle A(x^k - p^{k-1}) \mid u^k - u\rangle \\
&\quad + \|\Gamma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}\|^2 \frac{\|u^{k+1} - u^k\|^2_{\Gamma^{-1}}}{2} + \frac{\|x^k - p^{k-1}\|^2_{\Sigma^{-1}}}{2}. \qquad\qquad \text{(A.11)}
\end{aligned}$$

Then, recalling that $\alpha = 1 - \|\Gamma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}\|^2$, we have the following estimate:

$$\begin{aligned}
\mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) &+ \frac{\|x^{k+1} - p^k\|^2_{\Sigma^{-1}}}{2} - \frac{\|x^k - p^{k-1}\|^2_{\Sigma^{-1}}}{2} + \frac{\alpha}{2}\|u^{k+1} - u^k\|^2_{\Gamma^{-1}} + V(z^{k+1} - z) - V(z^k - z) \\
&\leq \delta\|\Gamma^{\frac{1}{2}}\|\|u^{k+1} - u\|_{\Gamma^{-1}} - \langle A(x^{k+1} - p^k) \mid u^{k+1} - u\rangle + \langle A(x^k - p^{k-1}) \mid u^k - u\rangle + \frac{e\delta^2}{2}.
\end{aligned}$$

Summing from 1 to $N - 1$, we obtain

$$\begin{aligned}
\sum_{k=1}^{N-1} (\mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1})) &+ \frac{\|x^N - p^{N-1}\|^2_{\Sigma^{-1}}}{2} - \frac{\|x^1 - p^0\|^2_{\Sigma^{-1}}}{2} \\
&+ \frac{\alpha}{2} \sum_{k=1}^{N-1} \|u^{k+1} - u^k\|^2_{\Gamma^{-1}} + V(z^N - z) - V(z^1 - z) - \langle A(x^1 - p^0) \mid u^1 - u\rangle \\
&\leq \delta\|\Gamma^{\frac{1}{2}}\| \sum_{k=1}^{N} \|u^k - u\|_{\Gamma^{-1}} - \langle \Gamma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} (x^N - p^{N-1}) \mid \Gamma^{-\frac{1}{2}}(u^N - u)\rangle + \frac{(N-1)e\delta^2}{2} \qquad \text{(A.12)} \\
&\leq \delta\|\Gamma^{\frac{1}{2}}\| \sum_{k=1}^{N} \|u^k - u\|_{\Gamma^{-1}} + \frac{\|x^N - p^{N-1}\|^2_{\Sigma^{-1}}}{2} + \|\Gamma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}\|^2 \frac{\|u^N - u\|^2_{\Gamma^{-1}}}{2} + \frac{(N-1)e\delta^2}{2}.
\end{aligned}$$

Now, by choosing $k = 0$ in (A.10), we get

$$\begin{aligned}
\mathcal{L}(x^1, u) - \mathcal{L}(x, u^1) &+ \frac{\|x^1 - p^0\|^2_{\Sigma^{-1}}}{2} + \frac{\alpha}{2}\|u^1 - u^0\|^2_{\Gamma^{-1}} + V(z^1 - z) - V(z^0 - z) + \langle A(x^1 - \bar{p}^0) \mid u^1 - u\rangle \\
&\leq \delta\|\Gamma^{\frac{1}{2}}\|\|u^1 - u\|_{\Gamma^{-1}} + \frac{e\delta^2}{2}.
\end{aligned} \tag{A.13}$$

Adding (A.12) and (A.13), we obtain

$$\begin{aligned}
\sum_{k=0}^{N-1} (\mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1})) &+ \frac{\alpha}{2}\|u^N - u\|^2_{\Gamma^{-1}} + \sum_{k=1}^{N} \frac{\alpha}{2}\|u^k - u^{k-1}\|^2_{\Gamma^{-1}} + \frac{\|x^N - x\|^2_{\Sigma^{-1}}}{2} \\
&\leq \delta\|\Gamma^{\frac{1}{2}}\| \sum_{k=1}^{N} \|u^k - u\|_{\Gamma^{-1}} + V(z^0 - z) + \frac{Ne\delta^2}{2}.
\end{aligned} \tag{A.14}$$

Next, by (A.11), we have the following estimate:

$$\begin{aligned}
\frac{\|x^{k+1} - p^k\|^2_{\Sigma^{-1}}}{2} &- \langle A(x^k - p^{k-1}) \mid u^{k+1} - u^k\rangle + \frac{\|u^{k+1} - u^k\|^2_{\Gamma^{-1}}}{2} \\
&+ \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) + V(z^{k+1} - z) - V(z^k - z) \\
&\leq \delta\|\Gamma^{\frac{1}{2}}\|\|u^{k+1} - u\|_{\Gamma^{-1}} - \langle A(x^{k+1} - p^k) \mid u^{k+1} - u\rangle + \langle A(x^k - p^{k-1}) \mid u^k - u\rangle + \frac{e\delta^2}{2}.
\end{aligned}$$

Summing from 1 to $N-1$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{N-1}\Big(\frac{\|x^{k+1}-p^{k}\|^2_{\Sigma^{-1}}}{2}-\langle A(x^{k}-p^{k-1})\mid u^{k+1}-u^{k}\rangle+\frac{\|u^{k+1}-u^{k}\|^2_{\Gamma^{-1}}}{2}\Big)\\
&\qquad+\sum_{k=1}^{N-1}(\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1}))+V(z^{N}-z)-V(z^{1}-z)-\langle A(x^{1}-p^{0})\mid u^{1}-u\rangle\\
&\quad\le\delta\|\Gamma^{\frac{1}{2}}\|\sum_{k=1}^{N-1}\|u^{k+1}-u\|_{\Gamma^{-1}}-\langle A(x^{N}-p^{N-1})\mid u^{N}-u\rangle+\frac{(N-1)e\delta^2}{2}\\
&\quad=\delta\|\Gamma^{\frac{1}{2}}\|\sum_{k=1}^{N-1}\|u^{k+1}-u\|_{\Gamma^{-1}}-\langle\Gamma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}(x^{N}-p^{N-1})\mid\Gamma^{-\frac{1}{2}}(u^{N}-u)\rangle+\frac{(N-1)e\delta^2}{2}\\
&\quad\le\delta\|\Gamma^{\frac{1}{2}}\|\sum_{k=1}^{N-1}\|u^{k+1}-u\|_{\Gamma^{-1}}+\frac{\|\Gamma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\|^2}{2}\|x^{N}-p^{N-1}\|^2_{\Sigma^{-1}}+\frac{\|u^{N}-u\|^2_{\Gamma^{-1}}}{2}+\frac{(N-1)e\delta^2}{2}.
\end{aligned}\tag{A.15}$$

Now, since

$$u^{k+1}-u^{k}=\Gamma(A\bar{p}^{k}-b^{\delta}),$$

we derive that

$$\begin{aligned}
&\sum_{k=1}^{N-1}\Big(\frac{\|x^{k+1}-p^{k}\|^2_{\Sigma^{-1}}}{2}-\langle A(x^{k}-p^{k-1})\mid u^{k+1}-u^{k}\rangle+\frac{\|u^{k+1}-u^{k}\|^2_{\Gamma^{-1}}}{2}\Big)\\
&\quad=\sum_{k=1}^{N-1}\Big(\frac{\|x^{k}-p^{k-1}\|^2_{\Sigma^{-1}}}{2}-\langle A(x^{k}-p^{k-1})\mid u^{k+1}-u^{k}\rangle+\frac{\|u^{k+1}-u^{k}\|^2_{\Gamma^{-1}}}{2}\Big)\\
&\qquad+\frac{\|x^{N}-p^{N-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|x^{1}-p^{0}\|^2_{\Sigma^{-1}}}{2}\\
&\quad=\sum_{k=1}^{N-1}\Big(\frac{\|\Gamma^{\frac{1}{2}}A(x^{k}-p^{k-1})\|^2}{2}-\langle\Gamma^{\frac{1}{2}}A(x^{k}-p^{k-1})\mid\Gamma^{\frac{1}{2}}(A\bar{p}^{k}-b^{\delta})\rangle+\frac{\|\Gamma^{\frac{1}{2}}(A\bar{p}^{k}-b^{\delta})\|^2}{2}\Big)\\
&\qquad+\frac{\|x^{N}-p^{N-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|x^{1}-p^{0}\|^2}{2}+\sum_{k=1}^{N-1}\Big(\frac{\|x^{k}-p^{k-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|\Gamma^{\frac{1}{2}}A(x^{k}-p^{k-1})\|^2}{2}\Big)\\
&\quad=\sum_{k=1}^{N-1}\frac{\|\Gamma^{\frac{1}{2}}(Ap^{k}-b^{\delta})\|^2}{2}+\frac{\|x^{N}-p^{N-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|x^{1}-p^{0}\|^2_{\Sigma^{-1}}}{2}\\
&\qquad+\sum_{k=1}^{N-1}\Big(\frac{\|x^{k}-p^{k-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|\Gamma^{\frac{1}{2}}A(x^{k}-p^{k-1})\|^2}{2}\Big).
\end{aligned}$$

Furthermore, since

$$\alpha=1-\|\Gamma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\|^2>0,$$

we obtain

$$\begin{aligned}
&\sum_{k=1}^{N-1}\frac{\|\Gamma^{\frac{1}{2}}(Ap^{k}-b^{\delta})\|^2}{2}+\frac{\|x^{N}-p^{N-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|x^{1}-p^{0}\|^2_{\Sigma^{-1}}}{2}\\
&\qquad+\sum_{k=1}^{N-1}\Big(\frac{\|x^{k}-p^{k-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|\Gamma^{\frac{1}{2}}A(x^{k}-p^{k-1})\|^2}{2}\Big)\\
&\quad\ge\sum_{k=1}^{N-1}\frac{\|\Gamma^{\frac{1}{2}}(Ap^{k}-b^{\delta})\|^2}{2}+\frac{\|x^{N}-p^{N-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|x^{1}-p^{0}\|^2_{\Sigma^{-1}}}{2}+\frac{\alpha}{2}\sum_{k=1}^{N-1}\|\Gamma^{\frac{1}{2}}A(x^{k}-p^{k-1})\|^2\\
&\quad\ge\sum_{k=1}^{N-1}\frac{\|\Gamma^{\frac{1}{2}}(Ap^{k}-b^{\delta})\|^2}{2}+\frac{\|x^{N}-p^{N-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|x^{1}-p^{0}\|^2_{\Sigma^{-1}}}{2}\\
&\qquad-\frac{\alpha}{2}\|\Gamma^{\frac{1}{2}}A(x^{N}-p^{N-1})\|^2+\frac{\alpha}{2}\|\Gamma^{\frac{1}{2}}A(x^{1}-p^{0})\|^2+\frac{\alpha}{2}\sum_{k=1}^{N-1}\|\Gamma^{\frac{1}{2}}A(x^{k+1}-p^{k})\|^2.
\end{aligned}$$

In turn, by the convexity of $\|\cdot\|^2$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{N-1}\frac{\|\Gamma^{\frac{1}{2}}(Ap^k-b^\delta)\|^2}{2}+\frac{\|x^N-p^{N-1}\|^2_{\Sigma^{-1}}}{2}-\frac{\|x^1-p^0\|^2_{\Sigma^{-1}}}{2}\\
&\qquad-\frac{\alpha}{2}\|\Gamma^{\frac{1}{2}}A(x^N-p^{N-1})\|^2+\frac{\alpha}{2}\|\Gamma^{\frac{1}{2}}A(x^1-p^0)\|^2+\frac{\alpha}{2}\sum_{k=1}^{N-1}\|\Gamma^{\frac{1}{2}}A(x^{k+1}-p^k)\|^2\\
&\geq\frac{\alpha}{4}\sum_{k=1}^{N-1}\|\Gamma^{\frac{1}{2}}(Ax^{k+1}-b^\delta)\|^2-\frac{\|x^1-p^0\|^2_{\Sigma^{-1}}}{2}+\frac{\alpha}{2}\|\Gamma^{\frac{1}{2}}A(x^1-p^0)\|^2+\frac{\alpha^2+\|\Gamma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\|^2}{2}\|x^N-p^{N-1}\|^2_{\Sigma^{-1}}\\
&\geq\frac{\alpha}{4}\sum_{k=2}^{N}\|\Gamma^{\frac{1}{2}}(Ax^{k}-b^\delta)\|^2-\frac{\|x^1-p^0\|^2_{\Sigma^{-1}}}{2}+\frac{\alpha}{2}\|\Gamma^{\frac{1}{2}}A(x^1-p^0)\|^2+\frac{\alpha^2+\|\Gamma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\|^2}{2}\|x^N-p^{N-1}\|^2_{\Sigma^{-1}}.
\end{aligned}\tag{A.16}$$

On the other hand, we get

$$\|\Gamma^{\frac{1}{2}}(Ax^k-b^\delta)\|^2\geq\frac{\|Ax^k-b^\delta\|^2}{\|\Gamma^{-1}\|}\geq\frac{1}{\|\Gamma^{-1}\|}\Big(\frac{\|Ax^k-b\|^2}{2}-\|b^\delta-b\|^2\Big).\tag{A.17}$$

Combining (A.13), (A.15), (A.16), and (A.17), we have that

$$\begin{aligned}
&\sum_{k=0}^{N-1}(\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1}))+\frac{\alpha^2}{2}\|x^N-p^{N-1}\|^2_{\Sigma^{-1}}\sum_{k=1}^{N}\frac{\alpha}{8\|\Gamma^{-1}\|}\|Ax^{k+1}-b\|^2+\frac{\|x^N-x\|^2_{\Sigma^{-1}}}{2}\\
&\qquad\leq\delta\|\Gamma^{\frac{1}{2}}\|\sum_{k=1}^{N}\|u^k-u\|_{\Gamma^{-1}}+V(z^0-z)+\frac{Ne\delta^2}{2}+N\frac{\alpha}{4\|\Gamma^{-1}\|}\delta^2.
\end{aligned}\tag{A.18}$$

It remains to bound $\delta\|\Gamma^{\frac{1}{2}}\|\sum_{k=1}^{N}\|u^k-u\|_{\Gamma^{-1}}$. From (A.14) and since (x,u) is a saddle-point of the Lagrangian, we deduce that

$$\|u^N-u\|^2\leq\frac{2\|\Gamma^{\frac{1}{2}}\|\delta}{\alpha}\sum_{k=1}^{N}\|u^k-u\|+\frac{2V(z^0-z)}{\alpha}+\frac{Ne\delta^2}{\alpha}.\tag{A.19}$$

Applying [62, Lemma A.1] to equation (A.19) with

$$\lambda_k:=\frac{2\|\Gamma^{\frac{1}{2}}\|\delta}{\alpha}\quad\text{and}\quad S_N:=\frac{2V(z^0-z)}{\alpha}+\frac{Ne\delta^2}{\alpha},$$

we get

$$\begin{aligned}
\|u^N-u\|&\leq\frac{N\|\Gamma^{\frac{1}{2}}\|\delta}{\alpha}+\Big(\frac{2V(z^0-z)}{\alpha}+\frac{Ne\delta^2}{\alpha}+\Big(\frac{N\|\Gamma^{\frac{1}{2}}\|\delta}{\alpha}\Big)^2\Big)^{\frac{1}{2}}\\
&\leq\frac{2N\|\Gamma^{\frac{1}{2}}\|\delta}{\alpha}+\Big(\frac{2V(z^0-z)}{\alpha}\Big)^{\frac{1}{2}}+\Big(\frac{Ne\delta^2}{\alpha}\Big)^{\frac{1}{2}}.
\end{aligned}$$

Insert the previous in equation (A.14) to obtain

$$\begin{aligned}
&\sum_{k=0}^{N-1}(\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1}))\\
&\qquad\leq\frac{2(N\|\Gamma^{\frac{1}{2}}\|\delta)^2}{\alpha}+N\|\Gamma^{\frac{1}{2}}\|\delta\Big(\frac{V(z^0-z)}{\alpha}\Big)^{\frac{1}{2}}+N\|\Gamma^{\frac{1}{2}}\|\delta\Big(\frac{Ne\delta^2}{\alpha}\Big)^{\frac{1}{2}}+V(z^0-z)+\frac{Ne\delta^2}{2}.
\end{aligned}$$

Analogously, from (A.18),

$$\begin{aligned}
\sum_{k=1}^{N}\|Ax^k-b\|^2&\leq\frac{16N^2\|\Gamma\|\|\Gamma^{-1}\|\delta^2}{\alpha^2}+8N\delta\|\Gamma^{-1}\|\Big(\frac{2\|\Gamma\|V(z^0-z)}{\alpha^3}\Big)^{\frac{1}{2}}+8N\delta^2\|\Gamma^{-1}\|\Big(\frac{\|\Gamma\|eN}{\alpha^3}\Big)^{\frac{1}{2}}\\
&\qquad+\frac{8\|\Gamma^{-1}\|V(z^0-z)}{\alpha}+2N\delta^2+\frac{4N\|\Gamma^{-1}\|e\delta^2}{\alpha},
\end{aligned}$$

and both results are straightforward from Jensen’s inequality. □

A.3 Proof of Theorem 4.2

Proof. It follows from (DPA) that

$$\begin{aligned}\Sigma^{-1}(x^k - x^{k+1}) - A^*\bar{v}^k &\in \partial J(x^{k+1}),\\ \Gamma^{-1}(v^k - u^{k+1}) + Ax^{k+1} &= b^\delta.\end{aligned}\tag{A.20}$$

Thus,

$$J(x^{k+1}) + \langle \Sigma^{-1}(x^k - x^{k+1}) - A^*\bar{v}^k \mid x - x^{k+1}\rangle \le J(x),\tag{A.21}$$

and (A.21) yields

$$\begin{aligned}0 &\ge J(x^{k+1}) - J(x) + \langle \Sigma^{-1}(x^k - x^{k+1}) - A^*\bar{v}^k \mid x - x^{k+1}\rangle\\ &= J(x^{k+1}) - J(x) + \frac{\|x^k - x^{k+1}\|^2_{\Sigma^{-1}}}{2} + \frac{\|x^{k+1} - x\|^2_{\Sigma^{-1}}}{2} - \frac{\|x^k - x\|^2_{\Sigma^{-1}}}{2} + \langle x^{k+1} - x \mid A^*\bar{v}^k\rangle.\end{aligned}\tag{A.22}$$

From (A.20), it follows that

$$\begin{aligned}0 &= \langle \Gamma^{-1}(v^k - u^{k+1}) + Ax^{k+1} - b^\delta \mid u - u^{k+1}\rangle\\ &= \frac{\|u^{k+1} - v^k\|^2_{\Gamma^{-1}}}{2} + \frac{\|u^{k+1} - u\|^2_{\Gamma^{-1}}}{2} - \frac{\|v^k - u\|^2_{\Gamma^{-1}}}{2} + \langle b^\delta - Ax^{k+1} \mid u^{k+1} - u\rangle.\end{aligned}\tag{A.23}$$

Recall that

$$z := (x, u) \in \mathcal{Z} \subset C \times \mathbb{R}^d, \quad z^k := (x^k, u^k), \quad V(z) := \frac{\|x\|^2_{\Sigma^{-1}}}{2} + \frac{\|u\|^2_{\Gamma^{-1}}}{2}.$$

Summing (A.22) and (A.23), we obtain

$$\begin{aligned}J(x^{k+1}) - J(x) + \frac{\|x^{k+1} - x^k\|^2_{\Sigma^{-1}}}{2} + \frac{\|u^{k+1} - v^k\|^2_{\Gamma^{-1}}}{2} + V(z^{k+1} - z) - V(z^k - z)&\\ + \langle A(x^{k+1} - x) \mid \bar{v}^k\rangle + \langle b^\delta - Ax^{k+1} \mid u^{k+1} - u\rangle&\\ \le 0.&\end{aligned}\tag{A.24}$$

Now, compute

$$\begin{aligned}&J(x^{k+1}) - J(x) + \langle A(x^{k+1} - x) \mid \bar{v}^k\rangle + \langle b^\delta - Ax^{k+1} \mid u^{k+1} - u\rangle\\ &\quad= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) - \langle Ax^{k+1} - b \mid u\rangle + \langle Ax - b \mid u^{k+1}\rangle\\ &\qquad + \langle A(x^{k+1} - x) \mid \bar{v}^k\rangle + \langle b^\delta - Ax^{k+1} \mid u^{k+1} - u\rangle\\ &\quad= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) - \langle Ax^{k+1} \mid u\rangle + \langle b \mid u\rangle + \langle Ax \mid u^{k+1}\rangle - \langle b \mid u^{k+1}\rangle\\ &\qquad + \langle A(x^{k+1} - x) \mid \bar{v}^k\rangle + \langle b^\delta \mid u^{k+1} - u\rangle - \langle Ax^{k+1} \mid u^{k+1}\rangle + \langle Ax^{k+1} \mid u\rangle\\ &\quad= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) + \langle b^\delta - b \mid u^{k+1} - u\rangle + \langle A(x^{k+1} - x) \mid \bar{v}^k - u^{k+1}\rangle\\ &\quad= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) + \langle b^\delta - b \mid u^{k+1} - u\rangle + \langle A(x^{k+1} - x) \mid v^k - u^{k+1}\rangle\\ &\qquad + \langle A(x^{k+1} - x) \mid u^k - v^{k-1}\rangle\\ &\quad= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) + \langle b^\delta - b \mid u^{k+1} - u\rangle + \langle A(x^{k+1} - x) \mid v^k - u^{k+1}\rangle\\ &\qquad + \langle A(x^k - x) \mid u^k - v^{k-1}\rangle + \langle A(x^{k+1} - x^k) \mid u^k - v^{k-1}\rangle\\ &\quad= \mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) + \langle b^\delta - b \mid u^{k+1} - u\rangle + \langle A(x^{k+1} - x) \mid v^k - u^{k+1}\rangle\\ &\qquad + \langle A(x^k - x) \mid u^k - v^{k-1}\rangle + \langle \Gamma^{\frac{1}{2}} A(x^{k+1} - x^k) \mid \Gamma^{-\frac{1}{2}}(u^k - v^{k-1})\rangle.\end{aligned}\tag{A.25}$$

From (A.25) and (A.24), we obtain

$$\begin{aligned}&\mathcal{L}(x^{k+1}, u) - \mathcal{L}(x, u^{k+1}) + \frac{\|x^{k+1} - x^k\|^2_{\Sigma^{-1}}}{2} + \frac{\|u^{k+1} - v^k\|^2_{\Gamma^{-1}}}{2} + V(z^{k+1} - z) - V(z^k - z)\\ &\quad\le -\langle b^\delta - b \mid u^{k+1} - u\rangle - \langle A(x^{k+1} - x) \mid v^k - u^{k+1}\rangle + \langle A(x^k - x) \mid v^{k-1} - u^k\rangle\\ &\qquad - \langle \Gamma^{\frac{1}{2}} A(x^{k+1} - x^k) \mid \Gamma^{-\frac{1}{2}}(u^k - v^{k-1})\rangle\\ &\quad\le \delta\|\Gamma^{\frac{1}{2}}\|\|u^{k+1} - u\|_{\Gamma^{-1}} - \langle A(x^{k+1} - x) \mid v^k - u^{k+1}\rangle + \langle A(x^k - x) \mid v^{k-1} - u^k\rangle\\ &\qquad + \frac{\|x^{k+1} - x^k\|^2_{\Sigma^{-1}}}{2} + \|\Gamma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}\|^2 \frac{\|u^k - v^{k-1}\|^2_{\Gamma^{-1}}}{2}.\end{aligned}$$

Therefore, we have that

$$\begin{aligned}\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1})&+\frac{\|u^{k+1}-v^k\|^2_{\Gamma^{-1}}}{2}-\|\Gamma^{\frac12}A\Sigma^{\frac12}\|^2\frac{\|u^k-v^{k-1}\|^2_{\Gamma^{-1}}}{2}+V(z^{k+1}-z)-V(z^k-z)\\&\le \delta\|\Gamma^{\frac12}\|\|u^{k+1}-u\|_{\Gamma^{-1}}-\langle A(x^{k+1}-x)\mid v^k-u^{k+1}\rangle+\langle A(x^k-x)\mid v^{k-1}-u^k\rangle.\end{aligned}$$

Summing from 1 to $N-1$, we obtain

$$\begin{aligned}\sum_{k=1}^{N-1}(\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1}))&+\frac{\alpha}{2}\sum_{k=1}^{N-1}\|u^{k+1}-v^k\|^2_{\Gamma^{-1}}+V(z^N-z)+\|\Gamma^{\frac12}A\Sigma^{\frac12}\|^2\frac{\|u^N-v^{N-1}\|^2_{\Gamma^{-1}}}{2}\\&\le \delta\|\Gamma^{\frac12}\|\sum_{k=1}^{N-1}\|u^{k+1}-u\|-\langle A(x^N-x)\mid v^{N-1}-u^N\rangle+\langle A(x^1-x)\mid v^0-u^1\rangle+V(z^1-z)\\&\le \delta\|\Gamma^{\frac12}\|\sum_{k=1}^{N-1}\|u^{k+1}-u\|+\|\Gamma^{\frac12}A\Sigma^{\frac12}\|^2\frac{\|u^N-v^{N-1}\|^2_{\Gamma^{-1}}}{2}+\frac{\|x^N-x\|^2_{\Sigma^{-1}}}{2}\\&\quad+\langle A(x^1-x)\mid v^0-u^1\rangle+V(z^1-z).\end{aligned}\tag{A.26}$$

Reordering (A.26), we obtain

$$\begin{aligned}\sum_{k=1}^{N-1}(\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1}))&+\frac{\alpha}{2}\sum_{k=1}^{N-1}\|u^{k+1}-v^k\|^2_{\Gamma^{-1}}+\frac{\|u^N-u\|^2_{\Gamma^{-1}}}{2}\\&\le \delta\|\Gamma^{\frac12}\|\sum_{k=1}^{N-1}\|u^{k+1}-u\|+\langle A(x^1-x)\mid v^0-u^1\rangle+V(z^1-z).\end{aligned}\tag{A.27}$$

On the other hand, from (A.24) and (A.25) we get

$$\mathcal{L}(x^1,u)-\mathcal{L}(x,u^1)+\frac{\alpha}{2}\|u^1-v^0\|^2\le\delta\|u^1-u\|-\langle A(x^1-x)\mid \bar{v}^0-u^1\rangle V(z^0-z)-V(z^1-z).\tag{A.28}$$

Summing (A.27) and (A.28) yields

$$\sum_{k=1}^{N}(\mathcal{L}(x^k,u)-\mathcal{L}(x,u^k))+\frac{\alpha}{2}\sum_{k=1}^{N}\|u^k-v^{k-1}\|^2_{\Gamma^{-1}}+\frac{\|u^N-u\|^2_{\Gamma^{-1}}}{2}\le\delta\|\Gamma^{\frac12}\|\sum_{k=1}^{N}\|u^k-u\|+V(z^0-z).\tag{A.29}$$

Moreover, since $u^{k+1}-v^k=\Gamma(Ax^{k+1}-b^\delta)$, we have

$$\begin{aligned}\|u^{k+1}-v^k\|^2_{\Gamma^{-1}}&=\langle\Gamma(Ax^{k+1}-b^\delta)\mid Ax^{k+1}-b^\delta\rangle\\&\ge\frac{\|Ax^{k+1}-b^\delta\|^2}{\|\Gamma^{-1}\|}\\&\ge\frac{1}{\|\Gamma^{-1}\|}\Big(\frac{\|Ax^{k+1}-b\|^2}{2}-\|b^\delta-b\|^2\Big).\end{aligned}\tag{A.30}$$

From (A.29) and (A.30), we obtain

$$\begin{aligned}\sum_{k=0}^{N-1}(\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1}))&+\frac{\alpha}{4\|\Gamma^{-1}\|}\sum_{k=1}^{N}\|Ax^k-b\|^2+\frac{\|u^N-u\|^2_{\Gamma^{-1}}}{2}\\&\le\delta\|\Gamma^{\frac12}\|\sum_{k=1}^{N}\|u^k-u\|_{\Gamma^{-1}}+V(z^0-z)+\frac{\alpha N\delta^2}{2\|\Gamma^{-1}\|}.\end{aligned}\tag{A.31}$$

From (A.29), it follows that

$$\|u^N-u\|^2_{\Gamma^{-1}}\le2\delta\|\Gamma^{\frac12}\|\sum_{k=1}^{N}\|u^k-u\|_{\Gamma^{-1}}+2V(z^0-z).\tag{A.32}$$

Apply [62, Lemma A.1] to equation (A.32) with $\lambda_k:=2\delta\|\Gamma^{\frac12}\|$ and $S_k:=2V(z^0-z)$ to get

$$\|u^k-u\|_{\Gamma^{-1}}\le N\|\Gamma^{\frac12}\|\delta+(2V(z^0-z)+(N\|\Gamma^{\frac12}\|\delta)^2)^{\frac12}\le2N\|\Gamma^{\frac12}\|\delta+(2V(z^0-z))^{\frac12}.\tag{A.33}$$

Insert (A.33) into equation (A.29) to obtain

$$\sum_{k=0}^{N-1}(\mathcal{L}(x^{k+1},u)-\mathcal{L}(x,u^{k+1})) \le 2\|\Gamma^{\frac{1}{2}}\|^2N^2\delta^2 + N\|\Gamma^{\frac{1}{2}}\|\delta(2V(z^0-z))^{\frac{1}{2}} + V(z^0-z).$$

By (A.31) and (A.33), we have

$$\sum_{k=1}^{N}\|Ax^k-b\|^2 \le \frac{4\|\Gamma^{-1}\|}{\alpha}\Big(2\|\Gamma^{\frac{1}{2}}\|^2N^2\delta^2 + N\|\Gamma^{\frac{1}{2}}\|\delta(2V(z^0-z))^{\frac{1}{2}} + V(z^0-z) + \frac{\alpha N\delta^2}{2\|\Gamma^{-1}\|}\Big),$$

and both results follows from Jensen's inequality. □

A.4 Proof of Lemma 5.2

Proof. Let us first recall that

$$P^\delta : x \mapsto x + \frac{b_j^\delta - \langle a_j \mid x\rangle}{\|a_j\|^2}a_j^*. \tag{A.34}$$

Note that the j-th equation of C and C_δ are parallel. Then, for every $j \in [d]$ and $\bar{x} \in C$, we get

$$\begin{aligned}\|P_j^\delta x - \bar{x}\|^2 &= \|P_jx - \bar{x}\|^2 + 2\langle P_jx - \bar{x} \mid P_j^\delta x - P_jx\rangle + \|P_jx - P_j^\delta x\|^2\\ &= \|P_jx - \bar{x}\|^2 + \|P_jx - P_j^\delta x\|^2.\end{aligned} \tag{A.35}$$

Analogously, we have that

$$\|x-\bar{x}\|^2 = \|x - P_jx\|^2 + \|P_jx - \bar{x}\|^2. \tag{A.36}$$

It follows from (A.35) and (A.36) that

$$\|P_j^\delta x - \bar{x}\|^2 + \|x - P_jx\|^2 = \|x-\bar{x}\|^2 + \|P_j^\delta x - P_jx\|^2.$$

Hence,

$$\begin{aligned}\|P_j^\delta x - \bar{x}\|^2 &\le \|x-\bar{x}\|^2 + \|P_j^\delta x - P_jx\|^2\\ &\le \|x-\bar{x}\|^2 + \frac{(b_j^\delta - b_j)^2}{\|a_j\|^2}\\ &\le \|x-\bar{x}\|^2 + \frac{\delta^2}{\|a_j\|^2}.\end{aligned}$$

(i) Since $T = P_{\beta_l}^\delta \circ \cdots \circ P_{\beta_1}^\delta$, it is clear that $C_\delta \subset \operatorname{Fix} T$ and, by induction, we have that

$$\|Tx-\bar{x}\|^2 \le \|x-\bar{x}\|^2 + e\delta^2,$$

where

$$e = \frac{l}{\max_{i=1,\dots,d}\|a_i\|}.$$

(ii) The proof follows from the convexity of $\|\cdot\|^2$, which is obtained with

$$e = \frac{1}{\max_{i=1,\dots,d}\|a_i\|}.$$

(iii) Let $\bar{x} \in C$. By (5.2), we have

$$\begin{aligned}\|Tx-\bar{x}\|^2 &= \|x-\bar{x}\|^2 - 2\alpha\langle x-\bar{x} \mid A^*(Ax-b^\delta)\rangle + \alpha^2\|A^*(Ax-b^\delta)\|^2\\ &= \|x-\bar{x}\|^2 - 2\alpha\langle Ax-b \mid Ax-b^\delta\rangle + \alpha^2\|A^*(Ax-b^\delta)\|^2\\ &\le \|x-\bar{x}\|^2 - 2\alpha\langle b^\delta-b \mid Ax-b^\delta\rangle + (\alpha^2\|A\|^2-2\alpha)\|Ax-b^\delta\|^2.\end{aligned}$$

Now, using the Young inequality with parameter $2-\alpha\|A\|^2$, we have that

$$\|Tx-\bar{x}\|^2 \le \|x-\bar{x}\|^2 + \frac{\alpha}{2-\alpha\|A\|^2}\|b^\delta-b\|^2 \le \|x-\bar{x}\|^2 + \frac{\alpha\delta^2}{2-\alpha\|A\|^2}.$$

It remains to prove that, if $C_\delta \ne 0$, then $C_\delta \subset \operatorname{Fix} T$, which is clear from (5.2).

(iv) Let $\bar{x} \in C$ and $x \in \mathbb{R}^p$. If $A^*Ax = A^*b^\delta$, then (3.3) immediately holds. Otherwise, we have

$$\begin{aligned} \|Tx - \bar{x}\|^2 &= \|x - \bar{x}\|^2 - 2\beta(x)\langle x - \bar{x} \mid A^*(Ax - b^\delta)\rangle + \beta(x)^2\|A^*(Ax - b^\delta)\|^2 \\ &= \|x - \bar{x}\|^2 - 2\beta(x)\langle Ax - b \mid Ax - b^\delta\rangle + \beta(x)^2\|A^*(Ax - b^\delta)\|^2 \\ &= \|x - \bar{x}\|^2 - 2\beta(x)\langle b^\delta - b \mid Ax - b^\delta\rangle - 2\beta(x)\|Ax - b^\delta\|^2 + \beta(x)^2\|A^*(Ax - b^\delta)\|^2. \end{aligned} \tag{A.37}$$

Now, using the Young inequality with parameter

$$2 - \beta(x)\frac{\|A^*(Ax - b^\delta)\|^2}{\|Ax - b^\delta\|^2},$$

we have that

$$\|Tx - \bar{x}\|^2 \le \|x - \bar{x}\|^2 + \frac{\beta(x)}{2 - \beta(x)(\|A^*(Ax - b^\delta)\|^2/\|Ax - b^\delta\|^2)}\|b^\delta - b\|^2 \le \|x - \bar{x}\|^2 + M\delta^2.$$

Finally, it is clear from (5.3) that, if $C_\delta \neq 0$, then $C_\delta \subset \operatorname{Fix} T$.

□

Funding: This project has been supported by the TraDE-OPT project, which received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 861137. The third author acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-18-1-7009 (European Office of Aerospace Research and Development), the EU H2020-MSCA-RISE project NoMADS-DLV-777826, and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. The third and fourth authors acknowledge the support of the AFOSR project FA8655-22-1-7034. The research by the first, second, and fourth authors has been supported by the MIUR Excellence Department Project awarded to Dipartimento di Matematica, Università di Genova, CUP D33C23001110001. The second and fourth authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). This work represents only the view of the authors. The European Commission and the other organizations are not responsible for any use that may be made of the information it contains.

References

[1] A. Alacaoglu, O. Fercoq and V. Cevher, On the convergence of stochastic primal-dual hybrid gradient, *SIAM J. Optim.* **32** (2022), no. 2, 1288–1318.

[2] M. Bachmayr and M. Burger, Iterative total variation schemes for nonlinear inverse problems, *Inverse Problems* **25** (2009), no. 10, Article ID 105004.

[3] M. A. Bahraoui and B. Lemaire, Convergence of diagonally stationary sequences in convex optimization, *Set-Valued Anal.* **2** (1994), 49–61.

[4] A. B. Bakushinsky and M. Y. Kokurin, *Iterative Methods for Approximate Solution of Inverse Problems*, Math. Appl. (New York) 577, Springer, Dordrecht, 2005.

[5] P. L. Bartlett and M. Traskin, AdaBoost is consistent, *J. Mach. Learn. Res.* **8** (2007), 2347–2368.

[6] F. Bauer, S. Pereverzev and L. Rosasco, On regularization algorithms in learning theory, *J. Complexity* **23** (2007), no. 1, 52–72.

[7] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, Cham, 2017.

[8] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Oper. Res. Lett.* **31** (2003), no. 3, 167–175.

[9] S. Becker, J. Bobin and E. J. Candès, NESTA: A fast and accurate first-order method for sparse recovery, *SIAM J. Imaging Sci.* **4** (2011), no. 1, 1–39.

[10] M. Benning and M. Burger, Error estimates for general fidelities, *Electron. Trans. Numer. Anal.* **38** (2011), 44–68.

[11] M. Benning and M. Burger, Modern regularization methods for inverse problems, *Acta Numer.* **27** (2018), 1–111.

[12] G. Blanchard and N. Krämer, Optimal learning rates for kernel conjugate gradient regression, in: *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, ACM, New York (2010), 226–234.

[13] R. I. Boţ and T. Hein, Iterative regularization with a general penalty term—theory and application to L^1 and *TV* regularization, *Inverse Problems* **28** (2012), no. 10, Article ID 104010.

[14] L. M. Briceño Arias, A Douglas–Rachford splitting method for solving equilibrium problems, *Nonlinear Anal.* **75** (2012), no. 16, 6053–6059.

[15] L. M. Briceño Arias, Forward-Douglas–Rachford splitting and forward-partial inverse method for solving monotone inclusions, *Optimization* **64** (2015), no. 5, 1239–1261.

[16] L. M. Briceño Arias, J. Deride and C. Vega, Random activations in primal-dual splittings for monotone inclusions with a priori information, *J. Optim. Theory Appl.* **192** (2022), no. 1, 56–81.

[17] L. M. Briceño Arias and S. López Rivera, A projected primal-dual method for solving constrained monotone inclusions, *J. Optim. Theory Appl.* **180** (2019), no. 3, 907–924.

[18] M. Burger, E. Resmerita and L. He, Error estimation for Bregman iterations and inverse scale space methods in image restoration, *Computing* **81** (2007), no. 2–3, 109–135.

[19] J.-F. Cai, E. J. Candès and Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM J. Optim.* **20** (2010), no. 4, 1956–1982.

[20] J.-F. Cai, S. Osher and Z. Shen, Linearized Bregman iterations for frame-based image deblurring, *SIAM J. Imaging Sci.* **2** (2009), no. 1, 226–252.

[21] L. Calatroni, G. Garrigos, L. Rosasco and S. Villa, Accelerated iterative regularization via dual diagonal descent, *SIAM J. Optim.* **31** (2021), no. 1, 754–784.

[22] E. J. Candès, Matrix completion with noise, *Proc. IEEE* **98** (2010), no. 6, 925–936.

[23] E. J. Candès and B. Recht, Exact matrix completion via convex optimization, *Found. Comput. Math.* **9** (2009), no. 6, 717–772.

[24] E. J. Candès, J. Romberg and T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. Inform. Theory* **52** (2006), no. 2, 489–509.

[25] E. J. Candes and T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, *IEEE Trans. Inform. Theory* **52** (2006), no. 12, 5406–5425.

[26] A. Chambolle, An algorithm for total variation minimization and applications, *J. Math. Imaging Vision* **20** (2004), 89–97.

[27] A. Chambolle, M. J. Ehrhardt, P. Richtárik and C.-B. Schönlieb, Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications, *SIAM J. Optim.* **28** (2018), no. 4, 2783–2808.

[28] A. Chambolle and P.-L. Lions, Image recovery via total variation minimization and related problems, *Numer. Math.* **76** (1997), no. 2, 167–188.

[29] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *J. Math. Imaging Vision* **40** (2011), no. 1, 120–145.

[30] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.* **4** (2005), no. 4, 1168–1200.

[31] L. Condat, A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms, *J. Optim. Theory Appl.* **158** (2013), no. 2, 460–479.

[32] C. De Mol, E. De Vito and L. Rosasco, Elastic-net regularization in learning theory, *J. Complexity* **25** (2009), no. 2, 201–230.

[33] D. L. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* **52** (2006), no. 4, 1289–1306.

[34] J. Duchi and Y. Singer, Efficient online and batch learning using forward backward splitting, *J. Mach. Learn. Res.* **10** (2009), 2899–2934.

[35] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic, Dordrecht, 1996.

[36] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Appl. Numer. Harmon. Anal., Birkhäuser/Springer, New York, 2013.

[37] G. Garrigos, L. Rosasco and S. Villa, Iterative regularization via dual diagonal descent, *J. Math. Imaging Vision* **60** (2018), no. 2, 189–215.

[38] G. H. Golub, M. Heath and G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21** (1979), no. 2, 215–223.

[39] E. B. Gutiérrez, C. Delplancke and M. J. Ehrhardt, Convergence properties of a randomized primal-dual algorithm with applications to parallel mri, in: *International Conference on Scale Space and Variational Methods in Computer Vision*, Springer, New York (2021), 254–266.

[40] M. Jaggi, Revisiting Frank-Wolfe: Projection-free sparse convex optimization, International conference on machine learning, *Proc. Mach. Learn. Res. (PMLR)* **28** (2013), 427–435.

[41] B. Jin, D. A. Lorenz and S. Schiffler, Elastic-net regularization: Error estimates and active set methods, *Inverse Problems* **25** (2009), no. 11, Article ID 115022.

[42] S. Kaczmarz, Angenäherte Auflösung von Systemen linearer Gleichungen, *Bull. Int. Acad. Pol. Sic. Let. Cl. Sci. Math. Nat.* **35** (1937), 335–357.

[43] B. Kaltenbacher, A. Neubauer and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, Radon Ser. Comput. Appl. Math. 6, Walter de Gruyter, Berlin, 2008.

[44] L. Landweber, An iteration formula for Fredholm integral equations of the first kind, *Amer. J. Math.* **73** (1951), 615–624.

[45] H. Li, N. Chen and L. Li, Error analysis for matrix elastic-net regularization algorithms, *IEEE Trans. Neural Netw. Learn. Syst.* **23** (2012), no. 5, 737–748.

[46] D. A. Lorenz, Convergence rates and source conditions for Tikhonov regularization with sparsity constraints, *J. Inverse Ill-Posed Probl.* **16** (2008), no. 5, 463–478.

[47] S. Matet, L. Rosasco, S. Villa and B. L. Vu, Don’t relax: Early stopping for convex regularization, preprint (2017), https://arxiv.org/abs/1707.05422.

[48] C. Molinari, J. Liang and J. Fadili, Convergence rates of forward-Douglas–Rachford splitting method, *J. Optim. Theory Appl.* **182** (2019), no. 2, 606–639.

[49] C. Molinari, M. Massias, L. Rosasco and S. Villa, Iterative regularization for convex regularizers, *Proc. Mach. Learn. Res. (PMLR)* **130** (2021), 1684–1692.

[50] C. Molinari, M. Massias, L. Rosasco and S. Villa, Iterative regularization for low complexity regularizers, preprint (2022), https://arxiv.org/abs/2202.00420.

[51] C. Molinari and J. Peypouquet, Lagrangian penalization scheme with parallel forward-backward splitting, *J. Optim. Theory Appl.* **177** (2018), no. 2, 413–447.

[52] C. Molinari, J. Peypouquet and F. Roldan, Alternating forward-backward splitting for linearly constrained optimization problems, *Optim. Lett.* **14** (2020), no. 5, 1071–1088.

[53] E. Moulines and F. Bach, Non-asymptotic analysis of stochastic approximation algorithms for machine learning, in: *Advances in Neural Information Processing Systems 24*, Morgan Kaufmann, Burlington (2011), 451–459.

[54] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley & Sons, New York, 1983.

[55] A. Neubauer, On Nesterov acceleration for Landweber iteration of linear ill-posed problems, *J. Inverse Ill-Posed Probl.* **25** (2017), no. 3, 381–390.

[56] S. Osher, M. Burger, D. Goldfarb, J. Xu and W. Yin, An iterative regularization method for total variation-based image restoration, *Multiscale Model. Simul.* **4** (2005), no. 2, 460–489.

[57] S. Osher, Y. Mao, B. Dong and W. Yin, Fast linearized Bregman iteration for compressive sensing and sparse denoising, *Commun. Math. Sci.* **8** (2010), no. 1, 93–111.

[58] S. Osher and L. I. Rudin, Feature-oriented image enhancement using shock filters, *SIAM J. Numer. Anal.* **27** (1990), no. 4, 919–940.

[59] G. Peyré, The numerical tours of signal processing-advanced computational signal and image processing, *IEEE Comput. Sci. Eng.* **13** (2011), no. 4, 94–97.

[60] T. Pock and A. Chambolle, Diagonal preconditioning for first order primal-dual algorithms in convex optimization, in: *2011 International Conference on Computer Vision*, IEEE Press, Piscataway (2011), 1762–1769.

[61] H. Raguet, J. Fadili and G. Peyré, A generalized forward-backward splitting, *SIAM J. Imaging Sci.* **6** (2013), no. 3, 1199–1226.

[62] J. Rasch and A. Chambolle, Inexact first-order primal-dual algorithms, *Comput. Optim. Appl.* **76** (2020), no. 2, 381–430.

[63] G. Raskutti, M. J. Wainwright and B. Yu, Early stopping and non-parametric regression: An optimal data-dependent stopping rule, *J. Mach. Learn. Res.* **15** (2014), 335–366.

[64] L. Rosasco and S. Villa, Learning with incremental iterative regularization, in: *Advances in Neural Information Processing Systems 28*, Curran Associates, Red Hook (2015), 1630–1638.

[65] M. Rudelson and R. Vershynin, Geometric approach to error-correcting codes and reconstruction of signals, *Int. Math. Res. Not. IMRN* **2005** (2005), no. 64, 4019–4041.

[66] L. I. Rudin and S. Osher, Total variation based image restoration with free local constraints, in: *Proceedings of 1st International Conference on Image Processing*, IEEE Press, Piscataway (1994), 31–35.

[67] L. I. Rudin, S. Osher and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Phys. D* **60** (1992), 259–268.

[68] O. Scherzer, A modified Landweber iteration for solving parameter estimation problems, *Appl. Math. Optim.* **38** (1998), no. 1, 45–68.

[69] F. Schöpfer and D. A. Lorenz, Linear convergence of the randomized sparse Kaczmarz method, *Math. Program.* **173** (2019), no. 1, 509–536.

[70] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University, Cambridge, 2014.

[71] A. Silveti-Falls, C. Molinari and J. Fadili, Generalized conditional gradient with augmented Lagrangian for composite minimization, *SIAM J. Optim.* **30** (2020), no. 4, 2687–2725.

[72] A. Silveti-Falls, C. Molinari and J. Fadili, Inexact and stochastic generalized conditional gradient with augmented Lagrangian and proximal step, *J. Nonsmooth Anal. Optim.* **2** (2021), 1–41.

[73] A. Silveti-Falls, C. Molinari and J. Fadili, A stochastic Bregman primal-dual splitting algorithm for composite optimization, *Pure Appl. Funct. Anal.* **8** (2023), no. 3, 921–964.

[74] I. Steinwart and A. Christmann, *Support Vector Machines*, Inform. Sci. Stat., Springer, New York, 2008.

[75] T. Strohmer and R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, *J. Fourier Anal. Appl.* **15** (2009), no. 2, 262–278.

[76] A. N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, *Soviet Math.* **4** (1963), 1035–1038.

[77] Y. Tsaig and D. L. Donoho, Extensions of compressed sensing, *Signal Process.* **86** (2006), no. 3, 549–571.

[78] B. C. Vũ, A splitting algorithm for dual monotone inclusions involving cocoercive operators, *Adv. Comput. Math.* **38** (2013), no. 3, 667–681.

[79] L. Xiao, Dual averaging methods for regularized stochastic learning and online optimization, *J. Mach. Learn. Res.* **11** (2010), 2543–2596.

[80] Y. Yao, L. Rosasco and A. Caponnetto, On early stopping in gradient descent learning, *Constr. Approx.* **26** (2007), no. 2, 289–315.

[81] W. Yin, Analysis and generalizations of the linearized Bregman model, *SIAM J. Imaging Sci.* **3** (2010), no. 4, 856–877.

[82] W. Yin, S. Osher, D. Goldfarb and J. Darbon, Bregman iterative algorithms for l_1-minimization with applications to compressed sensing, *SIAM J. Imaging Sci.* **1** (2008), no. 1, 143–168.

[83] T. Zhang and B. Yu, Boosting with early stopping: Convergence and consistency, *Ann. Statist.* **33** (2005), no. 4, 1538–1579.

[84] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** (2005), no. 2, 301–320.

[85] H. Zou and H. H. Zhang, On the adaptive elastic-net with a diverging number of parameters, *Ann. Statist.* **37** (2009), no. 4, 1733–1751.